分类号 G30　　密级

UDC　　编号

中国科学院研究生院

博士学位论文

知识结构演化深度分析的方法及其实现

韩 涛

指导教师 张晓林 教授

中国科学院文献情报中心

申请学位级别 博士学位 学科专业名称 图书馆学

论文提交日期 2008 年 5 月 论文答辩日期 2008 年 6 月

培养单位 中国科学院文献情报中心

学位授予单位 中国科学院研究生院

答辩委员会主席

Design and Implementation of Method for Deep Exploring the Evolution of Knowledge Structure

A Dissertation for the Doctoral Degree of Management

in the Graduate School of Chinese Academy of Sciences

By

Han Tao

Directed By

Professor Zhang Xiaolin

Chinese Academy of Sciences

June,2008

关于学位论文使用权声明

任何收存和保管本论文各种版本的单位和个人，未经著作权人授权，不得将本论文转借他人并复印、抄录、拍照、或以任何方式传播。否则，引起有碍著作权人著作权益之问题，将可能承担法律责任。

……

关于学位论文使用授权的说明

本人完全了解中国科学院文献情报中心有关保存、使用学位论文的规定，即：中国科学院文献情报中心有权保留学位论文的副本，允许该论文被查阅；中国科学院文献情报中心可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

（涉密的学位论文在解密后应遵守此规定）

签　名：　　导师签名：　　日　期：

……

关于学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：　　导师签名：　　日　期：

摘要

情报研究工作中目前有很多方法和工具进行知识结构演化的科学情报分析，但它们主要局限在静态关联强度阈值层聚类结构和静态时间窗聚类结构的分析上，主要描述知识结构的宏观状态，虽然也能借助多个静态宏观结构（例如不同阈值层的聚类结构）人工比较分析，但难以自动对不同聚类结构的差异、性质及相应的演变倾向进行探测，因此还难以有效支持对知识结构演化的深层次自动分析以及对潜在演变趋势的探测。

针对这个问题，对发现潜在知识、发现新知识、监测知识突变、分析知识转移、识别重要节点、展示演化结构等方面的研究进行了分析，指出现有的研究对潜在的、萌芽的、有强烈演化可能的知识的自动探测存在严重不足。

在共引聚类生成知识结构的基础上，提出不同阈值层聚类结构间差异性自动检测分析、自动发现宏观结构下的潜在结构的方法，提出不同时间窗聚类结构主题簇传承关系自动检测分析、自动发现不同时间聚类结构的演变趋势，并通过一个以这两种检测机制为核心模块的知识结构演化分析试验系统，对这两种方法进行初步验证。不同阈值层聚类结构间差异性自动检测方法称为阈值层潜在结构检测策略（简称阈值策略），不同时间窗聚类结构间主题簇传承关系自动检测分析方法称为时间窗演变结构检测策略（简称时间策略）。阈值策略的目标是寻找潜藏在宏观聚类结构下的孕育中、发展中的对象和关系，时间策略的目标是寻找在时间流中聚类结构包含的演变关系和演变过程。

阈值策略着重分析不同阈值层聚类结构中聚类簇的相互关系，定义同一阈值层中簇与簇的 relative 关系和不同阈值层之间的 twin 关系，划分并定义 5 个种类的潜在簇。进一步，设计阈值策略实施的步骤，包括突变阈值区间的确定、两个阈值层 relative 关系和 twin 关系的构建、5 种潜在簇的寻找、潜在关系的确定，最后是潜在结构的可视化展示。

时间策略着重分析不同时间窗聚类结构中聚类簇的相互关系，构建时间窗间簇与簇的 ancestor 关系和 offspring 关系，分析时间窗间的“遗传继承”性。利用 ancestor 关系和 offspring 关系，划分并定义 6 种演变簇。设计时间策略实施的步骤，包括时间窗间 ancestor 关系和 offspring 关系的构建、6 种演变簇的寻找、演变关系的确定，最后是潜在结构的可视化展示。

通过系统分析与建模，实现试验系统，初步证实两种策略的有效性。

本论文包括图幅 82、表 4 个、附录 2 个。

关键词：知识结构演化 阈值 时间窗 策略 潜在结构 演变结构

Design and Implementation of Method
for Deep Exploring the Evolution of Knowledge Structure

Abstract

Many methods and tools are available to analyze the evolution of knowledge structure, but limited to cluster structure analysis in terms of static relation threshold and static time slice. And manual comparison of static structures (e.g. cluster structures of different threshold) can't analyze differences, property and corresponding trend of evolution automatically, is difficult to support automatic and deep-seated analysis of the evolution structure and detect the latent evolution direction.

To the problem, after research of methods (exploring latent, new, bursty, transfering, important knowledge and displaying result), shortage is pointed out, including that automatic discovery of latent knowledge is of low efficiency, and that category and character of evolution are not depicted definitely.

Based on knowledge structure derived from co-citation and cluster, two methods are proposed. One is to automatically analyze the differences of multi-threshold-level cluster structures and detect the latent structure under the higher level, which is called threshold level structure difference detection strategy (abbr. threshold strategy) and aimed to explore the gestating and developing objects and their relations underlying the main objects and relations. The other is to automatically analyze the evolution relation of clusters of different time slices and detect the evolution direction in various periods, which is called time slice evolution structure analysis strategy (abbr. time strategy) and aimed to explore the relation and process of evolution structure in time stream. Moreover a test system is built with two kernel modules of threshold strategy and time strategy. Methods of tow strategies are certified by the system preliminarily.

In terms of threshold strategy, "relative" and "twin" relations are defined, the former for the clusters on the same threshold level, the latter for the different levels. Based on two relations, latent cluster is defined and divided into 5 categories. Next, implementary and display process of threshold strategy is designed.

In terms of time strategy, "ancestor" and "offspring" relations are defined, which ensure the quality of heredity and inherity. Based on the two relations, evolving cluster is defined and divided into 6 categories. Next, implementary and display process of time strategy is designed.

82 diagrams, 4 tables and 2 appendices are included.

Keywords: knowledge structure evolution, threshold, time slice, strategy, latent structure, evolution structure

目 录

图目录

表目录

1 引言

1.1 研究背景

1.1.1 科技情报工作面临的挑战

在“全面推进中国特色国家创新体系建设”的要求下，由于现代信息技术的飞速发展，科技情报工作面临机遇与挑战。一方面，信息技术的发展、信息环境的变化为情报工作提供了许多工具与资源，如数据挖掘与可视化技术为科学计量学的发展提高了数据处理的效率和质量；另一方面，面对信息爆炸和情报不足并存的现象，如何有效利用海量信息，是需要解决的迫切问题；另外，科学知识体系逐步庞大、细化，各类学科之间相互交叉渗透、汇聚融合的频率增加，联系不断增强，这些都使得科技情报研究的对象和需求日益复杂，要求情报人员为政府、企业等决策提供高增值的情报支撑。

科学发展日新月异，某个学科领域的突破会对其它领域产生巨大的影响，国家、研究所对科研的战略选择也会受到学科之间关系的影响。而以往知识演化状况的分析多依赖于学科专家的知识，但是由于专家自身专业知识范畴的限制以及方法有效性不足等等问题，使得对科学发展的分析结果会缺乏一定的全局性和系统性，影响了决策的科学性。因此，为科研战略选择提供有效的支持，要求科技情报研究必须重视某领域的、学科领域之间的知识演化关系的分析揭示。

目前科技信息分析正在充分利用数学、信息技术的最新成果，如何有效利用最新的工具方法、更全面有效地揭示科学知识的时空特征、发展演化关系，是情报研究中的主要课题。

1.1.2 情报研究对知识结构演化分析的需求层次

在科学情报研究方法和技术的不同发展阶段，情报研究工作中对知识结构演化分析的能力处于不同的发展阶段，相应的，情报研究工作对知识结构演化分析的需求也处于不同的发展阶段。归纳起来，情报研究工作对知识结构演化分析的需求可以分为以下几个层次：

（1）第一个层次是基于宏观数量统计的知识结构演化分析

基于宏观数量统计的知识结构分析是在宏观数量统计的基础上以描述知识结构宏观变化趋势为目的的情报研究工作。这种知识结构演化分析的目标是在统计量上全面了解主题或者学科领域的研究内容，从整体上考察主题或者学科领域

的研究体系，从宏观上分析主题或者学科领域的发展趋势。分析的对象通常是一个较长时间段内的所有相关的物理对象，如一篇篇的论文或者专著。情报研究工作中的“态势分析与研究”，就属于这种静态的知识结构分析。

（2）第二个层次是基于演化特征的知识结构演化分析

基于演化特征的知识结构分析是通过分析各种演化特征以描述知识微观结构与变化情况为目的的情报研究工作。演化特征包括融合、扩散、新增、消失等。这种知识结构演化分析的目标体现在两个方面：一个是已经发生的演化，即在宏观知识结构演化分析的基础上，考察主题或者学科领域的局部知识体系的变化，了解变化涉及到的对象和关系，探究变化前后的知识结构的差异，分析变化的产生原因；另一个是可能发生的演化，即如果继续向后延伸，还可以考察有哪些知识可能演化，会发生什么样的演化，进而分析知识体系以后变化的趋势。分析的对象依旧是所有相关的物理对象，但是这些对象应该以某种维度在逻辑上被划分成几个部分，将这几个部分之间的差异作为分析的主要关注点，以探讨已经发生的演化和可能发生的演化。这个层次还有一个需求是得到的结果应该用可视化方法来表达，增强可读性，帮助用户理解。同时还要建立与用户的互动反馈机制。

1.2 问题提出

现有的知识结构演化分析方法还不足以满足用户的需求：

（1）许多方法局限在宏观数量统计的基础上描述知识结构的宏观变化趋势。这些方法虽然从发表量、引用频次等各种角度在数量上进行统计分析，但是反映的只是知识体系演化的整体过程和宏观走向，无法体现知识体系演化的历史阶段特征和微观特征。

（2）多数方法局限在静态知识结构的分析上，这些结构或者产生于一定关系强度阈值的聚类分析，或者产生于特定时间窗的聚类分析（也依赖一定聚类强度阈值），主要描述知识结构的宏观状态，虽然也能借助多个静态宏观结构（例如不同阈值层的聚类结构）人工比较分析，但难以自动对不同聚类结构的差异、性质及相应的演变可能进行探测，因此还难以有效支持对知识结构演化的深层次自动分析以及对潜在演变趋势的探测。这种静态结果分析往往只能揭示已经发生的知识结构演化，无法揭示可能发生的演化。

（3）即使有个别研究涉及到了对知识演化动态过程的分析，但揭示演化特征的能力不足。一个原因是，知识结构演化分析系统无法用定量的方法明确划分各种演化特征的类型，第二个原因是，无法在数量上辨明各种演变特征的区别。因此，演化过程中知识结构发生了哪一类型变化，不能被很好地揭示。

（4）不能在一个连续完整的应用流程中完成知识结构两种演化方式的分析。知识结构演化包括可能发生的演化和已经发生的演化，还没有一个集成应用平

台，既能分析知识结构中已经发生的演化，又能分析知识结构中可能发生的演化。

1.3 研究内容和意义

1.3.1 研究内容

目前知识演化的具体研究内容和表现形式有多种方式，本研究探讨的知识演化是知识按照生物进化模式，随着时间推移不断地进行新老交替的演化[1]，侧重研究整个科学体系或者某个学科领域知识的发展变化情况。变化情况包括量的增长和质的发展两方面。量的增长表征宏观层面的演化趋势，质的发展包括知识的转移、扩展、融合，以及新知识的产生、旧知识的消亡等等。这些变化通过知识单元本身内容的更替，以及知识之间关系的变化来体现。而知识关系的变化通过微观知识单元之间的相互联系来揭示。本研究不探讨宏观上量的增长，而重点关注点将置于微观知识单元上，探讨知识演化规律分析中知识关系所揭示的质的发展过程。另外，本研究不深入知识的内容，不考察知识的主题，仅仅从众多知识单元对外表现出的结构进行研究。

知识结构演化不仅体现知识的现状和发展脉络，更重要的是其间蕴涵了可能正处于萌芽阶段的知识，或者可能将会成为热点的知识。首先掌握这些情报有重要的意义。因此，知识结构演化的跟踪、监测、挖掘等方法与工具的研究、开发就成为迫切之需。

本论文将主要针对情报研究对知识结构演化分析第二个层次的需求来开展，利用关系阈值和时间两个特性，从潜在和演变两个角度，对知识结构演化的分析方法进行拓展，以期实现深度分析，补充现有不足，更好满足需求。在对探寻知识结构演化相关方法的分析基础上，本文的主要工作集中在阈值策略、时间策略的设计和实现。具体表现在以下几个方面：

（1）分析和设计阈值策略

知识结构演化分析方法通常是采用共词、共引等方法对科学文献进行聚类，生成主题簇结构图（常被称为主题地图或知识地图，其中主题簇聚集程度和主题簇间关系往往由特定阈值的共词或共引关联强度来决定），以此来反映科学领域的主题分布结构。这种方法能较好地揭示科学领域的宏观结构，但是当涉及较大领域或跨领域、涉及大规模数据时，为了清晰地表现主题簇结构及其关系，一般会采用较高阈值的共词或共引关联强度，但这些较高阈值下的结构表示往往容易忽略掉一些关系暂时不强烈的潜在对象及其可能存在的潜在结构，然而这些潜在的对象和关系结构可能正代表了处于萌芽阶段或者消亡阶段的潜在主题簇或主题簇间关系，因此单独较高阈值下的宏观结构可能掩盖了潜在的演化趋势。

当然，可以采用人工逐步调整阈值的方法来生成多个阈值层聚类结构，对这

些聚类结构之间的差异进行比较，以发现宏观结构下可能存在的潜在对象或潜在结构。但是，人工调整阈值生成不同聚类结构、以及人工比较不同阈值层结构差异，不仅效率低，而且难以高效地对不同阈值层结构之间的差异特征进行准确和细致的分析，难以根据多个差异特征共同的存在和程度来探测复杂的演变倾向。

针对这个问题，本文在基于共引聚类生成知识结构的基础上，提出不同阈值层聚类结构间自动差异检测的方法，并进一步分析差异特征及其所代表的（潜在）演变趋势，并通过可视化技术来表现被发现的潜在结构。这种方法称为阈值层结构差异检测策略（简称阈值策略），其目标是寻找潜藏在宏观聚类结构下的正孕育中、发展中的对象和关系。

需要强调的是，阈值策略不是研究如何调整、设置阈值来改善聚类结果，也不是通过改善聚类方法体现潜在结构，而是探讨通过不同阈值层聚类结构的比较来自动发现潜在结构的方法。

（2）分析和设计时间策略

知识结构演化的动态分析经常采用按照不同时间窗、按照同一阈值进行多次聚类生成主题地图的方法来形象地反映知识结构随时间的演变。但是，不同时间窗的聚类结构之间存在着复杂的关系，往往可揭示出聚类结构中不同主题簇随时间的生长、消亡、分化、融合等趋势，从而反映知识的扩散、融合，以及新知识的产生、旧知识的消亡等知识演化的质变。

当然，可以采用人工观察和比较的方法来鉴别不同时间窗间主题簇的关系以及它们的生长、消亡、分化、融合趋势，但显然人工比较效率低，而且难以高效地对不同时间窗主题簇关系进行准确和细致的分析，难以根据多个关系特征共同的存在和程度来探测复杂的演变倾向。

针对这个问题，同样利用共引聚类生成知识结构，本文提出对不同时间窗聚类结构间主题簇关系、关系性质、变化程度及其所代表的演变趋势的自动分析方法，并通过可视化技术来表现被发现的变化趋势。对不同时间窗聚类结构间主题簇关系的自动分析方法称为时间窗结构检测策略（简称时间策略），其目标是寻找在聚类结构时间流中的演变关系和演变过程。

（3）方法实现

基于阈值策略和时间策略的设计要求，根据系统框架的对两种策略的具体位置安排，依照系统整体处理流程，分别对两个策略中的主要方法设计详细的实现流程。运行系统，进行测试，对两种策略的可用性和有效性进行实验验证。最后将获得的演化结构组织起来，以可视化的形式直观展示给用户，从而最大程度降低用户的认知负担和分析负担，帮助用户理解知识结构演化的分析结果。

1.3.2 研究意义

知识演化分析方法和系统的研究还不全面，由于缺乏专门技术方法和应用工具的支持，影响了情报研究的服务能力和服务水平的提高，这也成为情报研究工作的一个薄弱环节。为此，本论文研究意义表现在以下三个方面：

（1）通过对国内外知识演化研究成果进行分析、总结，为以后本领域的研究工作提供基本素材，并强调知识演化分析对当前情报研究工作的重要性。同时，结合信息环境的发展趋势，分析情报研究工作对知识演化分析的新需求，尝试提出情报分析方法的发展方向，为知识演化分析系统的建设提供指导。

（2）通过对知识演化分析系统的系统需求、目标定位、功能组成等方面的分析，阐述有效的知识演化分析系统对深化情报研究服务、提升情报分析能力的重要性，进一步引起人们对相关技术研究的关注。

（3）通过对知识演化分析系统的系统模型、体系结构、技术支撑体系以及实现策略等方面的研究，尝试探讨当前技术环境下知识演化分析系统的建设方案，在系统设计、技术选择以及应用集成等方面为相关研究提供借鉴经验，推动实际应用系统的设计与开发。

1.4 研究方法

本研究的主要任务是情报分析方法的设计、系统的构建与实现，涉及对已有方法、系统的调研分析、对研究中所设计方法的可靠性验证，以及对系统运行及其所获结果的案例测试等。因此，本文针对研究的目的和研究过程中所产生的具体问题采用了如下研究方法：

（1）系统分析法。在本文方法设计之前，对国内外现有的情报分析软件、工具等的功能、适用范围、方法基础、技术手段等进行了广泛的调研分析，为本文方法和系统的构建提供思路、基础和参考。

（2）实验验证法。为保证提出方法的有效性，在涉及到相关算法的分析讨论时，利用实际数据对算法进行对比分析、剖析论证。同时也为方法的修正提供思路，为论文的科学性、合理性奠定基础。通过对系统的实际实现，从实践的角度证明本论文提出的概念模型、设计思路、运行机制是可行的，从而证明系统设计的合理性和可行性。

1.5 论文章节安排

本文在分析现状的基础上，提出本文的研究问题，进而搭建本文的整体研究框架，同时从系统建设的角度，设计了试验系统的整体体系结构，整体流程、数

据流、核心功能模块。在此基础上，对研究内容的主要部分开展具体研究，包括阈值策略（关系阈值的调整来发现、展现潜在结构的方法），时间策略（时间窗的划分来发现、展现演变结构的方法），系统实现（主模块及其子模块在系统中的实现，结果的展示）。最后对本文研究作总结和展望。

本论文共分为 7 章，研究的主体思路和论文组织结构安排如图 1-1 所示。

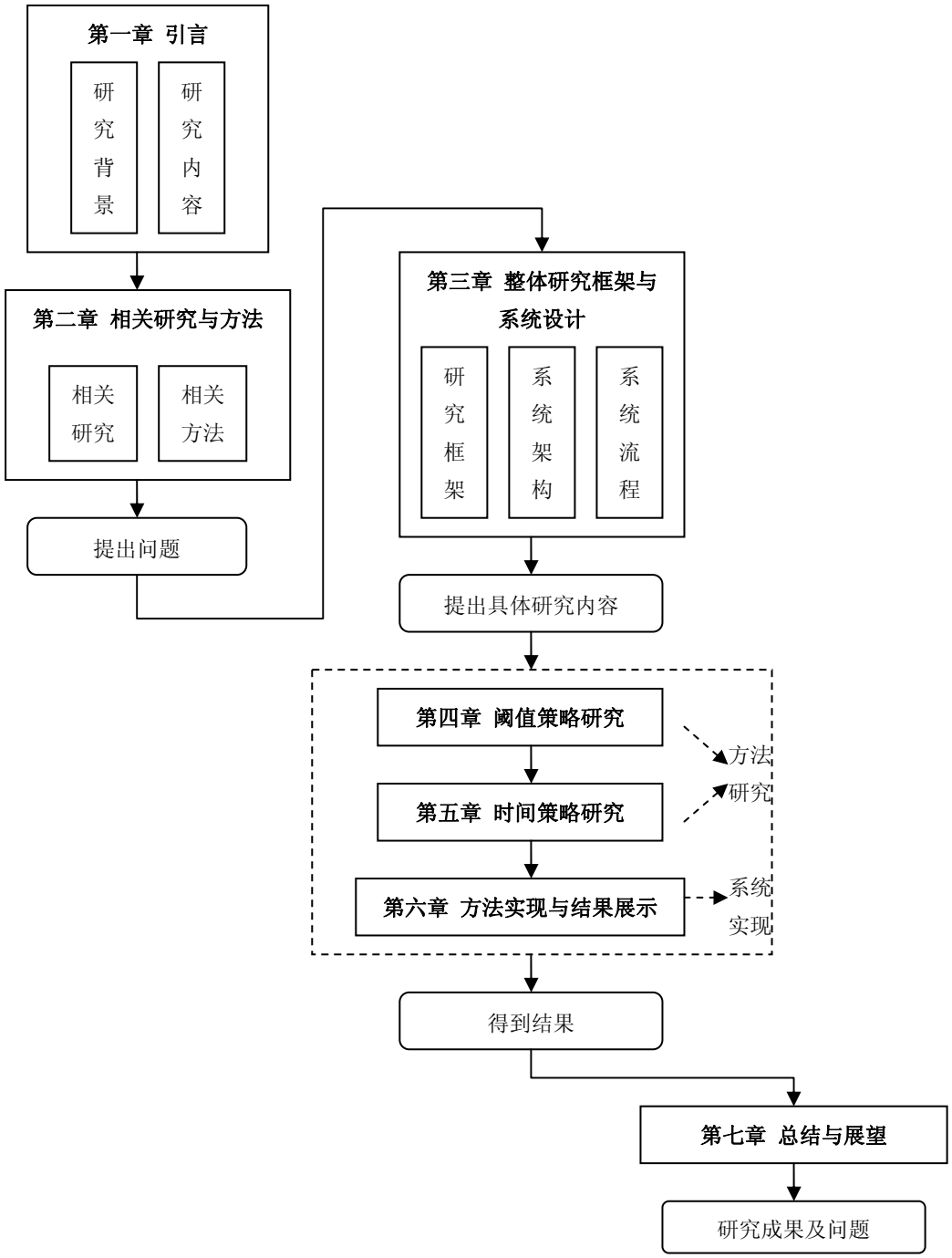


图 1-1 论文组织结构图

2 相关理论和研究方法

2.1 理论基础

2.1.1 科学知识进化

波普认为知识的发展同生物的进化非常相似，客观知识发展、进化是通过非自然的、非自发的或人为的选择进行的，具有遗传、继承、变异等特征。按照进化认识论的观点，知识是人的认识能力在自然界的长期发展中形成的。波普提出的客观知识世界[2]，即“信息世界”[3]，它的进化，是情报学借以分析科学知识结构演化规律的主要对象。从知识进化理论出发，通过阐明知识遗传、继承和变异的机制和规律性，可以为利用情报学方法揭示科学知识演化过程提供可靠的指导。

知识进化理论的基本思想是知识的创新发展体现为在遗传、继承、变异的复杂运动中实现新知识的出现、旧知识的淘汰。知识进化理论将科学知识作为遗传、继承与变异的基本单位，并假设知识一旦出现，就具有生命力并能代代遗传下去，新的知识通过继承已有的知识不断产生。同时，随着人们对客观事物认识的深化，知识也在不断地发生变异[4]。

从情报研究的角度，如果可以利用一定的方法总结知识的遗传、继承与变异的状况与机理，那么情报研究就可以揭示知识内部结构及其演化规律。在科学计量学领域，知识质的变化，由于其复杂性和方法的限制，还很难得到有效的揭示。

从实践上来讲，在知识进化理论的基础上，按照分析的目标分成可以获取和需要探究的知识单元，并通过特定的方法逐层反映出知识的进化情况，可以更深入揭示知识演化中质的发展过程，获得比较理想的分析效果。

2.1.2 潜在知识演化

学科领域知识可以被分为主体知识和潜在知识两个种类，主体知识是在科学研究中规模较大、相对稳定、被业界普遍接受、而且经常被引用的知识；潜在知识，相对于主体知识而言，通常是与领域有较高相关度但可能刚刚出现、或在业界存在争议讨论、或暂时不被经常引用的科学知识。

从Kuhn利用科学范式描述常规科学和科学革命的理论上看[5]，潜在知识极有可能演变成主体知识，它的演化可以认为是未来科学知识发展的主要方向，因而蕴涵非常丰富的情报内涵，需要利用科学知识结构演化的分析方法将其揭示。

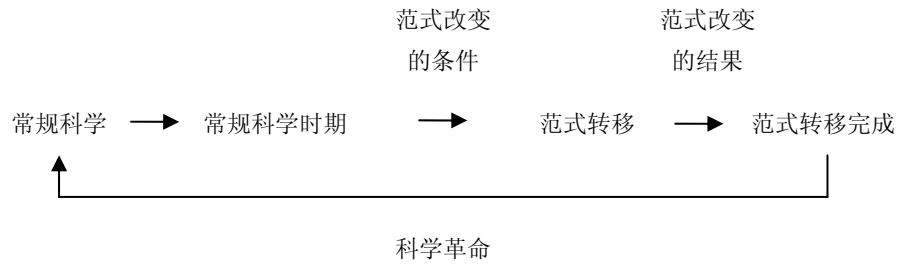


图 2-1 Kuhn 科学革命示意图

Kuhn阐述的科学革命结构，如图2-1所示。在科学革命的过程中每个阶段的知识结构可能包含了不同的演化特征，具体表现为：

（1）在常规时期，知识结构图稳定，可能有新词、新主题出现，但从整体上看对知识结构图的影响不大，结构图保持稳定。

（2）在危机时期，新知识处于萌芽阶段。此时，新知识因为自身内容不丰富、不完善，与其他知识的关联不强烈，因而往往位于结构图的边缘地带。考察边缘地带的知识变迁很有可能发现一个新的知识增长点。除了知识结构发生变化以外，代表知识的主题词在语义上也会有一定的变化。

（3）在演变时期，随着发展新知识进而出现在知识结构关键性的位置，强烈破坏原有的知识结构。具体表现可能有，新主题词大量出现，而且增速很大；词关系因为可能受到了其他学科的影响发生变化；某主题的文章发表数量发生异常变化，可能突增或者突减；代表一个新领域的新期刊可能会出现；结构图中关联较弱的“孤立主题簇”增多，表明新知识的刚刚出现，目前还处于发展阶段；结构图中某些主题簇的分裂或合并，表明主体知识在发生转移，扩散或者融合。

（4）在稳定时期，知识结构又趋于稳定，但是同常规时期相比有较明显的不同，可能产生一个新的大的主题簇，通过融合覆盖了原来的孤立主题簇，也可能是由原来的多个孤立主题簇合并而形成。

在某种意义上，科学革命的过程可以说是潜在知识从萌芽出生到发展壮大的过程。这个过程不仅带来潜在知识的演变，还引起主体知识的变化，但最终的结果还是通过词、语义、引文、期刊的变化，由知识结构布局的变化来体现。因此，可以通过对知识结构在不同粒度层间的变异（纵向变异）和在不同时间窗间的变异（横向变异）来检测和描述知识演化过程。如果我们共词、共引等对科学文献进行聚类，生成不同粒度（代表不同关联强度阈值）的主题结构布局，或者生成不同时间窗的主题结构布局，通过对这些结构之间的差异分析，就能检测可能的演变倾向。因此，从聚类结构差异分析入手，可以是探寻知识演化的有效切入点。

这样，从科学范式演变理论出发，通过对词汇、词汇语义、引文关系、期刊等的变化状态进行分析，可以对知识结构的演化进行深入分析。科学范式演变理论为揭示潜在知识演化提供了理论依据。

2.2 相关研究和方法

2.2.1 基本的研究和方法

2.2.1.1基于引文的方法

引文是知识交流和继承性的重要体现之一，具有明显的动态进化特征。基于引文对科学知识演化特征的分析，主要关注特定时间内科学前沿的跟踪分析，不同时间段内科学前沿的转变模式代表了学科的进化轨迹。

共引理论作为揭示科学结构的实用方法，在从不同层面揭示科学结构的过程中，应用范围在不断扩展，理论框架也在不断丰富[6][7][8][9][10][11][12][13][14][15][16]。比如，在揭示了单个学科结构的基础上，为了综合揭示自然科学和社会科学的结构，发展了多次聚类方法来反映不同层次学科间的关系，并通过比较连续年度的聚类图来分析学科间关系的动态变化情况[17]。另外，Small还利用共引聚类生成了多层次的科学结构图，即用一个总体图展示多学科的宽度，然后逐层下钻，直到文献层次[18]。

除了科学结构揭示，Small等人的研究还尝试用共引分析方法来展示知识结构，从而从知识结构中分析知识演化。用高被引的文献集合代表关键概念，它们之间的共引关系表示概念间的关系，这样共引聚类形成的共引簇便转化为由文献中包含的知识形成的知识库，在知识库中进一步进行知识搜寻可以导引新知识的发现。比如试图从共引网络中利用准最小遍历树和深度优先检索方法来合成专业描述。另外，1999年提出书目数据库中的知识发现，提供了一种依据强大的共引链接在科学文献间创建路径的方法，得到了一条从经济学到天体物理学的专门路径[19]。还有Garfield在1994年按时间序列生成一系列知识结构图，从横向描述知识的发展脉络[20]。

基于引文的揭示科学知识结构演化的方法中，除了共引，还有文献耦合。Morris 等人认为，引用了一组确定时间内特定文献的文献构成学科前沿，因此通过文献耦合聚类得出研究前沿及其动态变化状况，同样也能用于揭示科学知识结构演化[21]。有学者认为文献耦合更适合于研究前沿的揭示，而共引适合研究领域历史状况的揭示[22]。总的来说，文献耦合分析相比共引分析而言，理论和方法上都有待发展，在实践中的应用也不多，但是可以肯定的是它对演化特点的分析也是具有借鉴意义的。

2.2.1.2 基于词的方法

基于词的方法，最常使用的是共词分析。Callon、Law、Courtial、Bauin、Leydesdorff、Raan的研究代表了共词分析的主流[23][24][25][26][27][28][29][30]。从理论上讲，概念词可以依据其概念集合范围的大小分成不同层次，不同层次的概念词建

立的关系可以揭示不同层次的知识演化关系。词代表概念，是概念的具体表现，因此相比共引分析而言，共词分析揭示的是更微观的知识关系。

鉴于经典共词分析不能揭示全文中的丰富的语义关系，而且为了使任何格式的自由文本能在基于词的方法中得到利用，Kostoff等人发展了数据库内容结构分析方法[31]，即Database Tomography（DT）。DT中实现的自上而下的共现聚类的分类体系可以更加客观地揭示主题领域内的知识结构、研究层次以及科研活动的活跃程度，精确地确定主题领域的发展方向，短语的临近度分析有效揭示了推动领域内的科技关联，因此，DT可以在某种程度上揭示和挖掘特定领域内研究空白和知识创新点，是揭示知识演化关系、预测演化趋势的有效手段。然而，DT的局限是，强烈依赖专家的作用，虽然提高了准确性，但是易用性大大降低，操作起来复杂。

2.2.2 改进的研究和方法

2.2.2.1 发现潜在知识的研究和方法

Swanson在分析了知识结构宏观状况的三个表现，以及知识阵营的特性[32]的基础上，他将关注点聚焦在大量潜在的关联关系上，提出了非相关文献[33]的概念，并且通过实验证实了非相关关系的存在[34][35][36][37][38]。非相关关系说明，从引文或词的角度看似没有关系的知识，实际上存在间接关系，从而构成了潜在知识。这样的潜在知识是今后发展的一个可能方向，可能会使知识体系发生变化，导致知识结构的演化。

陈超美等人根据这种潜在性，认为知识发现和挖掘不应只停留在显著科学家及其多次被引论著的层面上，还应该多关注那些宏观引用模式中没有或者较少被引用的对象。相对于主体知识，潜在知识通常因为关系太弱被共引分析模式所忽略或者掩盖。陈超美对潜在知识的定义，如图2-2所示。

因此，陈超美提出了基于引文的潜在知识发现方法。图2-3描述了该方法的处理过程：首先按照共引方法确定由高被引文献形成的主体知识及其结构图。其次将领域从主体知识领域扩展到潜在知识领域。扩展过程的关键是从主体结构中基于知识结构和主题特征选择一个的“Exit”节点。这个“Exit”通过“Pull”（牵引出）显著相关但是相对引用率较低的文献到图中来，在跟踪潜在知识过程中起到桥接的作用。具体做法是以该“Exit”文献作为种子文献进行引文检索，找出直接引用和间接引用“Exit”的文献，形成新的结构图，从而发现潜在关联的知识[39]。

陈超美的研究特点是强调可视化技术在知识结构演化分析的应用[40][41][42]，此研究利用的是Pathfinder算法[43][44]。

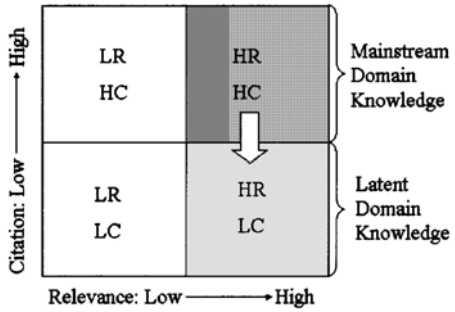


图 2-2 潜在知识定义的示意图[39]

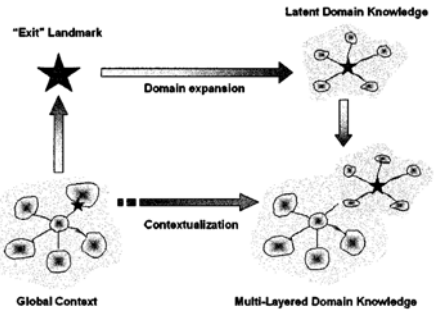


图 2-3 系统处理过程模型[39]

对于潜在领域知识的分析，还可以通过对比各种分析方法所得结果的不同，找到没有发现的联系和潜在的有价值链接。例如，如果共词分析揭示了一个相关知识链接，而该链接在引文网络中没有体现，或者仅仅体现了较弱的链接，那么该知识链接或许是一个比较重要的知识增长点。但是，两种方法结果的不同，可能是由方法的视角不同产生的，而且结果的比较多靠定性，非常复杂。

2.2.2.2 发现新生知识的研究和方法

对新生知识的发现研究，具有代表性的有Price和Crane利用回归曲线研究研究领域将来发展的程度[45][46]，Meadows寻找科学发展增长点的领先指标[47]，Goffman设计一个模型来预测短时间内领域的增长[48][49]，Tabah尝试使用混沌理论为不同领域文献的增长建模[50]，Garfield用historiograph研究领域文献的变化情况[51]，还有Leydesdorff提出用图论和成分分析方法研究社会科学引文索引的由上至下的分解模型[52]。

最近，Leydesdorff研究了在科学交流体系中知识的产生问题[53]。Leydesdorff认为新生的知识首先通过文献的流通来传播[54][55]。不断地创作和发表科技论文，科学家发现的新现象、新规律得以被他人认识。通过期刊交流，科学家丰富新事物的含义，增强对新事物的理解。至此，新事物得到很多科学家的认识，变成一个“有含义的信息体”。由于期刊的引用和被引用的关系，位于不同期刊的信息体在语义上发生很大的变化。从信息体挑出有具体含义的词汇，形成一个包含语义的关系结构图，此时一个新知识产生了[56][57]。进一步，知识历经考验，它的引用关系不断变化，当被普遍认同、广泛接受的时候，知识就演变成下一个知识发展的基础。至此，一个知识通过科研交流最终形成，如图2-4所示。值得注意

的是，应该更多关注非经典、很少被引用的期刊，位于这样期刊的知识也许正是一个研究前沿，蕴涵着剧烈的演变可能[58]。

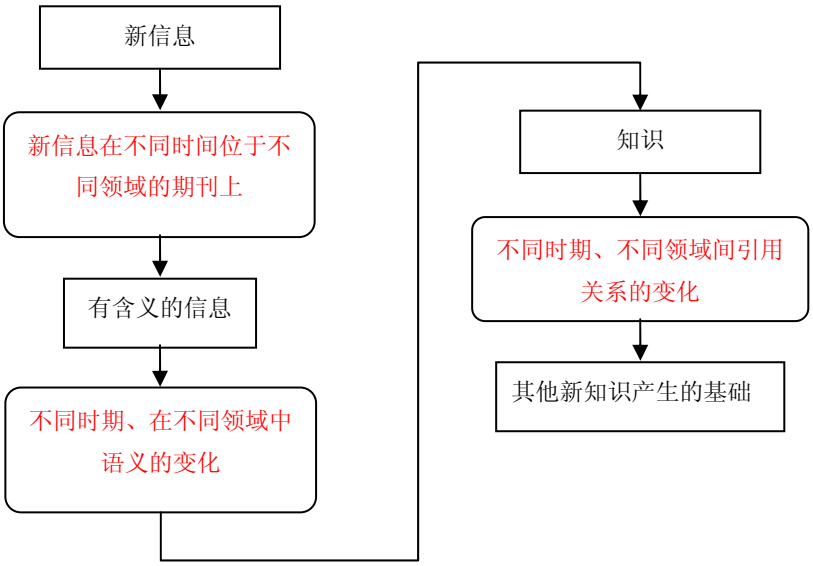


图 2-4 基于科学交流，由新信息到知识出现的演变过程

综合而言，新知识的出现可能伴随有期刊空间的变化，如新期刊的出现，期刊名的改变，期刊间引用关系的变化，发文数量的变化。同时还可能有主题空间的变化，如新词的出现，词与词关系的变化，语义表示的变化，学科体系的变化。新知识在知识结构中会现扰乱现有的知识关系网，改变科学家的交流用语（词及词的含义在变化）。经过一段时间，新知识在现有的知识关系网中被广泛传播，逐渐被接受。最后，新知识形成了自己的知识体系，并最为科学知识库促进科学的发展和更多新的发现。如图2-5所示。

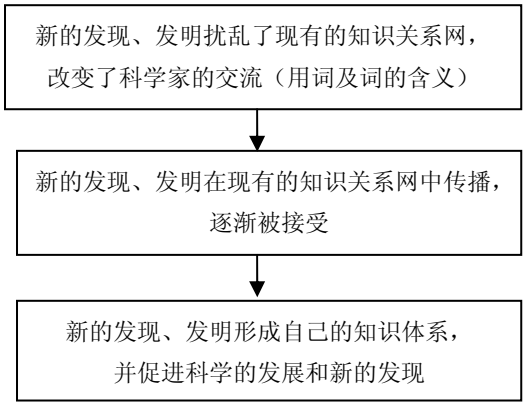


图 2-5 新知识在知识结构中的演化

值得一提的是，Small基于旧簇的消亡、新簇的出现、多个簇的交叉演变等过程，对多个时间窗的共引文献聚类结果进行深入的分析，揭示了科学知识演化的特点[59]。簇演变链（Cluster Strings）的线性模式和非线性模式，如图2-6所示。以3个时间窗，2或3个簇构成的演变链为例，根分别计算线性模式和非线性模式

占所有演变模式的比例。通过衡量非线性模式所占比例，分析知识结构的演变情况。通常情况下，非线性模式越多，知识结构演变越剧烈，新知识产生的可能性越大。

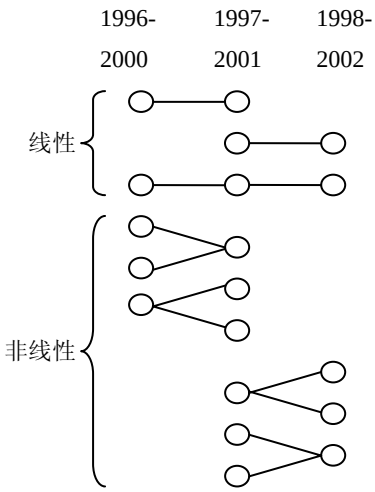


图 2-6 簇演变链的线性模式和非线性模式[59]

2.2.2.3 监测知识突发的研究和方法

监测知识突发的研究和方法，通常指的是基于主题突发监测的引文分析，首先识别突然引起关注的主题，然后将这些主题词同主体知识中高影响力的文献建立关系，并在时间维度上建立时间关系。其中突然引起关注的主题是通过突发监测算法（Burst Detection）得到的。突发监测算法源于用于监测电子邮件和文献题名中单词的变化而设计的。该算法建立在话题监测与跟踪[60][61][62][63][64]、文本挖掘[65][66][67][68]、可视化[69][70][71]等领域的相关研究之上，有很大的相似处。其主要思想是每当一个重要的事件发生或者将要发生的时候，应该有特定词汇的迅速增加标志着事件的发生，这样的迅速增加称为突变[72]。本质上来说，突变监测算法分析的是词频的增长率，并识别快速增长的词汇。

在一个给定的时间窗内，由引用单词和被引文献组成相关矩阵。在该矩阵中词与词之间的关系是由在引用文献中一对词的共现来确定的。文献与文献的关系是由共引关系定义的。词与文献之间的关系表明，词在引用文献中出现，并且该词在分析的时间窗内使用频率迅速增长[73]。该方法建立了一个基于热点词的研究前沿同基于引文的背景知识之间的链接，使得引用文献和被引文献所揭示的主题信息得以在一个统一的框架下被研究和表示。用此方法既可以监测一定时间段内主题的变化，又可以揭示背景知识同研究前沿之间的知识演化关系。

2.2.2.4 分析知识转移的研究和方法

转移涉及到两种知识，可能都属于科学知识或者技术知识，可能一个属于科学知识，一个属于技术知识。分析知识转移的研究和方法，最简单的思路是分别在两个知识体系上做出各自的共词结构图，进而进行定性比较分析，期望从图中

找到两个知识体系的相关部分，这个部分表明两个知识体系有一定的关系，其中的知识可能是从一方转移到另一方[74]。如图2-7所示。

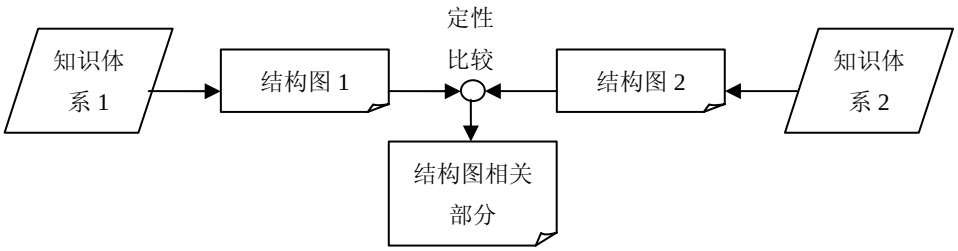


图 2-7 基于结构图比较的知识转移分析方法

词关联方法较上一个方法复杂，简单地说就是用两个知识体系中相同或相近的词的出现频次来作为相近度的测度指标，通过向量相似度将两个知识体系联系起来，来反映两个知识体系之间的关联程度。主要思路是：先利用文本挖掘的方法对专利和文献进行抽词和词频分析，然后利用词作为纽带进行专利和文献著者的关系揭示。以一个研究机构为分析对象，首先在一个时间段内抽取专利和文献中都出现的关键词，形成一个领域词表，基于向量的相似计算，将每个专利在文献数据库里的查询计算得分，得分较高的文献被认为是相关的。最后，对每项专利发明人的特征同相关的文献著者相联系，进行特征分析[75]。基于向量相似度比较的知识转移分析方法如图2-8所示。然而，由于专利的技术用语与文献中的科学用语存在语法、语义不一致的问题，该方法可靠性不足，必须事先在词的概念上进行统一。

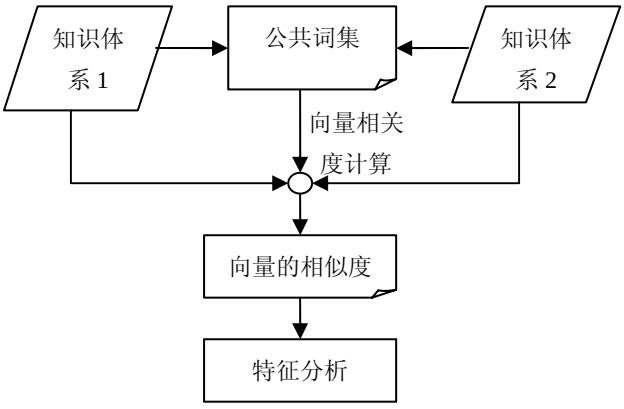


图 2-8 基于向量相似度比较的知识转移分析方法

除了上面用词构建两个知识体系之间联系的方法外，还有研究不用词而用引用关系通过期刊及其发文量研究知识在两个学科之间的转移[76]。知识在两个学科之间的转移已开展了很多研究，见参考文献[77]，[78]，[79]，[80]，[81]，[82]。该研究的特点是在两个基本计算公式

a、学科 i 引用学科 j 的数量 $R_{i,j}$，那么学科 j 占的份额

$$\gamma_{i,j}=\frac{R_{i,j}}{R_i}=\frac{R_{i,j}}{\sum_j R_{i,j}}$$

b、考虑学科间发文数量的差异，改进份额

$$RE_{i,j}=\gamma_{i,j}(\alpha_i)(1-\alpha_j)$$

的基础上，提出了两个核心计算公式

$$RE_j=\sum_{i\neq j}\gamma_{i,j}(\alpha_i)\left(\frac{1}{\alpha_j}\right) \qquad IER_j=\frac{C_j-R_{j,j}}{R_i-R_{i,i}}$$

这两个核心计算公式用以度量一个领域对另一个领域的影响强度。

另外，在科学与技术之间的知识转移方面，传统的专利引文分析很大程度上是在统计分析层次上的，如通过比较不同领域专利文献与期刊文献的引用数量关系，反映领域技术对科学的依赖程度，再如通过统计分析被引期刊文献发表时间与专利的申请时间，得出它们之间的时间差异等[83]。陈超美尝试综合使用复杂网络结构统计方法、可视化及引文分析的方法揭示科学与技术之间的知识转移关系[84]，即首先通过统计的方法揭示关联程度，再对科学文献进行共引可视化，找出连接两个聚类的中介点，然后对其被专利引用的情况进行定性分析。

为了避免专利引文带来的困难，Collon 将科学文献划分为两个集合，一个为侧重科学研究的集合，另一个为综合科学研究与技术开发的集合，对两个文献集合分别做共词聚类，并对聚类结果进行比较及相关性分析，以更清晰地揭示科学与技术之间的关联关系[85]。聚类结果的比较包括关联分析和动态分析，关联分析是将不同的共词聚类，通过簇中相同词的个数来计算相似度，以揭示科学与技术对应领域的关联程度；动态分析是建立簇变化的时间序列，跟踪中心度与密度指标的变化，内容的发展变化，以及在一定时间内比较稳定的簇。Collon 基于共词聚类揭示的科学与技术之间的关系，在某种程度上可以反映知识转移特点，但是这里网络比较还是更多地依赖定性分析，复杂度较高，因此，该方法的使用受到了限制。

2.2.2.5 识别结构图重要节点的研究和方法

识别重要节点是知识结构图分析的一个重要内容，也是两个知识结构图进行比较研究的一个前提条件。按照Otte和Rousseau的观点，此时的关注点落在结构图中的单个节点上，因此有必要借用社会网络分析[86]的若干方法，而不是科学计量的方法了[87]。因此，情报分析技术的研究引入了社会知识网络分析中“中心度”的概念，作为度量结构图中节点是否重要的标准。自七十年代末期Freeman提出第一个“Betweenness Centrality”的概念[88][89]以来，“中心度”的计算方式至今以总共发展到4个，并且还有学者在不断研究[90][91]。其中，Betweenness Centrality受到的关注较多，很多学者对它进行深入研究[92][93]。Leydesdorff利用

Betweenness Centrality计算期刊的中心度，识别期刊是否是交叉学科期刊[94]。Katarina Larsen则通过中心度的计算和比较，度量衔接节点的重要度[95]。

具有代表性的研究是陈超美等人在其分析系统Citespace II中所做的工作，Citespace不仅可以帮助用户识别引发概念变革的文献[96]，还可以识别重要枢纽节点来揭示邻接网络之间的转化状况[58]。Citespace认为，两个时间上的两个研究前沿通过共引聚类形成两个簇群。当两个簇群在一张图上显示时，它们中间会有一个或多个点衔接着这两个簇。此时这些点就是重要的枢纽节点。对于枢纽节点的识别，Citespace采用Betweenness Centrality指标来度量：节点的成员数达到一定要求，Betweenness Centrality足够大，衔接的是两个知识前沿，那么这个节点就是重要枢纽节点。另外，图2-9显示了Citespace对一些重要节点的处理。

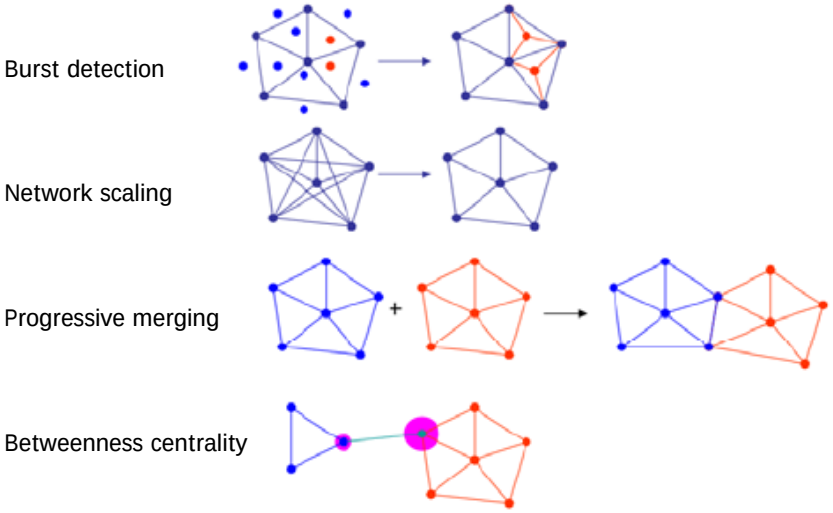


图 2-9 重要节点的处理方式[97]

2.2.2.6 展示演化结构的研究和方法

目前主流的可视化技术只能揭示一种特征项的关联结构。而很多情况下，不仅需要揭示两种特征项的关联结构，而且需要将两种特征项关联起来，获得交叉关联关系。为了揭示两种特征项之间的关联，Morris 借助于两个异共现矩阵相同特征项之间的关联，开发了时间线技术并进行了应用研究[98][99][100]，可以很好地弥补目前可视化技术不能揭示两种特征项关联的缺陷。

时间线分别用两个轴代表两个不同的维度，时间线技术用 x 轴代表时间轴，y 轴代表其它特征项，例如主题。研究主题之间的聚类结构用 y 轴可以揭示出来，每个研究主题包括一组论文簇，每簇论文给出名称，最后将属于每一个研究主题的论文与出版年关联起来[101]。时间线技术生成的时间线图可以直接观测研究主题的发展信息，揭示学科研究内容变化的情况[102]。

另外，陈超美等人对 minimum spanning trees(MSTs) 和 Pathfinder networks(PFNETs)两种可视化技术进行了比较研究。它们是广泛用于的结构图裁

减的算法。比较内容包括裁减能力的比较，体现结构演化能力的比较。通过比较，最后得出结论：虽然MSTs的执行效率要高于PFNETs，但是PFNETs在保持演变动态性方面效果更佳[103]。也正是这个原因，陈超美研究的系统CiteSpace采用了PFNETs算法。

2.3 存在的问题和不足

从上面的分析可以看到，针对情报研究对知识结构演化分析第二个层次的需求，对于可能发生的演变，即探寻潜在知识方面，陈超美关注的是与主体知识相关但被较少引用的知识，因而他提出的潜在知识是由引用高被引文献但自身较少被引用的文献构成。而本文认为，利用共引方法对高被引文献进行聚类分析生成主题结构图（代表了主体知识），这个过程本身就可能掩盖了某些高被引文献所代表的潜在知识。一些高被引文献虽然被较多引用和关注，但是可能由于发表时间不长，也可能由于阐述了一个新出现的知识而未与知识体系建立足够强的关联，导致与其他高被引文献的共引关系强度太低，因而在特定的关联阈值下就无法在主题结构图中体现。然而这样的高被引文献所代表的潜在知识比那些较少被引用的文献更有可能在未来获得重视，得到发展，发现这样的潜在知识具有更丰富的情报内涵。现有的方法还无法发现这样的潜在知识。

在分析知识随时间演变过程方面，Small虽然利用了簇演变链，研究对象降低到知识单元集（聚类簇）层面，但是采用的方法仍旧是在整体上通过对线性模式和非线性模式比例大小的判断来衡量演变的发生。而且，在复杂多样的非线性演变模式中没有明确指明多种演变类型的特征和区别，没有给出度量这些不同演变类型的计算方法，因而无法明确判断哪些知识发生了什么类型的演变。

总的来说，文献计量学、科学计量学对科学技术发展状况进行解读已经得到了较大的发展，但受研究方法、技术和工具等方面因素的限制，还有所欠缺。

（1）通常情况下，聚类过程作为研究知识结构演化的起点采用固定的、较高的簇内成员关系阈值。这种方式可以从数据全集中挑选出主要的对象和关系，反映出由它们形成的主体结构。再一方面，采用固定的、较高的阈值可以减少数据量，提高运算效率。

然而，选用固定的、较高的阈值进行聚类所形成的主体知识结构，往往容易忽略掉关联强度较低的相互关系，以及由较低关系联系在一起的对象，导致掩盖了潜在的有意义的结构和关联，因此，固定的较高的阈值下的知识结构可能掩盖了潜在的演化趋势。

正如前面指出，可以人工逐步调整阈值生成多个阈值层聚类结构，对这些聚类结构之间的差异进行比较，以发现可能存在的潜在对象或潜在结构。但是，人工调整阈值生成不同聚类结构、以及人工比较不同阈值层结构差异，不仅效率低，

而且难以高效地对不同阈值层结构之间的差异特征进行准确和细致的分析，难以根据多个差异特征共同的存在和程度来探测复杂的演变倾向。

（2）在时间演变方面，可以按照不同时间窗进行多次聚类生成主题地图的方法来形象地反映知识结构随时间的演变。可以采用自动的方法来鉴别不同时间窗间主题簇的总体继承与分化比例（如 Small 的研究），也可以分别比较分析具体主题簇间的生长、消亡、分化、融合趋势，然而前者虽然能揭示总体的演变“强烈程度”，但不能揭示具体的演变关系，后者尚不能高效地对不同时间窗主题簇关系进行准确和细致的分析，难以根据多个关系特征共同的存在和程度来探测复杂的演变倾向。

2.4 本文的研究思路

本文拟在共引聚类生成知识结构的基础上，提出不同阈值层聚类结构间差异性自动检测分析、自动发现宏观结构下的潜在结构的方法，提出不同时间窗聚类结构主题簇传承关系自动检测分析、自动发现不同时间下聚类结构的演变趋势的方法，并通过一个以这两种检测机制为核心模块的知识结构演化分析试验系统，对这两种方法进行初步验证。不同阈值层聚类结构间差异性自动检测方法称为阈值层潜在结构检测策略（简称阈值策略），不同时间窗聚类结构间主题簇传承关系自动检测分析方法称为时间窗演变结构检测策略（简称时间策略）。阈值策略的目标是寻找潜藏在宏观聚类结构下的孕育中、发展中的对象和关系，时间策略的目标是寻找在时间流中聚类结构包含的演变关系和演变过程。

阈值策略着重分析不同阈值层聚类结构中聚类簇的相互关系，定义同一阈值层中簇与簇的 relative 关系和不同阈值层之间的 twin 关系，划分并定义 5 个种类的潜在簇。进一步，设计阈值策略实施的步骤，包括突变阈值区间的确定、两个阈值层 relative 关系和 twin 关系的构建、5 种潜在簇的寻找、潜在关系的确定，最后是潜在结构的可视化展示。

时间策略着重分析不同时间窗聚类结构中聚类簇的相互关系，构建时间窗间簇与簇的 ancestor 关系和 offspring 关系，分析时间窗间的“遗传继承”性。利用 ancestor 关系和 offspring 关系，划分并定义 6 个演变簇划。设计时间策略实施的步骤，包括时间窗间 ancestor 关系和 offspring 关系的构建、6 种演变簇的寻找、演变关系的确定，最后是潜在结构的可视化展示。

3 整体研究框架和系统设计

根据前面确定的研究问题，本章的主要内容是，设计本文研究的整体框架，限定了本文的研究内容。从系统建设的角度，为试验系统设计系统整体体系结构，系统整体流程、系统数据流、系统的核心功能模块。

3.1 研究框架

本文研究的主要目的是对知识结构演化进行深度分析。从前面总结已有工作的不足，本文将从三个方面开展研究。

（1）针对知识结构演化的潜在性，本文将开展阈值策略研究，具体包括 4 个方面的研究工作：

a、不同阈值对聚类结果的影响研究

在不同阈值情况下将出现不同粒度且存在主题簇间差异的聚类结构，本部分将对这个现象进行验证，揭示潜在结构的可能特点。

b、潜在簇种类、潜在关系研究

归纳簇与簇的两种关系，为发现潜在结构打下基础。另外，潜在结构有不同的表现，各有特点，这是由潜在簇的多样性造成的，所以进一步对潜在簇作详细的定义，并对潜在簇的种类进行划分。

c、潜在结构寻找方法研究

首先探讨鉴别差异性明显的不同阈值层的方法，从大量的可能阈值层中确定部分相互差异明显的阈值层，提高在差异明显的两个阈值层中发现潜在结构的可能性。在此基础上，设计寻找潜在簇及其潜在关系的方法和步骤。

d、潜在结构可视化展示方法研究

由于差异是否明显的分析会产生两种不同的结果，所以本部分为两种结果设计两套展示方法。

（2）针对知识结构演化的演变性，本文将开展时间策略研究，具体包括 4 个方面的研究工作：

a、不同时间窗聚类结果的关系研究

对不同时间窗聚类结果的关系进行分析。使用不同的时间窗进行聚类，不同时间窗的聚类结果有何关系，是否隐藏着演变关系，是本部分的主要研究内容。

b、演变簇种类、演变关系研究

本部分研究内容将归纳簇与簇在不同时间窗上的两种关系，这两种关系是发现潜在结构的基础。接着针对多种演变方式，研究簇的演变特性，并结合特性对

各种演变簇作详细定义。每一种演变簇都对应一种典型的演变方式，这样，发现这些演变簇就能发现以这个簇为核心的演变结构。

c、演变结构寻找方法研究

时间窗聚类结果不存在差异性是否明显的问题，因此本部分研究的关注点是设计寻找演变簇和演变簇关系的方法和步骤。

d、演变结构可视化展示方法研究

从对演变结构理解的角度，本部分设计两种的展示方式，从不同的侧重面体现演变结构。

（3）针对前面设计的阈值策略和时间策略，本部分的研究内容包括 2 个方面：

a、设计两种策略的详细实现流程

根据设计策略时的模块划分，提取出两种策略的主要功能模块和子功能模块，依次为各个模块设计主要的方法函数和详细的系统实现流程，为系统建设打下基础。

b、考察两种策略的实施情况

通过系统运行，两种策略付诸实施。从得到的结果数据和展示图中考察两种策略的实际运行效果。

本文的整体研究框架如图 3-1 所示。

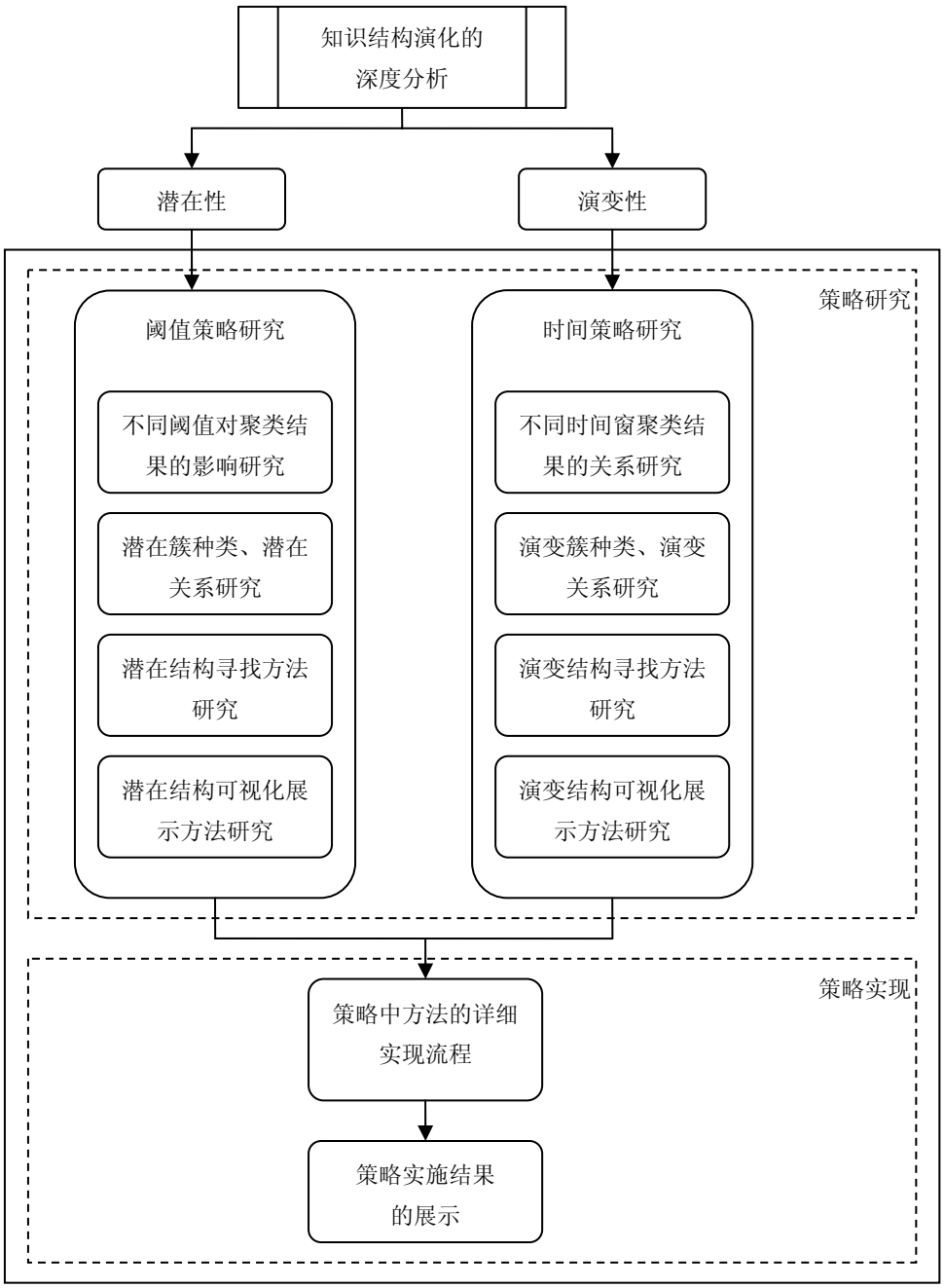


图 3-1 研究框架

3.2 系统设计

本研究将建立完整的系统流程来实现两种策略分析，以测试分析算法和过程的可用性和可靠性。

3.2.1 系统体系结构

系统整体体系结构如图 3-2 所示。系统从低到高共分为 5 个层次。数据层包

括待分析数据库，包括 ESI、SCI 两种结构化文献库。接口层分为两种，一种是用户接口，供用户设置系统运行参数和选择策略，另一种是数据接口，负责根据所选策略从数据库中提取、清洗数据。第三层是预处理层，负责共引聚类的实施，为阈值策略循环聚类调整阈值，为时间策略循环聚类划分时间窗。再上一层是功能层，也是本系统的核心层，负责两种策略核心功能的实现，即发现潜在结构或演变结构。最上一层是结果层，利用可视化方法展示得到的结果。

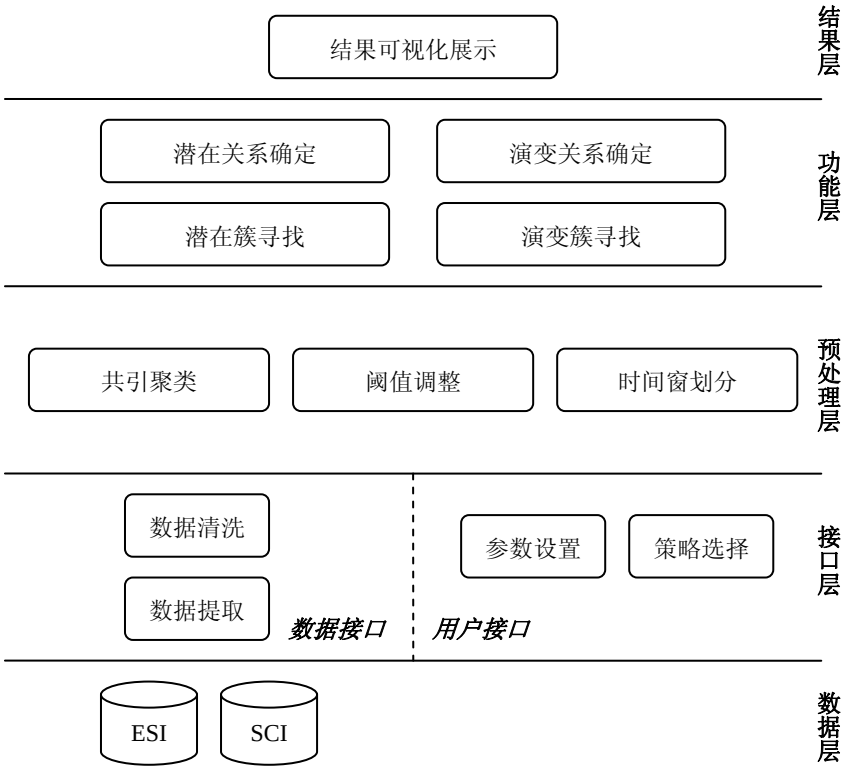


图 3-2 系统体系结构

3.2.2 系统运行流程

系统整体运行流程如图 3-3 所示。运行流程图展示了从数据集的选择与构建，到阈值策略的实施，再到最终展示给用户的结构图的全部处理过程，包括实现具体功能的各个子模块以及生成的中间数据。从整体上看，流程分为三个阶段。首先是准备阶段，其任务是选择策略、构建数据集、实施聚类分析。第二阶段是发现阶段，其任务是利用两种策略中的具体方法，来发现潜在结构和演变结构。最后一个阶段是展示阶段，其任务是用可视化方法展示结果。三个阶段中处理过程的描述，将在后续章节中详细介绍。

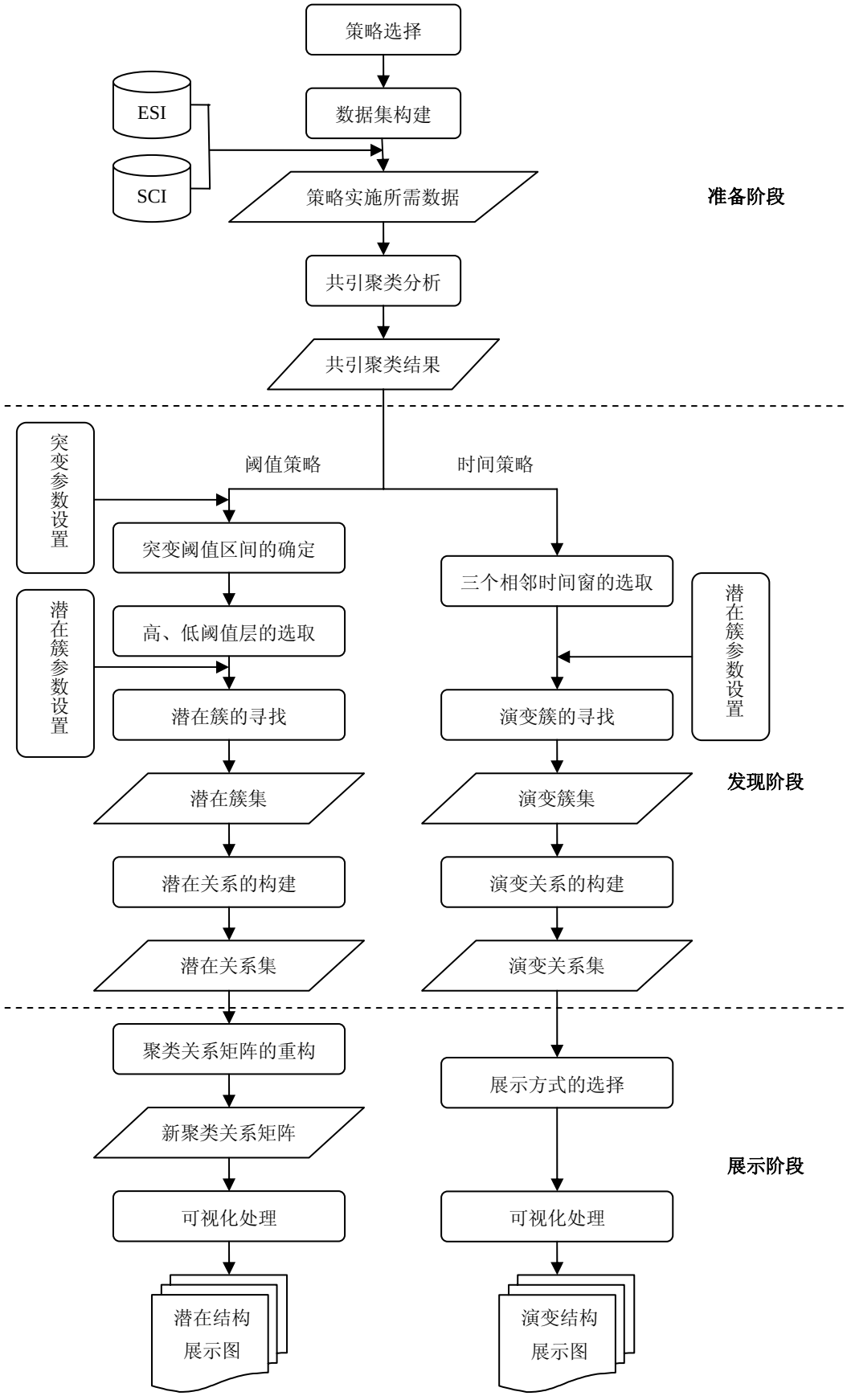

策略选择
ESI
数据集构建
SCI
策略实施所需数据
准备阶段
共引聚类分析
共引聚类结果
突变参数设置
阈值策略
时间策略
突变阈值区间的确定
三个相邻时间窗的选取
潜在簇参数设置
高、低阈值层的选取
潜在簇参数设置
潜在簇的寻找
演变簇的寻找
发现阶段
潜在簇集
演变簇集
潜在关系的构建
演变关系的构建
潜在关系集
演变关系集
聚类关系矩阵的重构
展示方式的选择
新聚类关系矩阵
展示阶段
可视化处理
可视化处理
潜在结构展示图
演变结构展示图

图 3-3 系统运行流程

3.2.3 系统数据流

系统整体数据流如图 3-4 所示。首先从数据库中获取待分析数据，在一条主线上形成策略所需的待分析数据集，经过聚类计算形成聚类结果。到此时，数据流将选择流向，根据何种策略选择进入哪个策略的处理流程。最后进入展示阶段，此时也会产生不同的展示结果。但是不论哪种策略、哪种流向，中间数据都将保存在结果数据库中。

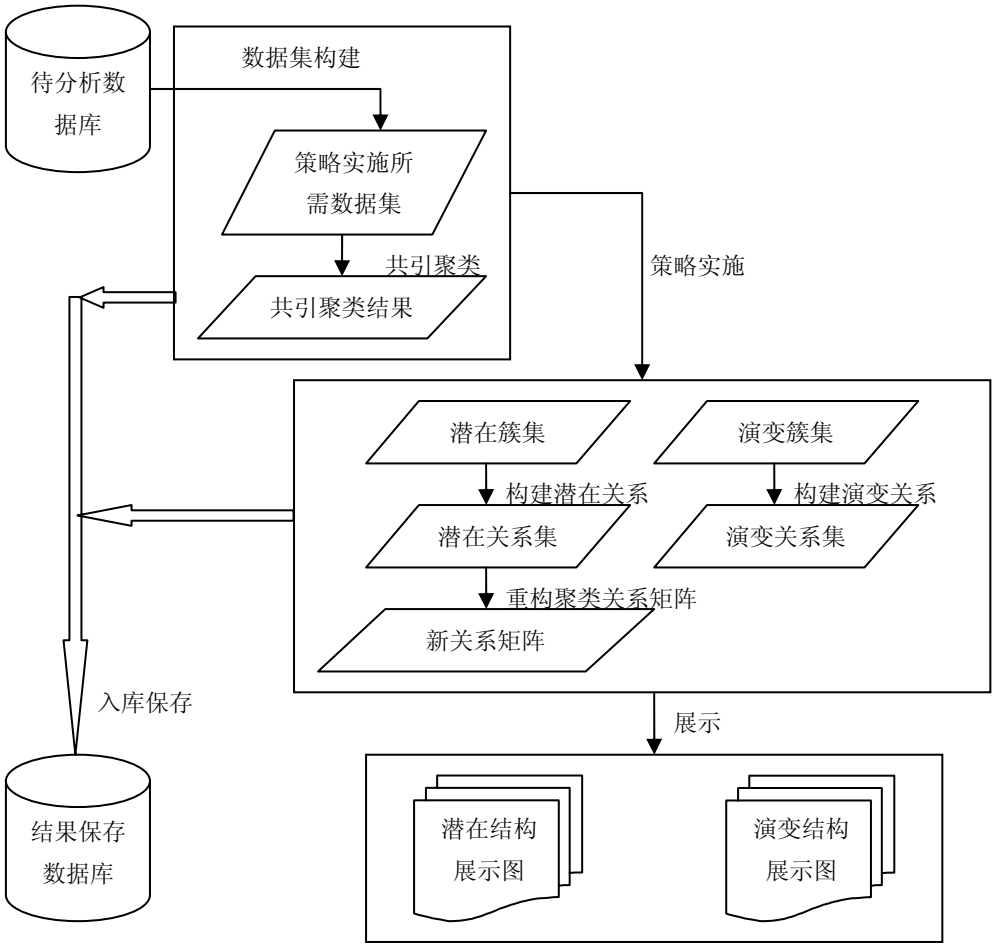


图 3-4 系统数据流

3.2.4 系统核心功能模块

系统核心部分的主要功能模块如图 3-5 所示。系统核心部分包括阈值策略模块和时间策略模块。前者较后者复杂，因为前者需要确定突变阈值区间、重构关系矩阵。但整体而言，两者的处理流程是相似的。

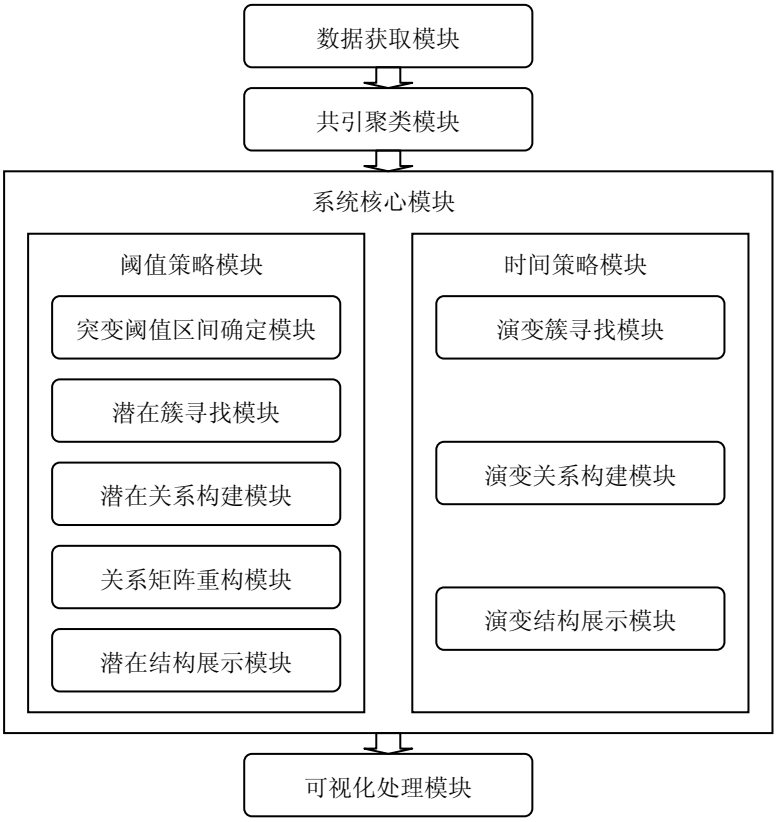


图 3-5 系统核心部分的主要功能模块

3.3 小结

本章节以提出的研究问题出发，设计了本文研究的整体框架，规定了本文的研究内容。还从建设系统的角度，设计了系统整体体系结构，系统整体流程、系统数据流、系统的核心功能模块。

4 阈值策略研究

在揭示知识结构的时候，通常采用聚类的方法对一个数据集中相互有关系的对象进行分析。聚合而成的簇体现了簇内成员之间的具有较强的相互关联，簇与簇之间的关系则揭示了数据集中蕴涵的知识结构。

本章将研究在多个阈值层聚类结果中自动发现、展示潜在结构的方法。在设计研究路线的基础上，先讨论关系阈值的调整对聚类结果的影响，明确阈值调整可以用于潜在结构的发现。紧接着具体研究阈值调整发现潜在结构的方法，主要内容包括潜在簇的种类划分、潜在结构的发现、潜在结构的展示。

4.1 研究线路

本文阈值策略所谈到的阈值，特指聚类算法中能参与聚类的 ESI 文献间共引关系的最小值。不论是共引的相似关系还是距离的相异关系，不难理解，以前一种关系制定的阈值策略同样适用于后一种方式。

阈值策略的研究线路如图 4-1 所示。聚类在前期已有的工作上实现，因此图种灰色部分是本章的主要研究内容。

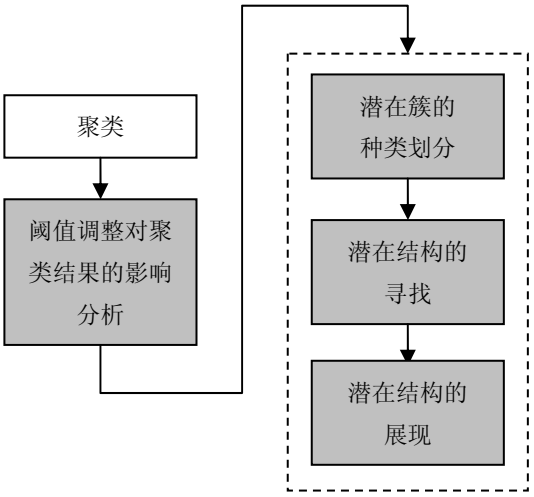


图 4-1 阈值策略研究线路

4.2 阈值调整对聚类结果的影响

本部分内容将重点讨论阈值的调整对聚类结果的影响，以及不同阈值产生的聚类结果能否用于发现潜在结构。另外，结合实际，讨论阈值调整的同时，其他参数一同变动，此时的阈值策略能否用于发现潜在结构。

4.2.1 阈值的影响分析

两个阈值层面的聚类结果之间原本没有相互关联，彼此相互独立。然而，当两个阈值层面的簇存在某种关系时，这两个阈值层面也就建立起了相互关系。

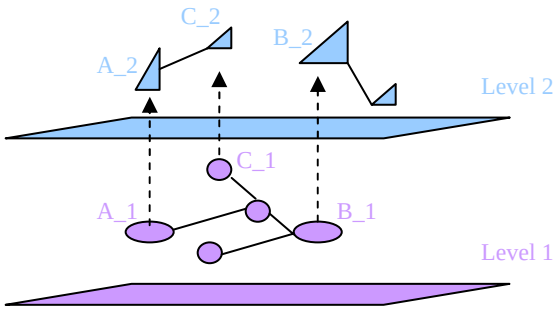


图 4-2 两阈值层聚类结果的差异与关联

图 4-2 所示，两个阈值层面的聚类结果从外在表现看差异很大，似乎没有任何关联。然而，当观察簇内成员时，可以发现：（1）第一层的簇 A_1 与第二层的簇 A_2 非常相似，A_1 中 83.1%成员与 A_2 中 95%成员是相同的。（2）第一层的簇 B_1 包含第二层的簇 B_2，即 B_2 中的成员在 B_1 中完全出现。（3）第一层的簇 C_1 与第二层的簇 C_2 等价，即 C_1 与 C_2 成员完全相同。

由此可见，看似毫无关系的、各自独立的两个阈值层次的聚类结果，在簇内成员上有非常大交融，或者说某阈值层上的簇在另一阈值层能找到它的“影子”。既然两个阈值层可以通过簇间的“影子”关系间接建立起一定的联系，那么对看似没有关系的多阈值层聚类结果进行比较分析来发现某种特殊的结构是可能的。

要能发现潜在结构，两个阈值层面聚类结果的“影子”关系应该在数量上进行度量，并进行必要的比较分析，归纳出阈值变化带来聚类结果变化的规律性。以 ESI 中的 COMPUTER 领域的数据为例，采用共引聚类方法，在其他参数相同的情况下，最高阈值设定为 0.4，阈值降低步长设定为 0.02。这样将最高阈值聚类结果分别与逐步降低的低阈值聚类结果进行 20 次比较。

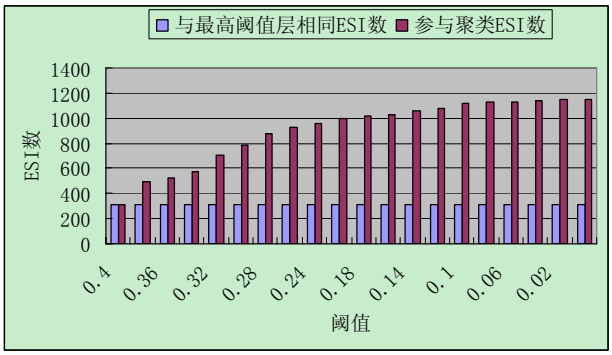


图 4-3 相同 ESI 与参与聚类 ESI 的比较

观察图 4-3 可以看到，随着阈值的降低，参与聚类的 ESI 文献篇数单调递增。然而，每一个低阈值层都与最高阈值层有相同的 ESI 文献，而且基本上就是最高

阈值层参与聚类的 ESI 文献。这些现象表明，阈值的降低使参与聚类的 ESI 文献的范围在逐渐扩大，而且低阈值层参与聚类的 ESI 文献包含高阈值层参与聚类的 ESI 文献，在高阈值层的基础上逐渐增加。

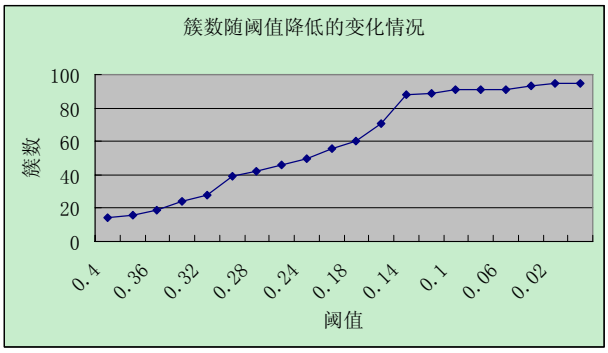


图 4-4 簇数随阈值降低的变化情况

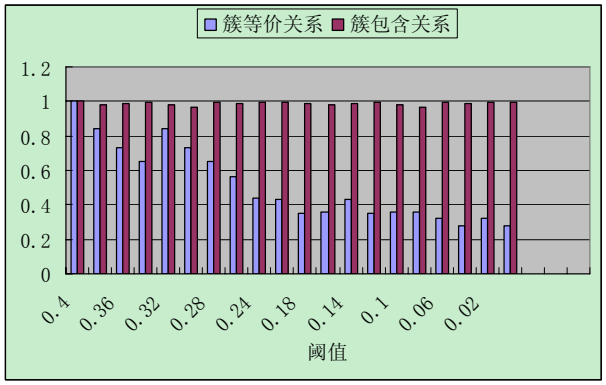


图 4-5 簇等价关系与簇包含关系的比较

观察图 4-4 可以看到，随着阈值的降低，聚类生成的簇数单调增加。从图 4-5 中观察到，低阈值层的簇对最高阈值层的包含关系都接近 100%，基本上是完全包含。这些现象说明，低阈值层的聚类结果基本完全保留高阈值层的聚类结果，在此基础上低阈值层还有新簇产生出来。

图 4-5 还可以看到，低阈值层与高阈值层的等价关系总体上处于下降的趋势。这个现象表明，低阈值层的聚类结果在等价关系上直接承接了高阈值层的部分聚类结果，但是等价关系只是包含关系的部分表现，还有相当一部分低阈值层簇是因为阈值的降低使新的 ESI 聚合到某些高阈值层簇中而形成的。

通过上面的分析，可以得到如下结论：

（1）从聚类结果整体看，低阈值层参与聚类的成员基本完全包含高阈值层参与聚类的成员，低阈值层聚类生成的簇基本完全包含高阈值层聚类生成的簇。

（2）从单个簇层面上看，低阈值层可能存在与高阈值层等价的簇，也可能存在由高阈值层变异而成的簇，也可能存在高阈值层没有的簇。

（3）随着阈值的降低，低阈值层对高阈值层的包含关系始终强烈，而等价关系逐渐减弱。

综合而言，因为强烈的包含关系，在低阈值层可以找到高阈值层的“影子”，保证低阈值层包含高阈值层的基本结构，这种“影子”关系是将两者联系起来的纽带。不仅如此，因为减弱的等价关系和越来越多新簇的产生，低阈值层还具有高阈值层没有的“内容”，保证在低阈值层发现高阈值层不具备的新结构，这些新的“内容”就可能是待探寻的、具有特殊意义的潜在结构。

4.2.2 阈值与其他参数的共同影响分析

聚类的实际过程中，不仅仅关系阈值需要调整，往往还有其他参数需要调整。此时会产生一个问题：阈值与其他参数一块调整，是否也能发现潜在结构。

本文选择“簇内成员数”这个参数，同关系阈值一起来讨论上面的问题。还是以前面的数据为例，不同的是，簇内成员数这个参数将随着阈值的下降做相应的调整，调整的原则是：阈值下降，参与聚类的 ESI 数目增多，会使聚类生成的簇数目太多，此时增大簇内成员数最小值和簇内成员数最大值，以保证低阈值层簇数目在最高阈值层簇数目上下 20%左右浮动，维持在一个比较稳定的水平。之所以采用这样的原则进行调整，是为了保证多次聚类结果处于一个相对稳定、可以进行比较的状态。除阈值和簇内成员数外，其他参数将固定不变。这样将最高阈值聚类结果分别与逐步降低的低阈值聚类结果进行 20 次比较，绘出同样的曲线图。

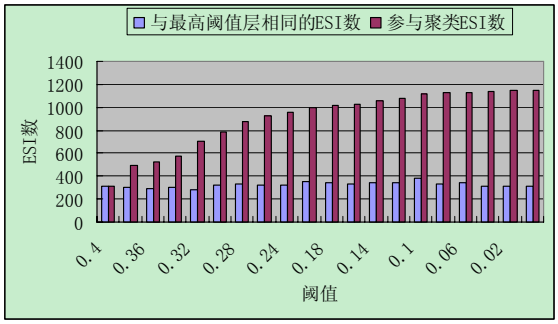


图 4-6 相同 ESI 与参与聚类 ESI 的比较（调整阈值和簇内成员数）

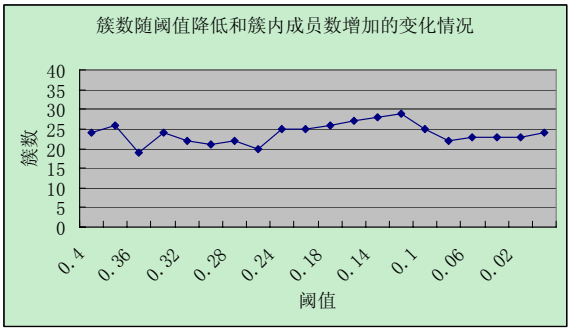


图 4-7 簇数的变化情况（调整阈值和簇内成员数）

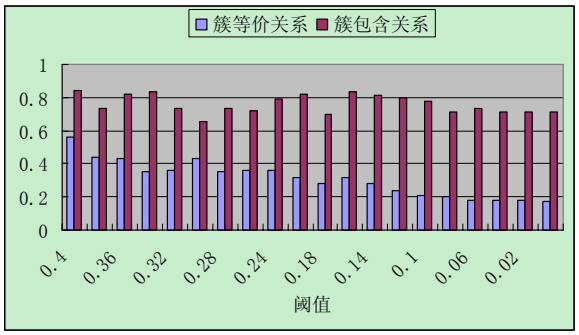


图 4-8 簇等价关系与簇包含关系的比较（调整阈值和簇内成员数）

从图 4-7、4-8 看到同时调整关系阈值和簇内成员数与仅仅调整阈值，对聚类结果没有根本性改变，只是生成簇数不是单调递增，而是保持基本稳定，如表 4-1 所示。

表 4-1 两种调整阈值方式（仅阈值调整；阈值、簇内成员数都调整）的效果比较

	仅调整阈值	调整阈值和簇内成员数	效果是否一致
参与聚类 ESI 数	单调递增	单调递增	√
相同 ESI 数	稳定	基本稳定	√
聚类产生簇数	单调递增	小幅波动变化；基本稳定	×
簇包含关系	完全包含；稳定	大部分包含；基本稳定	√
簇等价关系	单调递减	单调递减	√

通过分析，可以得出结论，从整体上看，同时调整关系阈值和簇内成员数参数，与只调整关系阈值，效果是一样的，都能保证低阈值层同高阈值层有“影子”关系，而且在低阈值层可能存在高阈值层没有的新“内容”，只要合理调整簇内成员数参数，阈值策略是可以在调整阈值的同时一块调整簇内成员数参数的。

4.3 阈值调整发现潜在结构的方法

“影子”关系说明不同阈值层之间存在相似性，新的“内容”说明不同阈值层之间存在差异性。本小节将在相似性和差异性的基础上，设计发现潜在结构的方法。首先，归纳簇与簇的两种关系，并定义潜在簇及其分类。然后，设计寻找潜在簇和潜在簇关系的步骤和流程。最后，给出潜在结构的展示方法。

4.3.1 潜在簇的种类划分

4.3.1.1 簇与簇的两种关系

在某一个阈值层面，簇与簇之间可能存在关系相关联。“影子”关系，表明

不同阈值层面簇与簇之间的也存在关系。因此，簇与簇关系归纳为两种类型，一种是同一个阈值层面中的 relative 关系，另一种是不同阈值层面之间的 twin 关系。

定义 4-1 簇与簇的 relative 关系

对于任意阈值层面的聚类簇 M，如果在本层存在另一个簇 N，并且 M 与 N 在关系矩阵中的关系不等于零，那么 M 与 N 之间具有 relative 关系。relative 关系的大小用 M 与 N 在关系矩阵中的关系大小来表示。

公式 4-1 簇与簇的 relative 关系

$$
\begin{aligned}
&M \xrightarrow{\text{Relative Relation}} N, if \\
&matrix[M, N] \neq 0 \\
&\text{其中} \\
&M \in CLUSTER\_RESULT \\
&N \in CLUSTER\_RESULT \\
&matrix\text{是簇的关系矩阵}
\end{aligned}
$$

从簇与簇的 relative 关系，可以引申出簇的 relatives，定义如下。

定义 4-2 簇的 relatives

对于任意阈值层面的聚类簇 M，如果在本层存在其他簇 N_1，N_2，……，N_p，并且 M 与 N_1，N_2，……，N_p 之间具有 relative 关系，那么 N_1，N_2，……，N_p 是 M 的 relatives。

公式 4-2 簇的 relatives

$$
\begin{aligned}
&relatives(M) = \left\{ (N_1, N_2, \cdots, N_p) \mid M \xrightarrow{\text{Relative Relation}} N_i \right\} \\
&\text{其中} \\
&M \in CLUSTER\_RESULT \\
&N_i \in CLUSTER\_RESULT \\
&i = 1,2\cdots, p
\end{aligned}
$$

定义 4-3 簇与簇的 twin 关系

对于 level_k 阈值层的簇 M，另一个 Level_g 阈值层中的簇 N，如果满足条件：

（1）M 与 N 的成员交集数目最大，

（2）M 与 N 的成员交集数目占 M 成员数的比例达到一定水平，

那么 M 与 N 之间具有 twin 关系。twin 关系的大小用 M 与 N 的成员交集数目占 M 成员数的比例来表示。

公式 4-3 簇与簇的 twin 关系

$$
\begin{aligned}
&M \xrightarrow{\text{Twin Relation}} N, if \\
&Num(M \cap N) = MAX\left\{ Num\{M \cap \alpha\}, a \in LEVEL\_g \right\}, \\
&\frac{Num(M \cap N)}{Num(M)} \geq \text{MIN\_SIMILARITY\_TWIN} \\
&\text{其中} \\
&M \in CLUSTER\_RESULT\_LEVEL\_k \\
&N \in CLUSTER\_RESULT\_LEVLE\_g
\end{aligned}
$$

同样，从簇与簇的 twin 关系，可以引申出簇的 twin，定义如下。

定义 4-4 簇的 twin

对于 level_k 阈值层的簇 M，如果另一个 Level_g 阈值层中的簇 N 与 M 具有 twin 关系，那么 M 的 twin 是 N。

公式 4-4 簇的 twin

$$twin(M)=N, if$$
$$M \xrightarrow{Twin\,Relation} N$$
其中
$$M \in CLUSTER\_RESULT\_LEVEL\_k$$
$$N \in CLUSTER\_RESULT\_LEVLE\_g$$

4.3.1.2 5种潜在簇

低阈值层具有高阈值层不能体现的潜在内容，应该在低阈值层中将考察目标定位在潜在簇上。首先，潜在簇具有位置性。潜在簇位于低阈值层的聚类结果中。其次，潜在簇具有相对性。某一低阈值层面的潜在簇是相对于一个特定的高阈值层面而言的。还有，潜在簇具有新颖性。潜在簇的某些簇内成员只有在低阈值层才能被聚合到簇中，而在高阈值层聚类时不参与聚类。这一点体现了低阈值层与高阈值层聚类之间的本质差别，是潜在簇最重要的特性。

根据这些特性，结合不同阈值层面簇与簇的 twin 关系，潜在簇的定义是：

定义 4-5 潜在簇

低阈值层（level_k）的簇 M，如果某一高阈值层（level_g）中找不到它的 twin，那么簇 M 是相对于这一高阈值层的潜在簇，或者说高阈值层 level_g 在低阈值层 level_k 中存在潜在簇 M。

公式 4-5 潜在簇

$$M是潜在簇，\ if$$
$$twin(M)\neq N_i$$
其中
$$M \in CLUSTER\_RESULT\_LEVEL\_k$$
$$N_i \in CLUSTER\_RESULT\_LEVLE\_g$$
$$i=1,2,\cdots$$

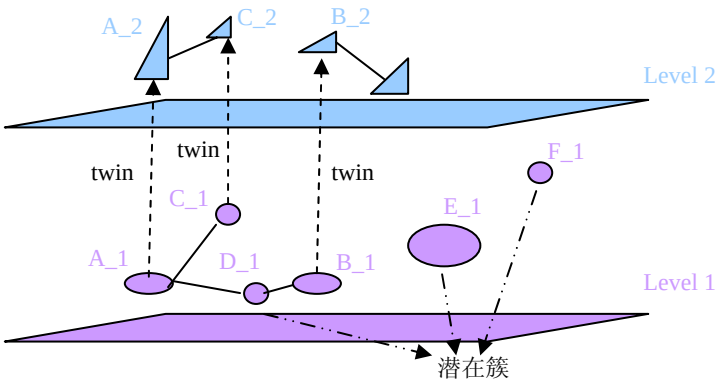


图 4-9 潜在簇示意图

图 4-9 展示，低阈值层 level_1 中，簇 A_1、B_1、C_1 均能在高阈值层 level_2

的中找到各自的 twin，A_2、B_2、C_2，因此簇 A_1、B_1、C_1 都不是相对于高阈值层 level_2 的潜在簇。而簇 D_1、E_1 在高阈值层 level_2 中没有 twin，因此簇 D_1、E_1 是相对于高阈值层 level_2 的潜在簇，或者说高阈值层 level_2 在低阈值层 level_1 中有潜在簇 D_1、E_1 存在。

潜在簇可能会有很多，应该特别关注有两类，一类是潜在孤立簇，另一类是潜在衔接簇。

潜在孤立簇是那些与其他簇没有关系的簇，或者是那些与其他簇有较少、较弱关系的簇。在知识结构的形成过程中，孤立簇可能代表某些主题的萌芽出现，也可能代表某些主题在整个知识结构中自成体系，这两种簇都有较强的独立性。虽然孤立簇与其他簇没有关系或者关系很弱，但是随着时间的推移，萌芽阶段的孤立簇有可能发展壮大，内容越来越丰富，体系越来越复杂，逐步成为以后的主要研究点。而自成体系孤立簇也具有重要的意义，它所代表的研究主题是整个主题结构中重要的一部分，也是与其他主题在研究内容上有明显差异的一部分。

潜在衔接簇是那些与其他簇有关系的簇，而且这些簇在发现知识结构潜在性方面具有特殊的作用。在高阈值层中，衔接簇不出现在聚类结果中。但是，在低阈值层中，衔接簇出现在聚类结果中，而且与其他簇之间具有一定的关系。通过这些关系，衔接簇可能将某些重要的簇联系在一起，而这些重要的簇在高阈值层面可能表现为没有联系。

从图 4-9 中可以看到，低阈值层的簇 E_1 与本层中的其他簇之间没有关系，是个潜在孤立簇。然而，簇 E_1 内成员数较多，应该具有一定的意义，值得关注。低阈值层的簇 D_1 是个潜在衔接簇。在低阈值层，簇 D_1 与两个重要的簇 A_1、B_1 之间建立了联系。而簇 A_1、B_1 在高阈值层上有 twin，分别是簇 A_2、B_2，而且簇 A_2、B_2 之间没有关系。此时高阈值层会产生一个错觉，簇 A_2、B_2 之间是没有联系的，两者是割裂开的。但实际上，在较低的阈值关系层面，簇 A_2、B_2 经过重新聚合，得到它们的变异簇 A_1、B_1，变异簇之间有一个簇 D_1 联系了两者，值得关注。

本文将低阈值层聚类结果中潜在簇归纳为 5 类，包括绝对孤立簇、自成体系孤立簇、马鞍衔接簇、分支衔接簇和直接衔接簇。详细定义和描述如下。

（1）绝对孤立簇

定义 4-6 绝对孤立簇

某阈值层的簇 M，如果 M 是潜在簇，M 的簇内成员数小于预先设定的孤立簇的最小成员数，并且与本阈值层其他簇无关系，那么簇 M 称作绝对孤立簇。

公式 4-6 绝对孤立簇

$$
\begin{aligned}
&M\text{是绝对孤立簇，}\ if\\
&twin(M)\neq N_i\\
&Num(M)<\text{MIN\_SIZE\_ISOLATE}\\
&relatives(M)=\varnothing\\
&\text{其中}\\
&M\in CLUSTER\_RESULT\_LEVEL\_k\\
&N_i\in CLUSTER\_RESULT\_LEVLE\_g\\
&i=1,2,\cdots
\end{aligned}
$$

从图 4-9 可以看到，低阈值层 level_1 中的簇 F_1 的簇内成员数较少，小于预先设定的孤立簇最小成员数。在本阈值层中簇 F_1 没有 relatives。在 level_2 高阈值层中没有 twin。因此，F_1 是 level_1 低阈值层相对于 level_2 高阈值层的一个绝对孤立簇。

（2）自成体系孤立簇

定义 4-7 自成体系孤立簇

某阈值层的簇 M，如果 M 是潜在簇，M 的簇内成员数大于预先设定的孤立簇最小成员数，并且与本阈值层其他簇无关系，或者有关系，且关系数小于预先设定的最小关系数，关系值小于预先设定的最小关系值，那么簇 M 称作自成体系孤立簇。

公式 4-7 自成体系孤立簇

$$
\begin{aligned}
&M\text{是自成体系孤立簇，}\ if\\
&twin(M)\neq N_i\\
&Num(M)\geq\text{MIN\_SIZE\_ISOLATE}\\
&relatives(M)=\varnothing\\
&\text{or}\\
&twin(M)\neq N_i\\
&Num(M)\geq\text{MIN\_SIZE\_ISOLATE}\\
&\text{Num}(\text{relatives(M)})<MIN\_NUM\_RELATIVES\\
&\text{Value}\left(\text{M}\xrightarrow{\text{Relative Relation}}M_j\right)<MIN\_RELATION\_RELATIVES\\
&\text{其中}\\
&M\in CLUSTER\_RESULT\_LEVEL\_k\\
&M_j\in relatives(M)\\
&N_i\in CLUSTER\_RESULT\_LEVLE\_g\\
&i=1,2,\cdots\\
&j=1,2,\cdots
\end{aligned}
$$

观察图 4-9，低阈值层 level_1 中的簇 E_1 的簇内成员数较多，大于或者等于预先设定的孤立簇最小成员数。在本阈值层中簇 E_1 没有 relatives。在 level_2 高阈值层中没有 twin。因此，E_1 是 level_1 低阈值层相对于 level_2 高阈值层的一个自成体系孤立簇。

（3）马鞍衔接簇

定义 4-8 马鞍衔接簇

某阈值层的簇 M，如果 M 是潜在簇，不是孤立簇，M 的簇内成员数大于预先设定的最小成员数，M 的 relatives 都是潜在簇，那么簇 M 称作马鞍衔接簇。

公式 4-8 马鞍衔接簇

$$
\begin{aligned}
&M\text{是马鞍衔接簇，}\ if\\
&twin(M)\neq N_i\\
&Num(M)\geq MIN\_SIZE\_LINK\\
&relatives(M)\neq\varnothing\\
&twin(M_j)\neq N_i\\
&\text{其中}\\
&M\in CLUSTER\_RESULT\_LEVEL\_k\\
&M_j\in relatives(M)\\
&N_i\in CLUSTER\_RESULT\_LEVLE\_g\\
&i=1,2,\cdots\\
&j=1,2,\cdots
\end{aligned}
$$

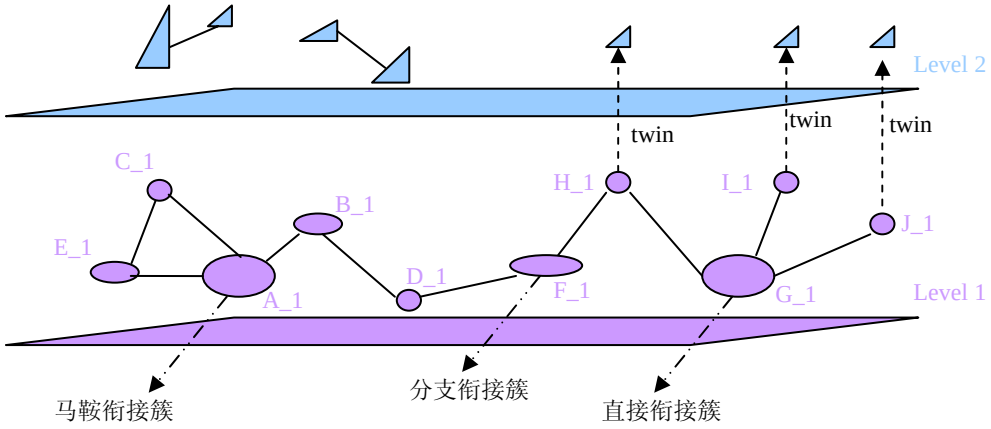


图 4-10 衔接簇示意图

观察图 4-10，低阈值层 level_1 中的簇 A_1 的簇内成员数达到最小成员数的要求。在本阈值层中簇 A_1 有三个 relatives 是 B_1、C_1、E_1。三个 relatives 在 level_2 高阈值层中都没有 twin，即 B_1、C_1、E_1 都是潜在簇。因此，A_1 是 level_1 低阈值层相对于 level_2 高阈值层的一个马鞍衔接簇。

（4）分支衔接簇

定义 4-9 分支衔接簇

某阈值层的簇 M，如果 M 是潜在簇，不是孤立簇，M 的簇内成员数大于预先设定的最小成员数，M 的 relatives 中有且仅有一个非潜在簇，那么簇 M 称作分支衔接簇。

公式 4-9 分支衔接簇

M是分支衔接簇，if

$twin(M) \neq N_i$

$Num(M) \geq MIN\_SIZE\_LINK$

$relatives(M) \neq \varnothing$

$twin(M_p) = N_q$，∃唯一一对M_p, N_q

其中

$M \in CLUSTER\_RESULT\_LEVEL\_k$

$M_p \in relatives(M)$

$N_i \in CLUSTER\_RESULT\_LEVLE\_g$

$i = 1,2,\cdots,q,\cdots$

观察图 4-10，低阈值层 level_1 中簇 F_1 达到最小成员数的要求。在本阈值层中簇 A_1 有两个 relatives 是 D_1、H_1，有且只有 H_1 在 level_2 高阈值层中有 twin，不是潜在簇。因此，F_1 是 level_1 低阈值层相对于 level_2 高阈值层的一个分支衔接簇。

（5）直接衔接簇

定义 4-10 直接衔接簇

某阈值层的簇 M，如果 M 是潜在簇，不是孤立簇，M 的簇内成员数大于预先设定的最小成员数，M 的 relatives 中至少有两个非潜在簇，那么簇 M 称作直接衔接簇。

公式 4-10 直接衔接簇

M是直接衔接簇，if

$twin(M) \neq N_i$

$Num(M) \geq MIN\_SIZE\_LINK$

$relatives(M) \neq \varnothing$

$twin(M_p) = N_q$，∃至少两对M_p, N_q

其中

$M \in CLUSTER\_RESULT\_LEVEL\_k$

$M_p \in relatives(M)$

$N_i \in CLUSTER\_RESULT\_LEVLE\_g$

$i = 1,2,\cdots$

$q \in i$

观察图 4-10，低阈值层 level_1 中簇 G_1 达到最小成员数的要求。在本阈值层中簇 G_1 有三个 relatives 是 H_1、I_1、J_1，三个 relatives 在 level_2 高阈值层中都有 twin，都不是潜在簇。因此，G_1 是 level_1 低阈值层相对于 level_2 高阈值层的一个直接衔接簇。

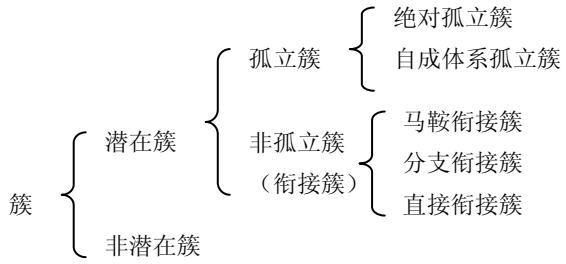


图 4-11 簇及潜在簇的种类划分

表 4-2 阈值策略中各种簇之间的区别

簇种类	区别
潜在簇	在高阈值层的聚类结果中是否有 twin
非潜在簇	
孤立簇	在本阈值层的聚类结果中是否有 relative
非孤立簇	
绝对孤立簇	簇成员数不同
自成体系孤立簇	
马鞍衔接簇	relatives 中是非潜在簇的数目不同
分支衔接簇	
直接衔接簇	

根据本文的定义，聚类结果中簇的种类划分体系如图 4-11 所示，各种簇的区别如表 4-2 所示。

4.3.2 潜在结构的发现

通过簇的 relative 和 twin 关系，提炼出阈值策略最需要关注的 5 类潜在簇。应用阈值策略发现潜在结构的主要思路就是在对不同阈值层面聚类结果进行比较的基础上，发现高阈值层面中无法体现的潜在结构。这个过程的核心问题就是为高阈值层面在低阈值层面中寻找到可能存在的潜在簇及其潜在关系。

4.3.2.1 发现潜在结构的主要步骤

阈值策略实际上是通过逐步降低待聚类对象的关系阈值，对同一数据集中挑选待聚类的对象，分别进行聚类计算。然后对不同阈值层面的聚类结果进行比较，找到聚类结果存在较明显差异的阈值层次，本文称作“突变阈值区间”。从突变阈值区间中确定一个高阈值层和一个低阈值层。在这两个阈值层面上，应用判断标准计算低阈值层中簇与簇的 relative 关系，低阈值层与高阈值层之间簇与簇的

twin 关系。在此基础上，再根据潜在簇的判断标准寻找低阈值层相对于高阈值层的各种潜在簇。图 4-12 描述的是应用阈值层略，在多阈值层面的聚类结果中发现潜在结构的主要步骤。

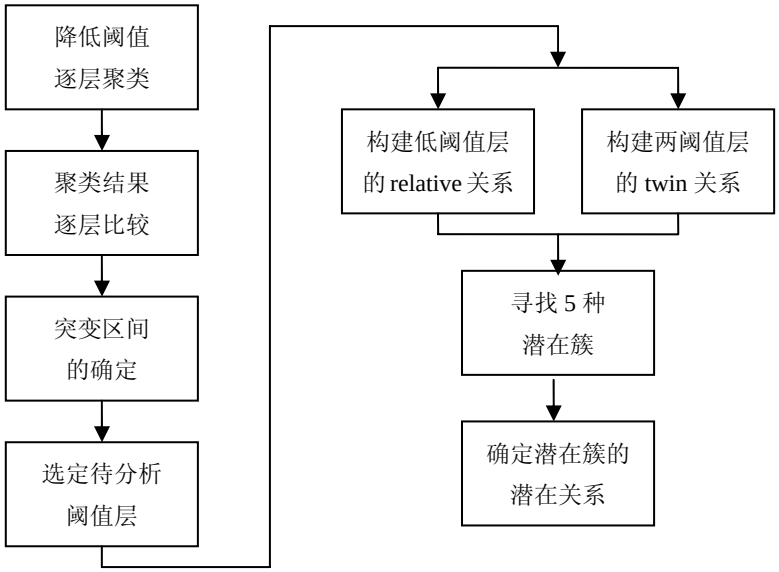


图 4-12 发现潜在结构的主要步骤

4.3.2.2 聚类结果的比较以及突变区间的确定

不难想象，相对于同一个高阈值层，并不是所有的低阈值层聚类结果都会存在潜在的内容，只有当新簇产生，或者两阈值层簇与簇的包含关系、等价关系发生剧烈变化，此时的低阈值层相对于高阈值层差异明显，蕴涵潜在结构的可能性大大增强。因此，有必要确定与高阈值层有明显差异的一个或多个低阈值层。目的在于：（1）提高发现潜在结构的可能性。并不是所有低阈值层的聚类结果中都有蕴涵潜在结构的可能性，那么先从中挑选出可能性大的数个低阈值层，对它们再启动潜在结构的寻找过程，会提高发现潜在结构的成功率；（2）提高发现潜在结构的运算速度。如果高阈值层的阈值较高，阈值降低的步长较小，那么在大量的低阈值层中逐个寻找潜在结构，运算速度非常低。如果先确定出部分差异明显的，缩小后续的寻找范围，会减少整个过程的运算时间，提高运算速度。

为了判断两个阈值层面的聚类结果是否差异明显，需要比较的项目包括：

（1）低阈值层与高阈值层簇数之差异度

定义 4-11 低阈值层与高阈值层簇数之差异度

低阈值层聚类结果生成簇的数目是 m，高阈值层聚类结果生成簇的数目是 n，两阈值层簇数之差为 m－n 的绝对值，那么两阈值层簇数之差异度为簇数之差与高阈值层簇数目之比。

公式 4-11 低阈值层与高阈值层簇数之差异度

$$PercentClusterNumDif = \frac{|m-n|}{n}$$

其中

$$m = ClusterNum(CLUSTER\_RESULT\_LEVEL\_LOW)$$

$$n = ClusterNum(CLUSTER\_RESULT\_LEVEL\_HIGH)$$

（2）低阈值层对高阈值层的包含度

定义 4-12 低阈值层对高阈值层的包含度

低阈值层聚类结果生成簇的数目是 m，高阈值层聚类结果生成簇的数目是 n，低阈值层中的 p 个簇同高阈值层中有 q 个簇是包含关系，那么低阈值层对高阈值层的包含度为 q 同 m，n 中的较大者之比。

公式 4-12 低阈值层对高阈值层的包含度

$$PercentContainClusterNum = \frac{q}{MAX\{m,n\}}$$

其中

$$q = ContainClusterNum$$

$$m = ClusterNum(CLUSTER\_RESULT\_LEVEL\_LOW)$$

$$n = ClusterNum(CLUSTER\_RESULT\_LEVEL\_HIGH)$$

计算包含度，分母之所以取 m，n 中的较大者，有两个原因。一个原因是，若分母取高阈值层的簇数 n，当高阈值层的所有簇都能被低阈值层包含，此时不论低阈值层的聚类结果如何变化，包含度均为 1。显然在这种情况下计算得到的包含度不能体现两个阈值层的差异，与实际情况不符合。还有一个原因是，采用 m，n 中的较大者作为分母，可以把低阈值层与高阈值层的聚类差异带入包含度的计算，分母越大，说明两阈值层聚类的结果差异越大，相应的，包含度越小，此时计算的包含度则与实际情况相符合。

（3）低阈值层对高阈值层的等价度

定义 4-13 低阈值层对高阈值层的等价度

低阈值层聚类结果生成簇的数目是 m，高阈值层聚类结果生成簇的数目是 n，低阈值层中的 r 个簇同高阈值层中有 r 个簇是等价关系，那么低阈值层对高阈值层的等价度为 r 同 m，n 中的较大者之比。

公式 4-13 低阈值层对高阈值层的等价度

$$PercentSameClusterNum = \frac{r}{MAX\{m,n\}}$$

其中

$$r = SameClusterNum$$

$$m = ClusterNum(CLUSTER\_RESULT\_LEVEL\_LOW)$$

$$n = ClusterNum(CLUSTER\_RESULT\_LEVEL\_HIGH)$$

计算等价度，分母取 m，n 中的较大者的原因同上。

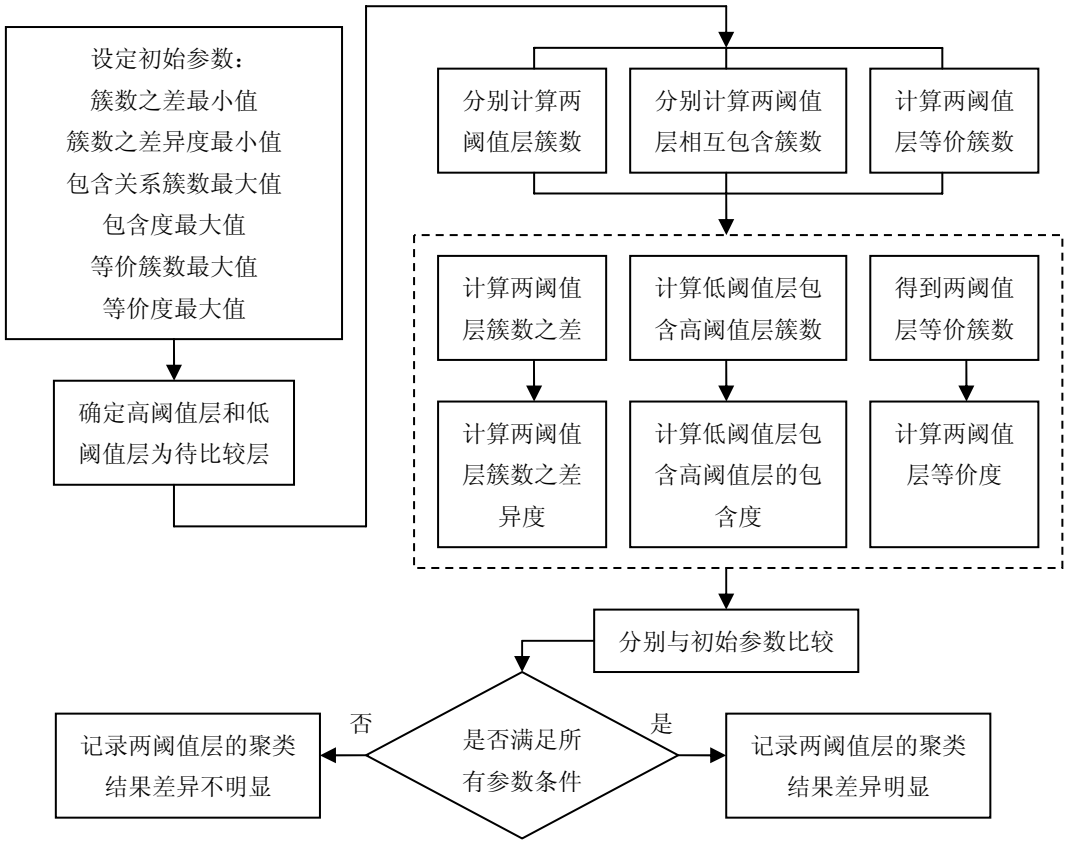


图 4-13 两个阈值层面聚类结果差异比较的算法流程

应用上述 3 个项目的比较，可以度量低阈值层和高阈值层之间聚类结果的差异程度，具体算法流程如图 4-13 所示。当差异度大于最小值、包含度小于最大值、等价度小于最大值这三个条件同时满足时，就认为比较的两个阈值层聚类结果是差异明显的。当差异程度较大时，在低阈值层发现潜在结构的可能性也较大。通过差异程度的计算，可以从所有阈值层中挑选出部分差异明显的聚类结果，以及这些结果所处的阈值层次。本文定义这些聚类结果差异明显的阈值层次为阈值策略的“突变阈值区间”。聚类结果差异度的两种不同比较方式，会使阈值策略的实施过程产生两种不同的“突变阈值区间”：

（1）固定的高阈值层

通常，将最高阈值层设置为高阈值层，固定不变，而低阈值层将随着阈值的降低逐步变换。本处理过程的步骤描述为：

第一步：获取最高阈值和阈值降低步长。

第二步：将最高阈值层设置为高阈值层。

第三步：逐步降低阈值，将当前阈值层设置为低阈值层。

第四步：计算高阈值层与低阈值层的聚类结果差异性。

第五步：判断两阈值层的聚类结果是否差异明显。

第六步：是，记录低阈值层所处的阈值并将该阈值推入突变阈值区间队列，保存低阈值层的聚类结果；否，删除低阈值层的聚类结果。

第七步：判断当前阈值层是否是最低阈值层。

第八步：是，突变阈值区间寻找过程结束；否，转入第三步继续寻找。

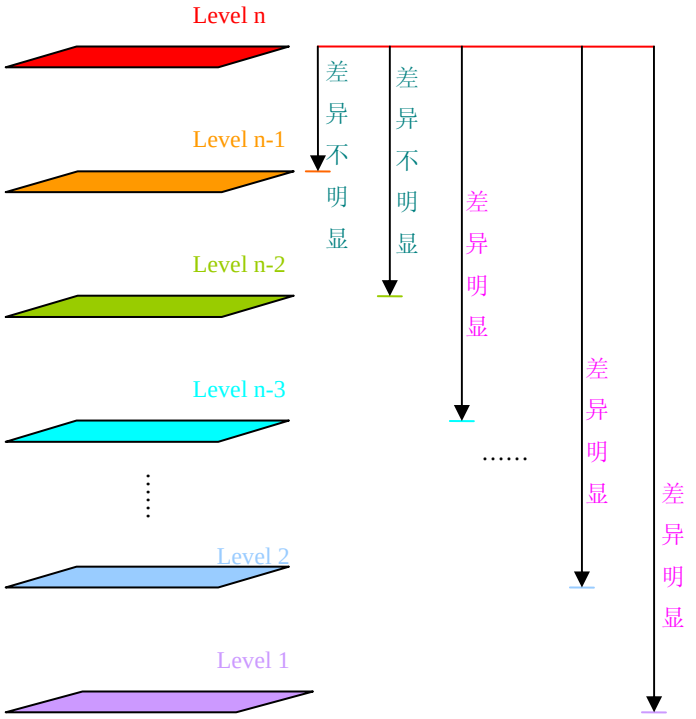


图 4-14 寻找突变阈值区间的步骤（高阈值层固定不变）

从图 4-14 可以看到，最高阈值层 Level_n 始终被确定为高阈值层，与下面的多个低阈值层逐一进行比较。从结果上看，高阈值层（也就是最高阈值层）与 level_n-3 差异明显，而且从 level_n-3 往后，下面的所有低阈值层都是差异明显的。产生这个现象的原因，正如本章前面“阈值变化对聚类结果的影响”的分析，是因为低阈值层的聚类结果对高阈值层几乎是完全的包含关系，level_n-3 下面的阈值层基本上具有 level_n-3 的结构特征，因而与最高阈值层都是差异明显的。

（2）降低的高阈值层

高阈值层变动，而非固定设置。高阈值层首先被设置在最高阈值层。发现一个差异明显的低阈值层时，要重新设置高阈值层，将此时的低阈值层设置为高阈值层。在此基础上，后面紧接着的低阈值层都将与这个更新的高阈值层进行差异性比较。本处理过程的步骤为：

第一步：获取最高阈值和阈值降低步长。

第二步：高阈值层初始设置在最高阈值层。

第三步：逐步降低阈值，将当前阈值层设置为低阈值层。

第四步：计算高阈值层与低阈值层的聚类结果差异性。

第五步：判断两阈值层的聚类结果是否差异明显。

第六步：是，记录低阈值层所处的阈值并将该阈值推入突变阈值区间队列，保存低阈值层的聚类结果，同时，重新设置高阈值层，将此时的低阈值层设置为

高阈值层；否，删除低阈值层的聚类结果。

第七步：判断当前阈值层是否是最低阈值层。

第八步：是，突变阈值区间寻找过程结束；否，转入第三步继续寻找。

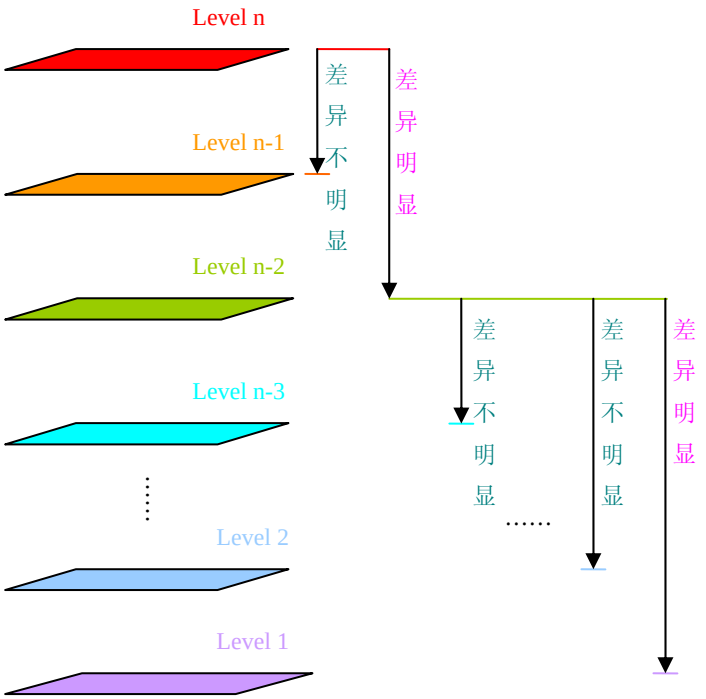


图 4-15 寻找突变阈值区间的步骤（高阈值层逐步降低）

从图 4-15 可以看到，最高阈值层 Level_n 被确定为高阈值层的初始位置，高阈值层与 level_n-2 差异明显，因此 level_n-2 的聚类结果被保存，并且将 level_n-2 记录在突变阈值区间中。除此之外，高阈值层将被重新设置，更新到 level_n-2 的位置。随着阈值一次次降低，每到发现差异明显的两个阈值层，高阈值层将降低到当前低阈值层的位置，并以此为比较的基点，继续进行差异性比较。

差异是否明显的两种判断方式，会产生两种形式的突变阈值区间。这里，以 ESI 中的 COMPUTER 领域的数据为例，在其他参数相同的情况下，最高阈值设定为 0.4，降低阈值的步长设定为 0.02。采用高阈值层固定不变和高阈值层逐步降低两种方式，在$[0,0.4]$的阈值区间中，分别寻找突变阈值区间。

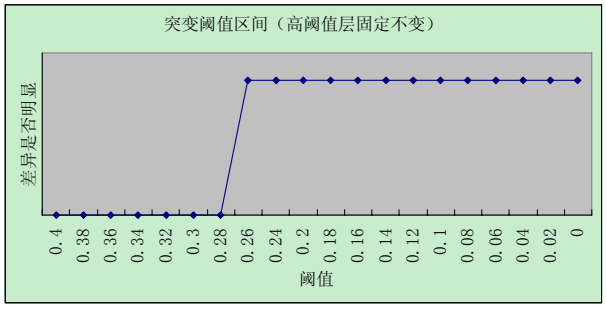


图 4-16 突变阈值区间（高阈值层固定不变）

第一种方式产生的结果如图 4-16 所示。曲线的前面一段一直处于低位，表明前 7 步的低阈值层与最高阈值层没有明显的差异。当阈值降低到第 8 步的时候，此时低阈值层位于 0.26，相对于最高阈值层有明显差异，曲线陡升，表明差异性由不明显阶越式突变为明显。从此处开始，曲线将一直处于高位，表明后面余下的低阈值层始终与最高阈值层由明显差异。

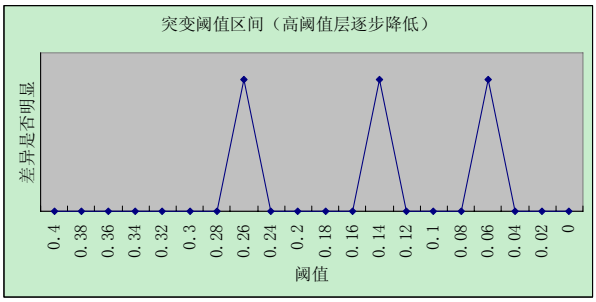


图 4-17 突变阈值区间（高阈值层逐步降低）

第二种方式产生的结果如图 4-17 所示。曲线的前面一段与第一种方式相同，前 7 步一直处于低位，当阈值降低到第 8 步的时候陡升。按照第二种方式的要求，此时高阈值层将由 0.4 阈值层降低为 0.26 阈值层。阈值继续降低，图中曲线在随后紧接着的 0.24 阈值层陡降，表明此时的低阈值层（0.24）与高阈值层（0.26）没有明显的差异。阈值经过多次降低，曲线在 0.16 阈值层再一次陡升，表明此处与 0.26 高阈值层差异明显。此时再一次重设高阈值层，由 0.26 变为 0.16，并继续与前面相同的比较过程，直到低阈值层到达最低阈值层为止。

从表现上看，两种方式的不同之处在于，前一种方式只体现最高阈值层聚类结果在阈值的整个可变范围内的突变情况，因此图中表现为曲线有且只有一次陡升，一旦升起曲线将始终处于高位一直到最低阈值层。而后一种方式体现了在阈值的整个可变范围内聚类结果的突变层次，因此图中表现为曲线可能有多次陡升，曲线升起后立即陡降，到下一个突变阈值层再次陡升，这个过程一直循环进行到最低阈值层。

从结果上看，两种方式的不同之处在于，前一种方式可以寻找到相对于最高阈值层首次有明显差异的低阈值层，以及下面相对于最高阈值层有明显差异的所有低阈值层，在这些低阈值层中能逐层发现最高阈值层没有的潜在结果。因为都是相对于最高阈值层的明显差异，所以突变阈值区间表现为持续性。而后一种方式可以寻找到此次阈值策略过程的差异层次，同样的数据集随着阈值的降低，聚类结果有明显差异的阈值层能通过这种方式确定。因为每一差异明显的阈值层在前面是低阈值层，在后面是高阈值层，所以突变阈值区间表现为跳跃性。

根据寻找潜在结构的主体步骤，突变阈值区间被确定后，应该应用于确定一个高阈值层面和一个低阈值层面，以深入分析潜在结构。前一种方式发现的都是相对于最高阈值层的潜在结构，因而高阈值层始终位于最高阈值层，而低阈值层

从差异明显的第一个阈值层开始，并依次被下一个阈值层更替。后一种方式发现的是层次更迭的潜在结构，因而高阈值层和低阈值层是紧邻的有明显差异的两个阈值层，而且两者依次由下一对紧邻的有明显差异的两个阈值层更替。

4.3.2.3 寻找潜在簇及其潜在关系的算法流程

通过差异明显的判断，发现本次阈值层略的突变阈值区间。从区间中确定用于潜在结构深入分析的两个阈值层面之后，就可以开始在低阈值层面的聚类结果中寻找相对于高阈值层面聚类结果的潜在结构。

（1）构建 relative 关系和 twin 关系。

a、构建低阈值层面中的 relative 关系，详细算法流程如图 4-18 所示。

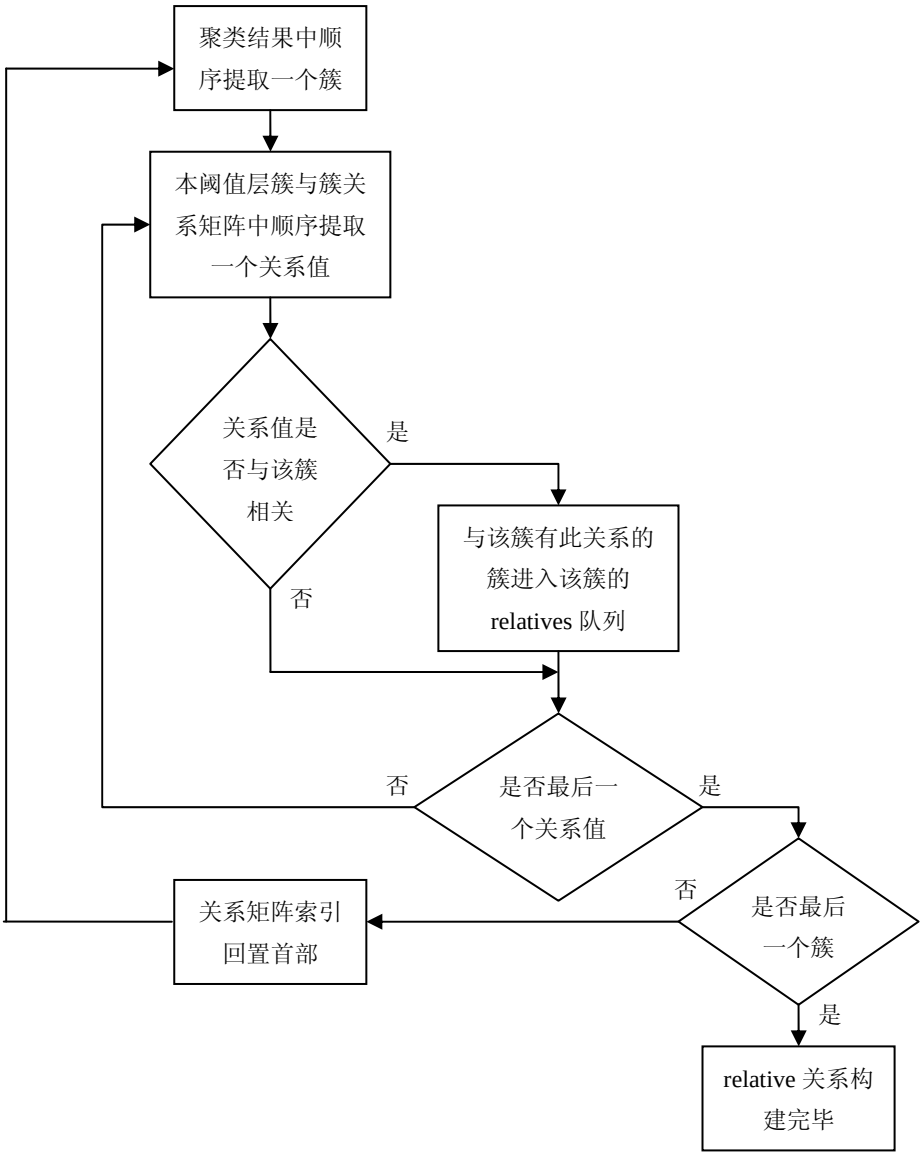


图 4-18 构建 relative 关系的算法流程

b、构建 twin 关系的详细算法流程如图 4-19 所示。

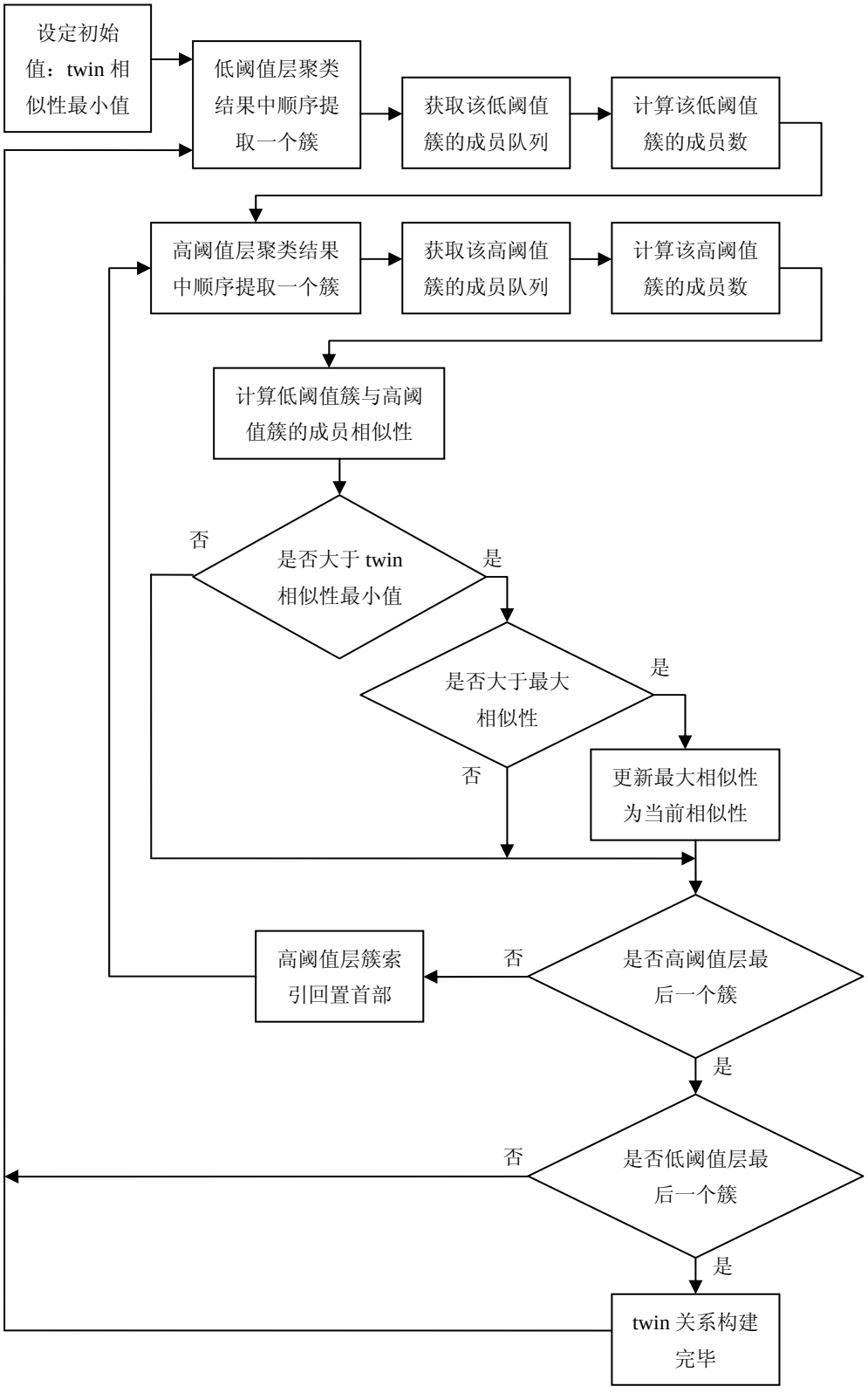


图 4-19 构建 twin 关系的算法流程

（2）在低阈值层的聚类结果中寻找潜在簇。详细算法流程如图 4-20 所示。

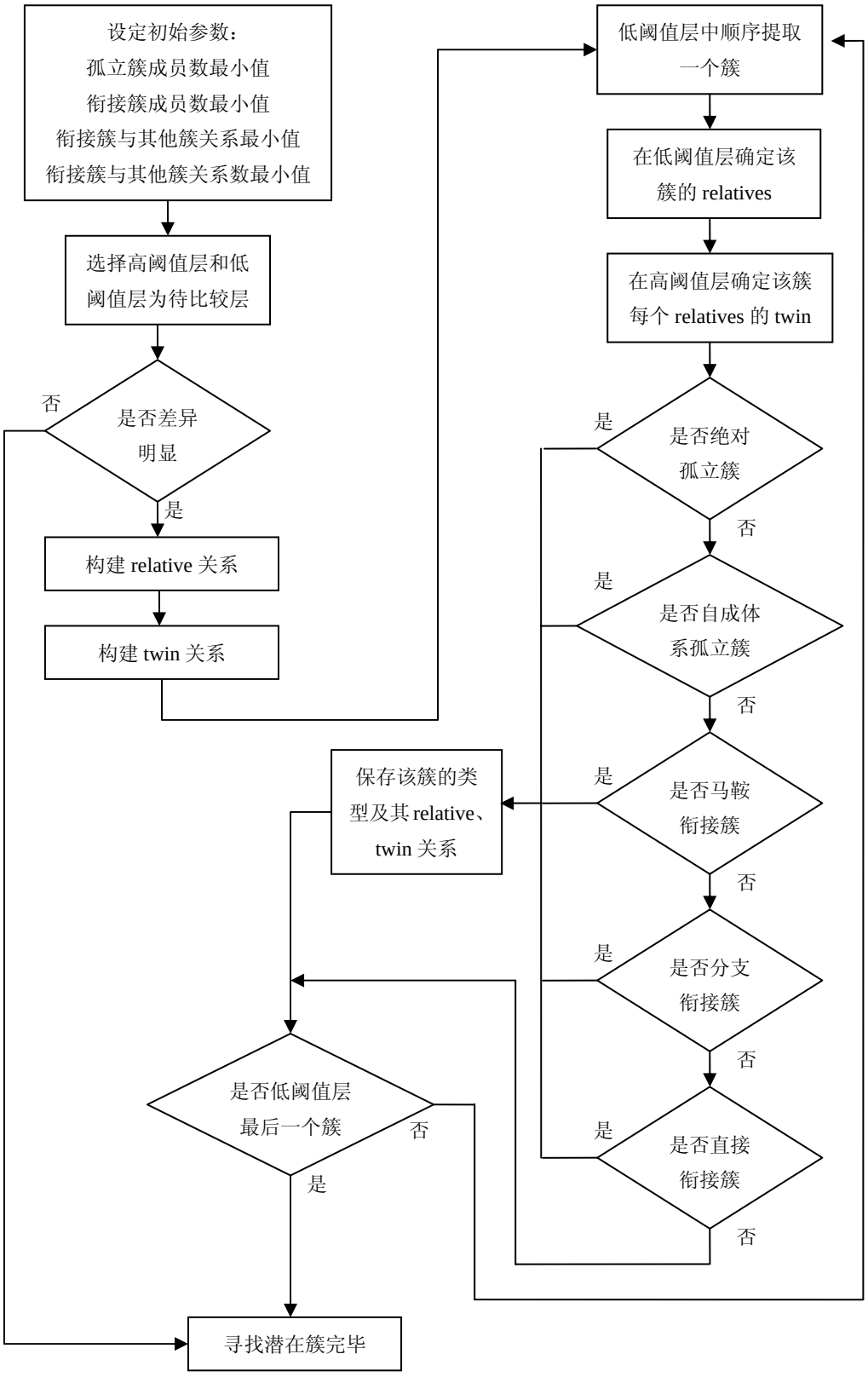


图 4-20 寻找潜在簇的算法流程

在寻找潜在簇的过程中，核心步骤是判断一个簇属于哪一种类型的潜在簇。其中，判断绝对孤立簇和自成体系孤立簇的详细算法流程如图 4-21 所示。

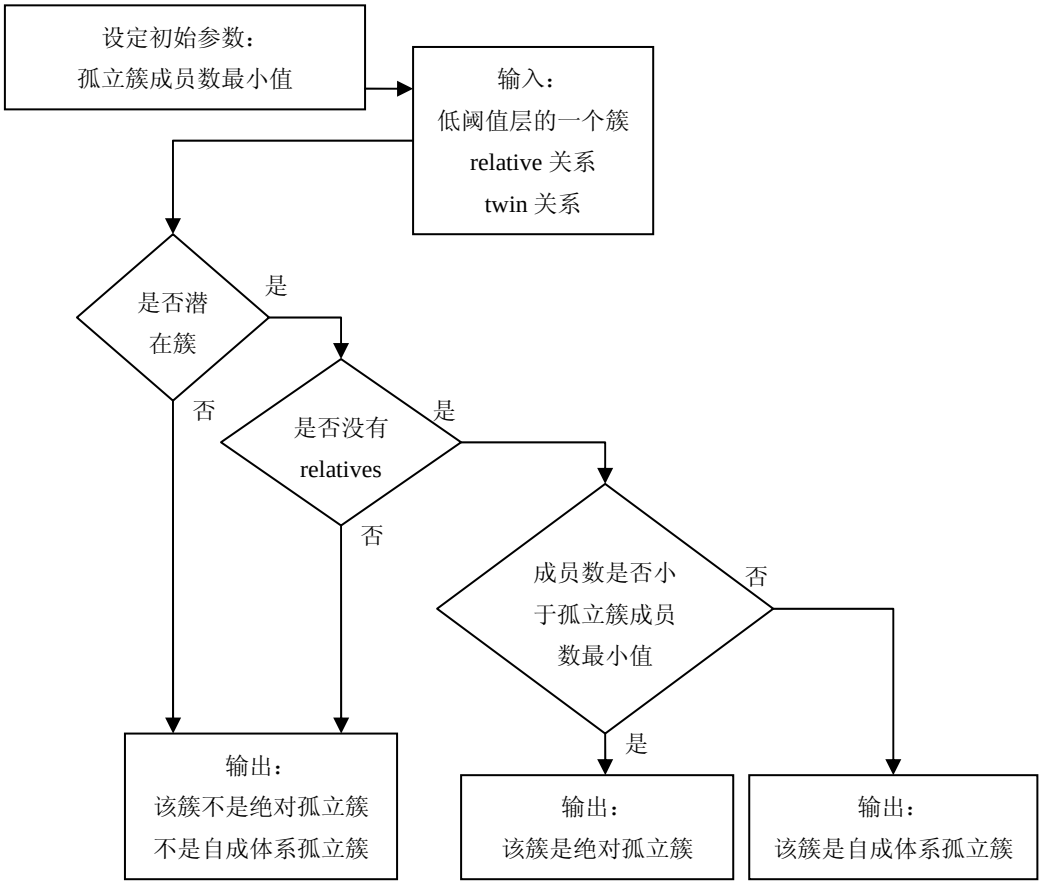


图 4-21 判断绝对孤立簇和自成体系孤立簇的算法流程

判断马鞍衔接簇、分支衔接簇、直接衔接簇的详细算法流程如图 4-22 所示。

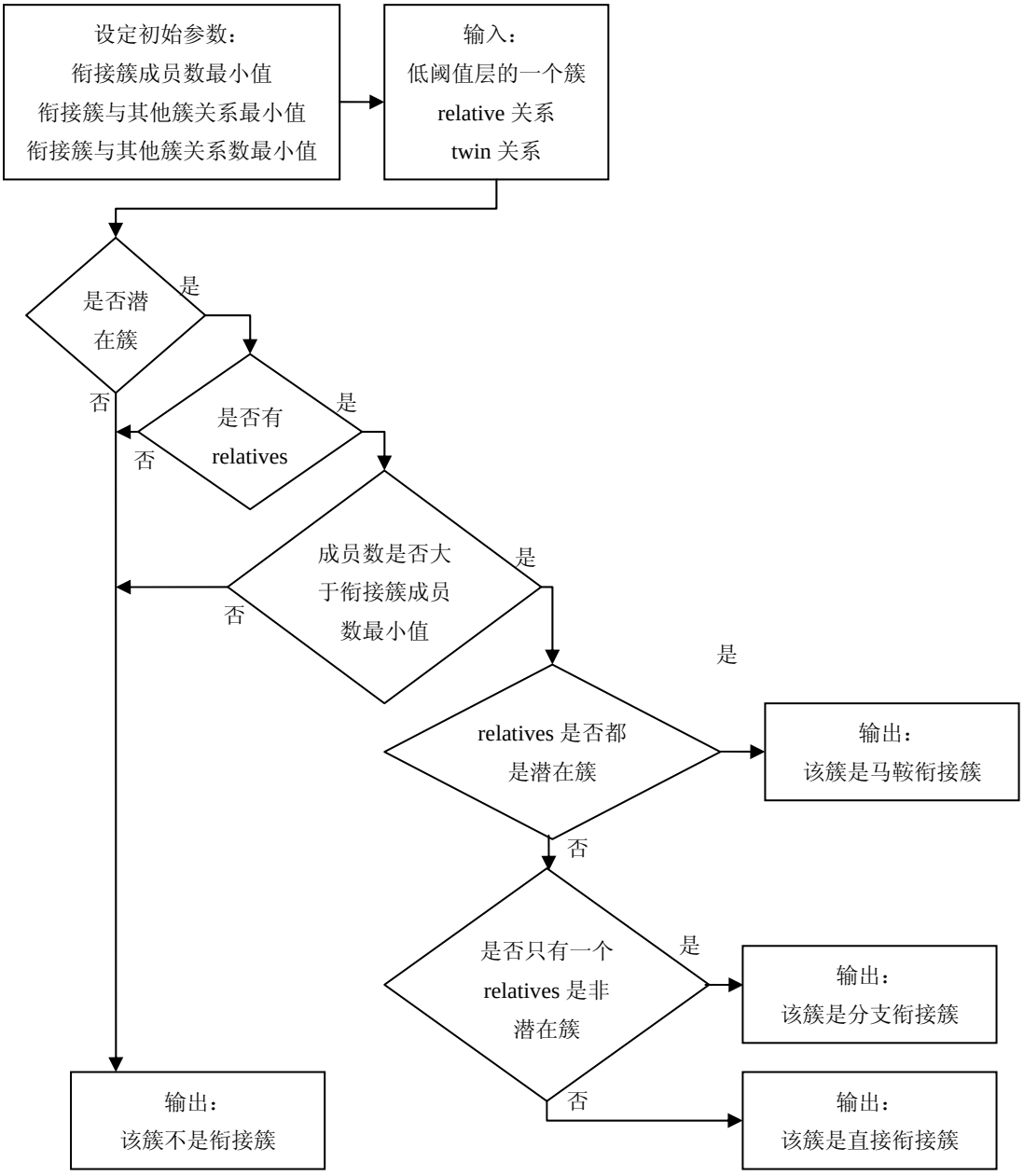


图 4-22 判断马鞍衔接簇、分支衔接簇和直接衔接簇的算法流程

（3）寻找潜在结构有两个主要任务。一个是在低阈值层面的聚类结果中寻找 5 种潜在簇，另一个任务是要确定这些潜在簇同高阈值层簇的潜在关系。潜在簇如果是孤立簇，它自身在高阈值层没有 twin，在本阈值层也没有 relatives，因此孤立簇同高阈值层的簇之间没有潜在关系。潜在簇如果是衔接簇，虽然它自身在高阈值层中没有 twin，但是在本阈值层中有 relatives，而且这些 relatives 可能在高阈值层中有 twin。由此，可以通过潜在簇同本阈值层簇的 relative 关系，以及 relatives 同高阈值层簇的 twin 关系，构建潜在簇同高阈值层簇的潜在关系。

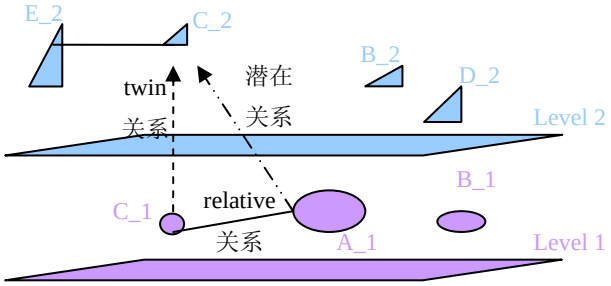


图 4-23 确定潜在簇同高阈值层簇的潜在关系

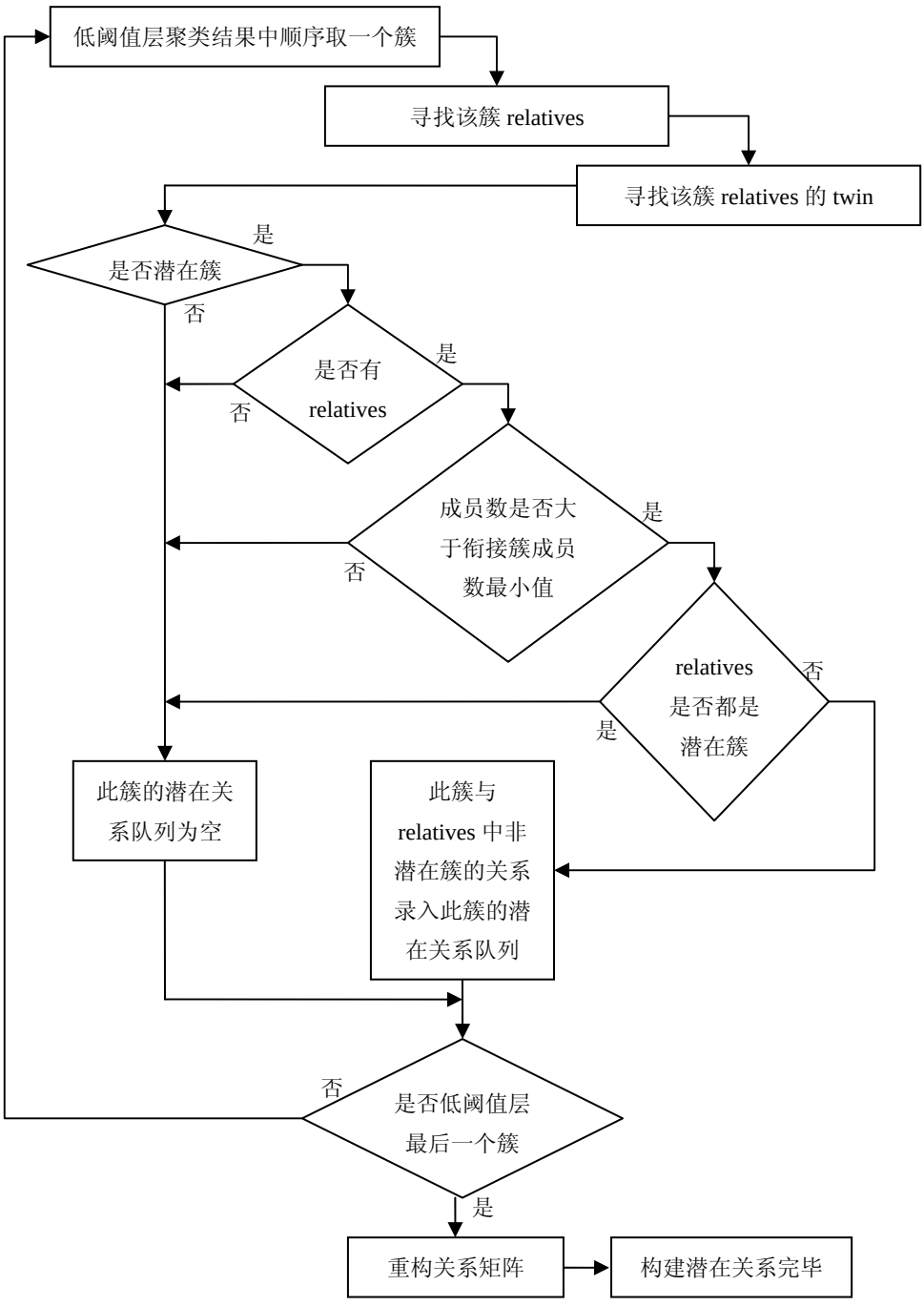


图 4-24 构建潜在关系的算法流程

从图 4-23 可以看到，经由 relative 关系和 twin 关系，原本没有关系的低阈值层簇 A_1 同高阈值层簇 C_2 建立了潜在关系，潜在关系的大小无法精确计算，但可以近似等于簇 A_1 同簇 C_1 的 relative 关系大小。

经过上面的分析，确定潜在关系的详细算法流程如图 4-24 所示。

4.3.3 潜在结构的展示

应用阈值策略的主要目的是在低阈值层面的聚类结果中寻找到高阈值层面不能反映的潜在结构。潜在结构包括潜在簇和潜在关系。因此，可视化展示的基本内容是：（1）展示高阈值层中的簇及其关系；（2）展示从低阈值层中寻找的潜在簇；（3）展示潜在簇同高阈值层簇的潜在关系。根据展示内容，可视化展示的基本要求是：（1）能区分高阈值层簇和来自低阈值层的潜在簇；（2）能区分高阈值层中簇与簇的关系和来自低阈值层的潜在关系；（3）能区分潜在簇的种类。

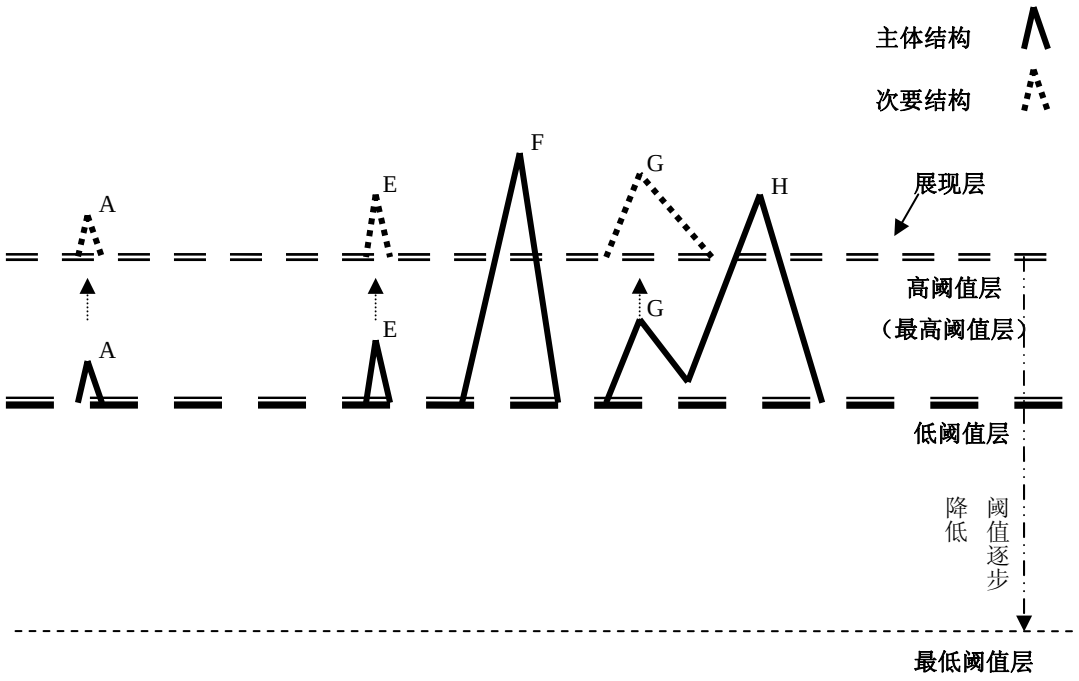
针对两种突变阈值区间，可视化展现的核心问题是高阈值层、低阈值层该如何确定。对于第一种方式，持续的突变阈值区间，高阈值层设置在本次阈值策略的最高阈值层，始终固定不变，低阈值层首先设置在第一次差异明显的低阈值层，从此开始，后续的所有低阈值层都将依次成为低阈值层，从中发现相对于最高阈值层的潜在结构。因此，在这种方式下，可视化展示应该采用下面的方法：

（1）主体结构由最高阈值层的簇及其关系构成，维持不变。

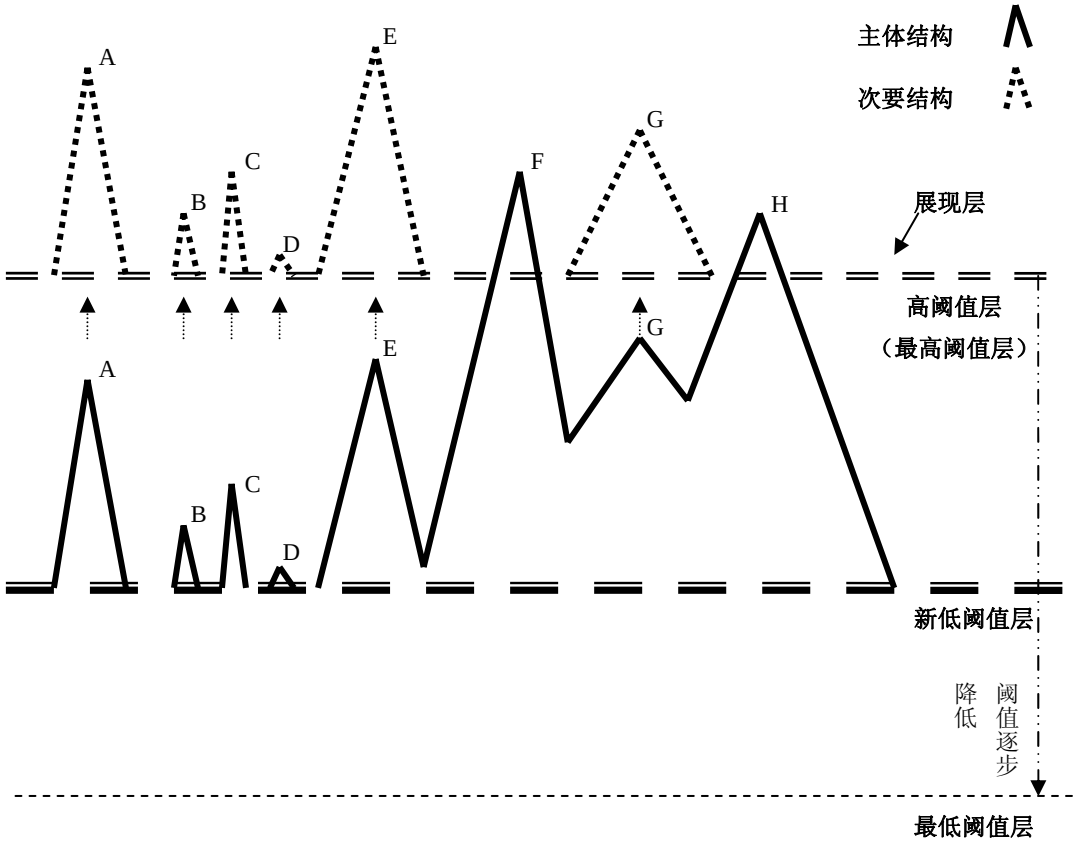
（2）次要结构由低阈值层的潜在簇，以及潜在簇同高阈值层的潜在关系构成，附属在主体结构上。

（3）次要结构随着低阈值层的变化而变化。

图 4-25 描述了采用持续的突变阈值区间，展示潜在结构的方法。（a）中高阈值层固定设置在最高阈值层上，其聚类结果簇 F、簇 H 体现了主体结构。对于差异明显的低阈值层，从中能找到簇 A、簇 E、簇 G 等潜在簇。将这些潜在簇向上移动至最高阈值层面中，形成次要结构，结合潜在关系，将次要结构附属于主体结构，最终在最高阈值层上体现此时低阈值层聚类结果中蕴涵的潜在簇和潜在关系。当低阈值层将定位在突变阈值区间中下一个有潜在簇的阈值层面上时，见图（b），从中能找到簇 A、簇 B、簇 C、簇 D、簇 E、簇 G 等潜在簇。展示层依旧定位在最高阈值层，按上面的方法形成结构图。随着阈值的降低，会形成多幅可视化图形。这些图形的主体结构始终不变，由最高阈值层聚类结果决定，而次要结构会低阈值层的变化而变化。



（a）



（b）

图 4-25 潜在结构展示（采用持续的突变阈值区间）

对于第二种方式，跳跃的突变阈值区间，高阈值层和低阈值层分别设置在相邻的差异明显的两个阈值层面上。而且随着阈值的降低，高阈值层和低阈值层同

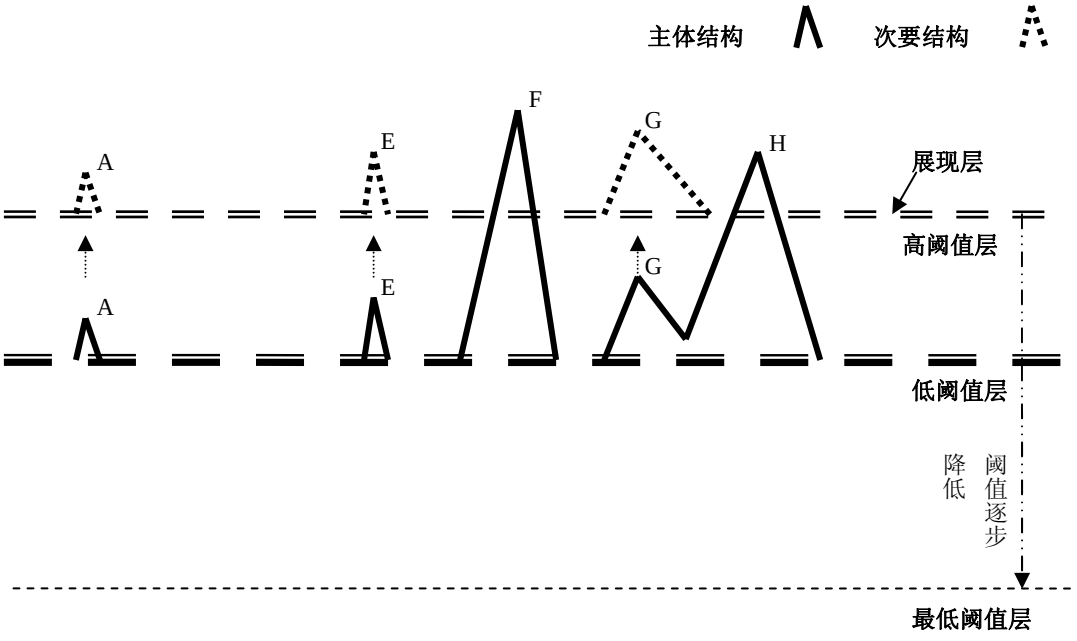
时向下推移，使当前的低阈值层成为新的高阈值层，使下一个相邻的差异明显的阈值层成为新低阈值层。因此，在这种方式下，可视化展示应该采用下面的方法：

（1）主体结构由高阈值层的簇及其簇与簇的关系构成。

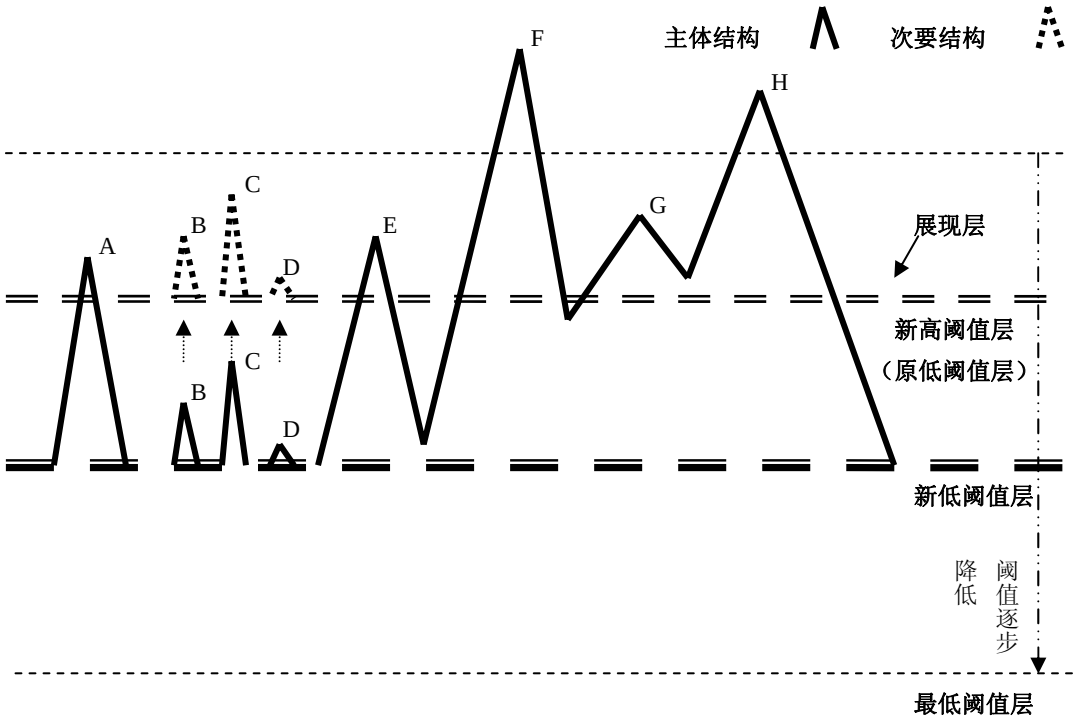
（2）次要结构由低阈值层的潜在簇，以及潜在簇同高阈值层的潜在关系构成，次要结构附属在主体结构上。

（3）图形的主体结构、次要结构随着高阈值层、低阈值层的跳跃而变化。

图 4-26 描述了采用持续的突变阈值区间，展示潜在结构的方法。从图（a）可以看到，高阈值层的聚类结果簇 F、簇 H 体现了主体结构。对于差异明显的低阈值层，从中能找到簇 A、簇 E、簇 G 等潜在簇。将这些潜在簇向上移动至高阈值层面中，形成次要结构，结合潜在关系，将次要结构附属于主体结构，最终在高阈值层上体现此时低阈值层中蕴涵的潜在簇和潜在关系。从图（b）可以看到，由于采用跳跃的突变阈值区间，高阈值层更新在（a）中的低阈值层上，主体结构改为簇 A、簇 E、簇 F、簇 G、簇 H 体现。而且低阈值层也重新定位在突变阈值区间中下一个有潜在簇的阈值层面上，从中能找到簇 B、簇 C、簇 D 等潜在簇。随着阈值的降低，会形成多幅可视化图形。这些图形的主体结构逐层变化，由每次的高阈值层聚类结果决定，次要结构也会低阈值层的变化而变化。



（a）



（b）

图 4-26 潜在结构展示（采用跳跃的突变阈值区间）

通过上面的分析，归纳潜在结构展示的主要流程如图 4-27 所示：

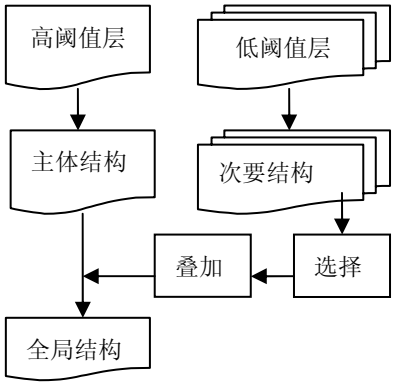


图 4-27 潜在结构展示的操作流程

4.4 小结

表达知识结构，通常使用聚类的方法绘制科学知识地图。为了实现聚类计算过程的较高效率、聚类结果表达的精简明了，一般使用较高的阈值（待聚类对象之间的关系）。然而，这样的得到的结果会忽略掉一些关系不强烈的潜在对象，掩藏住一些处于萌芽阶段的潜在结构。

本章的主要目标是想通过阈值策略，在原有结果的基础上发现潜在内容。阈

值策略的主要内容是：

（1）在逐步降低的阈值层面上进行多次聚类，经过比较发现，低阈值层的聚类结果与高阈值层通过簇的包含关系体现出两个阈值层面的相似性。更重要的是，除了相似性，低阈值层的聚类结果体现出与高阈值层的差异性，而且随着阈值的降低，差异性逐渐积累增强，越来越明显。这个现象说明，降低阈值多次聚类，不同阈值层面的聚类结果通过相似性被关联起来，同时又通过差异性两个阈值层面都有各自的特点，特别是低阈值层具有高阈值层没有的潜在内容。在此，用定量的方法证明了以阈值逐步降低为基础的阈值策略用于发现潜在结构是可行的。

（2）探讨潜在簇的种类。根据潜在簇种类划分的需要，先定义了同一阈值层中簇与簇的 relative 关系和不同阈值层之间的 twin 关系。利用这两种关系，以及聚类结果中簇的其他特性，首先对潜在簇给出定义，进而将潜在簇划分为 5 个种类并给出定义，它们分别是：绝对孤立簇、自成体系孤立簇、马鞍衔接簇、分支衔接簇、直接衔接簇。

（3）阈值策略实施的步骤。主体步骤包括突变阈值区间的确定、两个阈值层 relative 关系和 twin 关系的构建、5 种潜在簇的寻找、潜在关系的确定。这里，因为低阈值层与高阈值层比较方式的不同，突变阈值区间会有两种表现形式，一种是持续的突变阈值区间，另一种是跳跃的突变阈值区间。两种突变阈值区间，会直接影响后续寻找潜在结构的实施对象，即在哪一个低阈值层中相对于哪一个高阈值层寻找潜在结构。

（4）潜在结构的可视化展示。同寻找潜在结构一样，潜在结构展示也会因为突变阈值区间的两种不同类型而采用不同方式。但是，不论突变阈值区间以何种方式确定低阈值层和高阈值层，潜在关系关联的是哪个低高阈值层和高阈值层，可视化展示图都将以高阈值层的聚类结果作为主体结构，低阈值层的潜在簇作为次要结构，并通过潜在关系将次要结构附属在主体结构上。

5 时间策略研究

在揭示知识结构的时候，利用关联关系的强弱，可以从不同的阈值层面找到潜在结构。通过另一个特性——时间，则可以在一个时间段上发现待分析对象的演变结构，即确定发生演变的对象，及其演变的方式。

本章研究时间窗的划分来发现、展现演变结构的方法。先设计主要的研究线路。紧接着探讨不同时间窗聚类结果之间的关系，明确时间窗的划分用于探寻演变结构的可能性。在此基础上，提出基于时间窗划分的发现演变结构的具体方法，主要研究内容包括演变簇的种类划分、演变结构的发现、演变结构的展示。

5.1 研究线路

时间策略的研究线路如图 5-1 所示。聚类在前期已有的工作上实现，因此图中灰色部分是本章的主要研究内容。

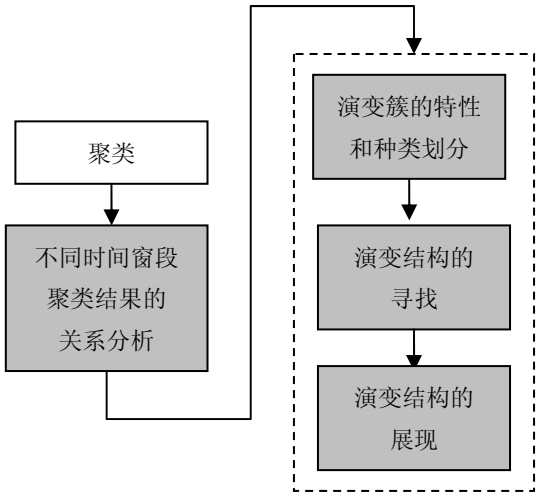


图 5-1 时间策略研究线路

5.2 不同时间窗聚类结果的关系分析

本小节将重点探讨不同时间窗聚类结果之间的关系。

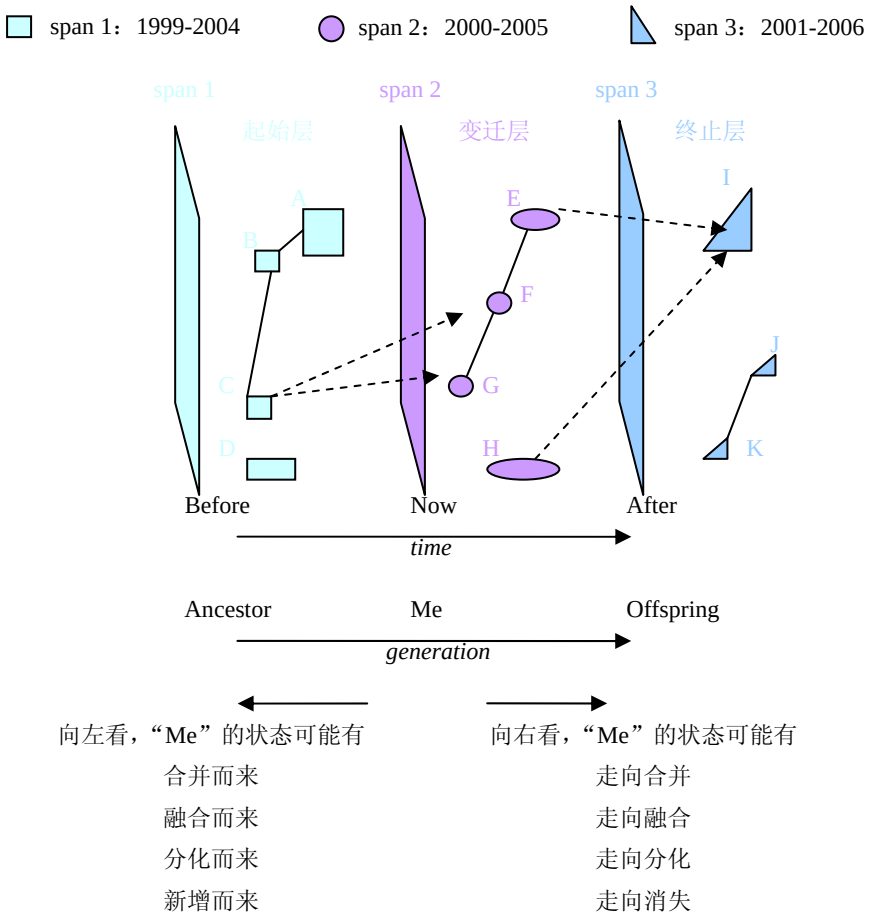


图 5-2 时间策略多时间窗聚类结果关系分析模型

时间窗聚类只反映当前时间窗的知识结构。然而，每一个时间窗不是孤立的。针对簇内成员，从图 5-2 可以发现，span_1 层中的簇 C 分别与 span_2 层中的簇 F、簇 G 有相同的成员，而且 span_2 层中的簇 E、簇 H 分别与 span_1 层中的簇 I 有相同的成员。这个现象说明，公共时间窗内的成员在两个时间窗的聚类结果中传承，而且以前聚合在一起的成员可能现在被分离，或者现在分离开的成员可能后来被聚合起来。成员在时间窗之间相继传承，同时又在聚类结果中被重新分配，这种现象正如知识进化理论提出的“遗传继承”。将 span_2 称作“Me”，span_1 可以视作“Me”的“Ancestor”（祖先），span_3 可以视作“Me”的“Offspring”（后代）。“Me”层中的簇 F、簇 G 有祖先簇 C，簇 E、簇 H 有后代簇 I。结合时间推移和遗传继承关系，任意三个相邻时间窗依次称作演变的起始层、变迁层、终止层。以变迁层视作分析的主体对象，向左看，“Me”的状态可能是由“Ancestor”合并、融合、分化、新增而来；向右看，“Me”的状态可能是走向合并、融合、分化、消失，生成“Offspring”。如“Me”层中的簇 F、簇 G 有可能是由祖先簇 C 分化而来，簇 E、簇 H 有可能是走向融合后生成后代簇 I。

上述分析可以看到，一个时间窗从祖先层继承一些成员，同时还会遗传一些

成员给后代层，从而保证从一个时间窗可以发现另一个时间窗的“影子”，时间窗与时间窗之间具有相似性。另外，因为不同时间窗内还包含不同的成员，时间窗之间会存在差异，尤其当两个时间窗的公共时间窗非常小时，这两个时间窗的差异性就更大。

下面将在数量上度量两个时间窗聚类结果的相似性和差异性，并进行比较，进而分析通过时间窗划分的方式来探寻演变结构的可能性。以 ESI 中的 COMPUTER 领域的数据为例，整个时间段为 1996 至 2005，每个时间窗的时间跨度为 3 年，共划分时间窗 8 个。依次将相邻两个时间窗的聚类结果进行比较，共 7 次。方法与阈值策略相同，绘出比较项目随时间推移的变化曲线。

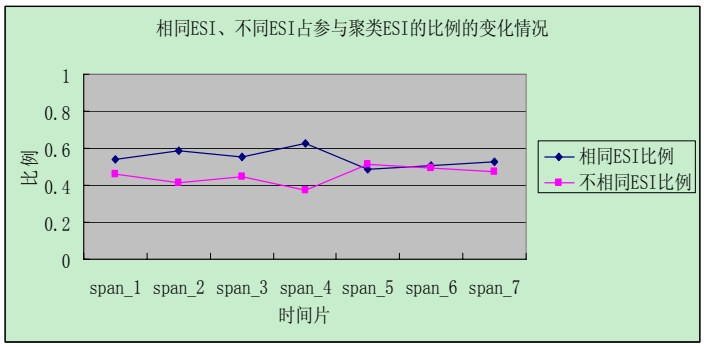


图 5-3 相同 ESI、不同 ESI 占参与聚类 ESI 的比例的变化情况

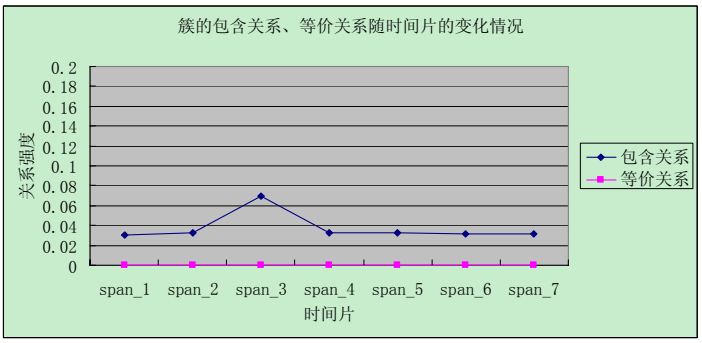


图 5-4 簇的包含关系、等价关系随时间窗的变化情况

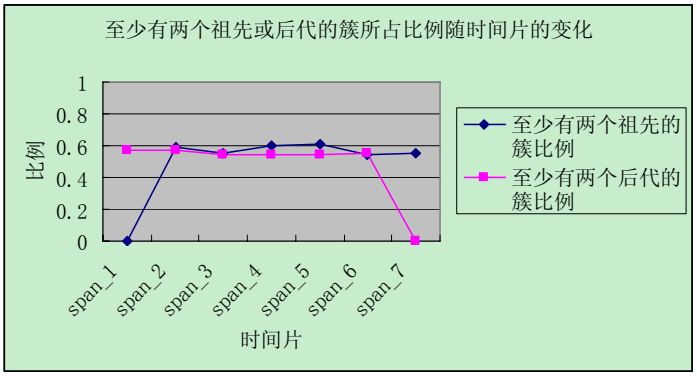


图 5-5 至少有两个祖先或后代的簇所占比例随时间窗的变化

综合而言，观察图 5-3、5-4、5-5 可以看到，划分时间窗产生多个聚类结果，

这些结果之间不是割裂开的，它们之间的关系总结如下：

（1）参与聚类的成员中相同成员所占比例是稳定，在后一时间窗可以找到前一时间窗的“影子”，这种“影子”关系在时间维度上称作“遗传继承”，是将两者联系起来的纽带。

（2）较大的不同成员所占比例、稳定的聚类簇数、非常弱的包含关系、为零的等价关系，表明在后一时间窗还具有前一时间窗不一样的“内容”，这些不一样的“内容”可能体现出的正是待探寻的、具有特殊意义的演变关系和演变过程。

（3）虽然不同时间窗的簇与簇之间有很强的非包含关系（即很弱的包含关系和等价关系），但是每一个时间窗的聚类结果中有多余一半的簇拥有多于一个祖先或后代，反映了非包含关系并不表明不同时间窗的簇之间完全不相关，相反，有很多簇与另一个时间窗的很多簇在簇内成员上是相似的。这种非包含关系中的相似性正反映了时间窗之间较强烈的“遗传继承”关系，是研究演变过程的基础。

5.3 时间策略发现演变结构的方法

前一小节讨论了不同时间窗聚类结果的相互关系，可以得到看到，簇与簇之间有共同的聚类成员，不同时间窗具有相似性，尽管如此，聚类结果从簇内成员看还是表现出强烈的差异性。强烈的非包含关系，以及较强的“遗传继承”性，说明可以在“遗传继承”关系的基础上，利用时间窗划分探寻演变结构。本小节设计在不同时间窗之间发现演变结构的方法。首先，归纳簇与簇在不同时间窗上的两种关系，以及簇的演变特性，并定义什么是演变簇，有哪些种类。然后，设计寻找演变簇和演变簇关系的步骤和流程。最后，给出演变结构的展示方法。

5.3.1 演变簇的种类划分

5.3.1.1 簇与簇的两种关系

在相邻的两个时间窗上，一种是前一个时间窗相对于后一个时间窗的的祖先关系，另一种是后一个时间窗相对于前一个时间窗的的后代关系。

定义 5-1 簇与簇的 ancestor 关系

对于某时间窗的簇 N，如果在祖先层存在一个簇 M，并且簇 N 与簇 M 有相同的簇内成员，那么 M 与 N 之间具有 ancestor 关系。ancestor 关系的大小用 M 与 N 之间的相同簇内成员数来表示。

公式 5-1 簇与簇的 ancestor 关系

$$M \xrightarrow{Ancestor\ Relation} N, if$$
$$member\_intersection(M, N) \neq \varnothing$$
其中
$$M \in CLUSTER\_RESULT\_LEVEL\_ANCESTOR$$
$$N \in CLUSTER\_RESULT\_LEVEL\_ME$$

从簇与簇的 ancestor 关系，可以引申出簇的 ancestors，定义如下。

定义 5-2 簇的 ancestors

对于某时间窗的簇 N，如果在祖先层存在一些簇 M_1，M_2，……，M_n，并且 N 与 M_1，M_2，……，M_n 之间具有 ancestor 关系，那么 M_1，M_2，……，M_n 是 N 的 ancestors。簇 N 的 ancestors 的个数是 n。

公式 5-2 簇的 ancestors

$$ancestors(N) = \{(M_1, M_2, \cdots, M_n) \mid M_i \xrightarrow{Ancestor\ Relation} N\}$$
其中
$$M_i \in CLUSTER\_RESULT\_LEVEL\_ANCESTOR$$
$$N \in CLUSTER\_RESULT\_LEVEL\_ME$$
$$i = 1,2\cdots, n$$

定义 5-3 簇与簇的 offspring 关系

对于某时间窗的簇 P，如果在后代层存在一个簇 Q，并且簇 P 与簇 Q 有相同的簇内成员，那么 Q 与 P 之间具有 offspring 关系。offspring 关系的大小用 P 与 Q 之间的相同簇内成员数来表示。

公式 5-3 簇与簇的 offspring 关系

$$Q \xrightarrow{Offspring\ Relation} P, if$$
$$member\_intersection(P, Q) \neq \varnothing$$
其中
$$P \in CLUSTER\_RESULT\_LEVEL\_ME$$
$$Q \in CLUSTER\_RESULT\_LEVEL\_OFFSPRING$$

从簇与簇的 offspring 关系，可以引申出簇的 offsprings，定义如下。

定义 5-4 簇的 offsprings

对于某时间窗的簇 P，如果在后代层存在一些簇 Q_1，Q_2，……，Q_p，并且 P 与 Q_1，Q_2，……，Q_p 之间具有 offspring 关系，那么 Q_1，Q_2，……，Q_p 是 N 的 offsprings。簇 P 的 offsprings 的个数是 p。

公式 5-4 簇的 offsprings

$$offsprings(P) = \{(Q_1, Q_2, \cdots, Q_n) \mid Q_i \xrightarrow{Ancestor\ Relation} P\}$$
其中
$$P \in CLUSTER\_RESULT\_LEVEL\_ME$$
$$Q_i \in CLUSTER\_RESULT\_LEVEL\_OFFSPRING$$
$$i = 1,2\cdots, p$$

簇与簇之间的关系一种是与前一个时间窗的 ancestor 关系，反映了当前时间窗是从哪些祖先演变而来，一种是与后一个时间窗的 offspring 关系，反映了当

前时间窗是将演化成哪些后代。两种关系的结合应用，将是后面演变簇定义和划分的基础。

5.3.1.2 簇的特性

从 ancestor 和 offspring 关系引申出一系列与“遗传继承”相关的特性。讨论簇的特性都将以三个相邻的时间窗为讨论对象，以中间时间窗为观察点，同前一个时间窗讨论与 ancestor 关系相关的特性，同后一个时间窗讨论与 offspring 关系相关的特性。

为了方便、清晰讨论簇的特性，先给出一些定义。

定义 5-5 公共时间窗内的簇成员

时间窗 span_1 的簇 M 和时间窗 span_2 的簇 N，span_1 和 span_2 的公共时间窗是 inter_span（span_1,span_2）。如果簇 M 的部分成员的时间属性在 inter_span（span_1,span_2）中取值，那么这部分成员称为簇 M 在公共时间窗 inter_span（span_1,span_2）内的簇成员。同样，如果簇 N 的部分成员的时间属性在 inter_span（span_1,span_2）中取值，那么这部分成员称为簇 N 在公共时间窗 inter_span（span_1,span_2）内的簇成员。

公式 5-5 公共时间窗内的簇成员

$$
\begin{aligned}
&member\_in\_inter\_span(M)\\
&=\{member\_M_1, member\_M_2, \cdots, member\_M_n\}, if\\
&member\_M_i \in M\\
&timeAttribute(member\_M_i) \in inter\_span(span\_1, span\_2)\\
&timeAttribute(M) = span\_1\\
&i = 1,2,\cdots,n\\
&\exists span\_1, span\_2, inter\_span(span\_1, span\_2)
\end{aligned}
$$

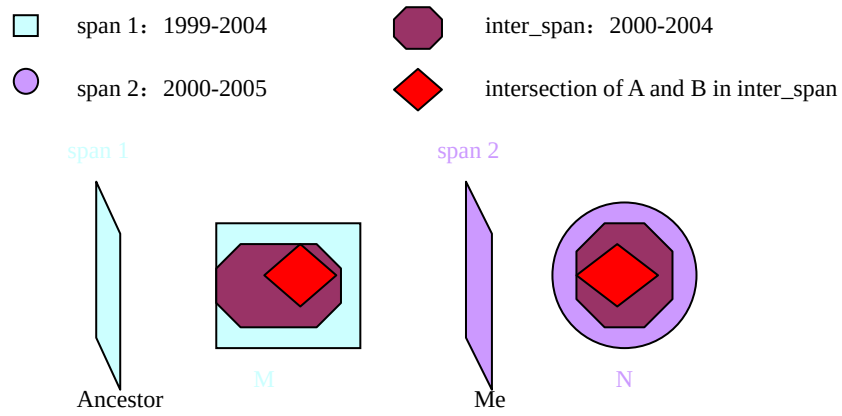


图 5-6 公共时间窗内的簇成员及其交集

图 5-6 描述，时间窗 span_1 和时间窗 span_2 的公共时间窗是 2000－2004，时间窗 span_1 中的簇 M 和时间窗 span_2 中的簇 N 在公共时间窗中都有各自的

簇内成员，图中由八边形表示。值得一提的是，M、N 在公共时间窗内的簇成员数可能不一样。簇 M 有 50 个公共时间窗的簇成员，簇 N 有 20。这说明原本聚合在一起的成员，经过时间推移，可能在另一个时间窗被分离，这正是有演变发生的重要标志。

定义 5-6 公共时间窗内的簇与簇的成员交集

时间窗 span_1 的簇 M 和时间窗 span_2 的簇 N，span_1 和 span_2 的公共时间窗是 inter_span（span_1,span_2）。如果簇 M 的公共时间窗内的某个成员也是簇 N 的公共时间窗内的成员，那么这样的成员集合称为簇 M、簇 N 在公共时间窗 inter_span（span_1,span_2）内的成员交集。

公式 5-6 公共时间窗内的簇与簇的成员交集

$$
\begin{aligned}
&\mathrm{int}er\,\mathrm{sec}tion\_member\_in\_\mathrm{int}er\_span(M,N)\\
&=\{member_1,member_2,\cdots,member_n\},if\\
&member_i\in member\_in\_\mathrm{int}er\_span(M)\\
&member_i\in member\_in\_\mathrm{int}er\_span(N)\\
&timeAttribute(M)=span\_1\\
&timeAttribute(N)=span\_2\\
&\exists span\_1,span\_2,\mathrm{int}er\_span(span\_1,span\_2)
\end{aligned}
$$

从图 5-6 可以看到，公共时间窗内簇 M 与簇 N 的成员有交集，用菱形表示。值得注意的是，公共时间窗内的簇与簇的成员交集实际上就是两个簇中相同成员的集合。

簇特性的具体定义为：

（1）祖先的遗传率

定义 5-7 祖先的遗传率

某时间窗内的簇 N，如果在簇 N 的祖先层有一个簇 M 是簇 N 的祖先，那么簇 M、簇 N 在公共时间窗内的成员交集数与簇 M 的在公共时间窗内的成员数之比，称作祖先簇 M 对簇 N 的遗传率。

公式 5-7 祖先的遗传率

$$
\begin{aligned}
&heredity(M,N)\\
&=\frac{num\_\mathrm{int}er\,\mathrm{sec}tion\_member\_in\_\mathrm{int}er\_span(M,N)}{num\_member\_in\_\mathrm{int}er\_span(M)}
\end{aligned}
$$

其中

$M\in CLUSTER\_RESULT\_LEVEL\_ANCESTOR$

$N\in CLUSTER\_RESULT\_LEVEL\_ME$

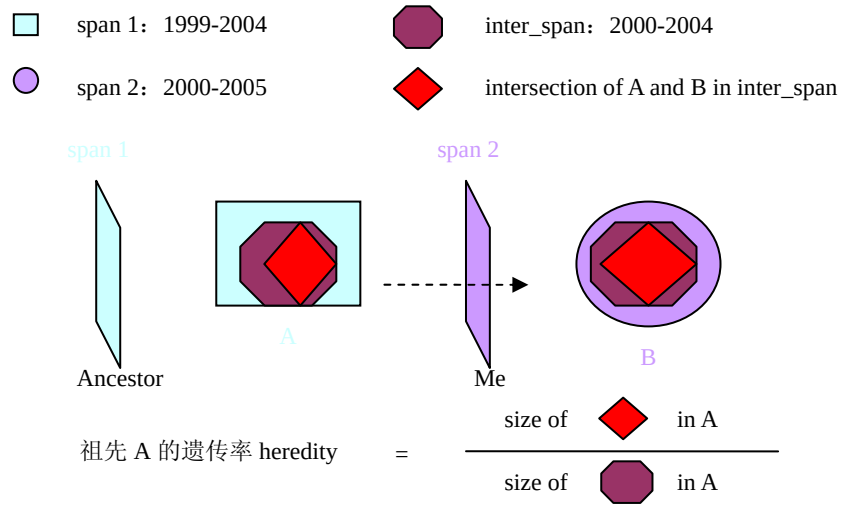


图 5-7 祖先的遗传率

（2）有效祖先

数个祖先的遗传率有大有小，各不相同。遗传率太小的祖先，对后代的影响很小。因此有必要从所有祖先中挑选出有效祖先。

定义 5-8 有效祖先

某时间窗内的簇 N，如果在祖先层有一个簇 M 是簇 N 的祖先，而且簇 M 的遗传率大于或者等于预先设定的最小遗传率，那么称作簇 M 是簇 N 的有效祖先。

公式 5-8 有效祖先

$$realAncestors(N)=(M_1,M_2,\cdots,M_n),if$$
$$heredity(M_i,N)\geq MIN\_HEREDITY\_RATE$$

其中

$$M_i\in CLUSTER\_RESULT\_LEVEL\_ANCESTOR$$
$$M_i\in ancestors(N)$$
$$N\in CLUSTER\_RESULT\_LEVEL\_ME$$
$$i=1,2,\cdots,n$$

从有效祖先可以进一步引申出有效祖先数，即有效祖先的个数。

（3）后代的继承率

定义 5-9 后代的继承率

某时间窗内的簇 P，如果在簇 P 的后代层有一个簇 Q 是簇 P 的后代，那么簇 P、簇 Q 在公共时间窗内的成员交集数与簇 Q 的在公共时间窗内的成员数之比，称作后代簇 Q 对簇 P 的继承率。

公式 5-9 后代的继承率

$$inherit(P,Q) = \frac{num\_inter section\_member\_in\_inter\_span(P,Q)}{num\_member\_in\_inter\_span(Q)}$$

其中

$P \in CLUSTER\_RESULT\_LEVEL\_ME$

$Q \in CLUSTER\_RESULT\_LEVEL\_OFFSPRING$

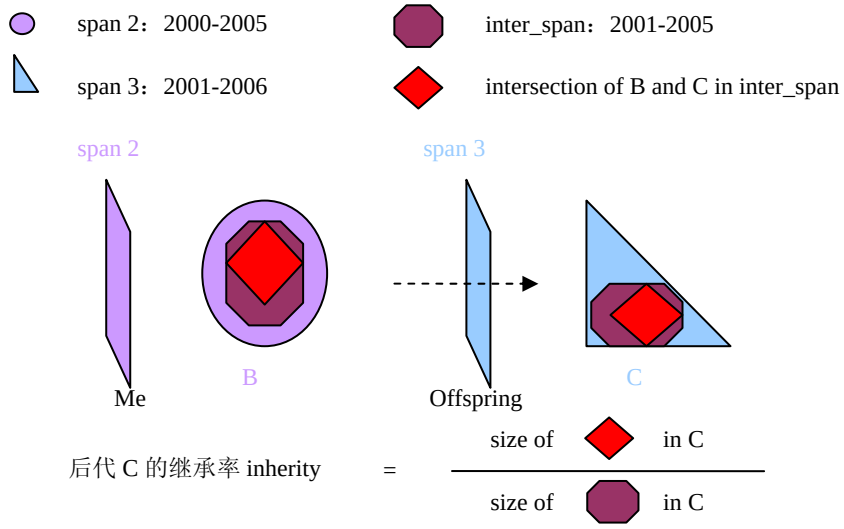


图 5-8 后代的继承率

（4）有效后代

数个后代的继承率有大有小，各不相同。继承率太小的后代，从祖先那里承接的特征太少。因此有必要从所有后代中挑选出有效后代。

定义 5-10 有效后代

某时间窗内的簇 P，如果在后代层有一个簇 Q 是簇 P 的后代，而且簇 Q 的继承率大于或者等于预先设定的最小继承率，那么称作簇 Q 是簇 P 的有效后代。

公式 5-10 有效后代

$$realOffsprings(P) = (Q_1, Q_2, \cdots, Q_p), if$$

$$inherit(P, Q_i) \geq MIN\_INHERIT\_RATE$$

其中

$P \in CLUSTER\_RESULT\_LEVEL\_ME$

$Q_i \in CLUSTER\_RESULT\_LEVEL\_OFFSPRING$

$Q_i \in offsprings(P)$

从有效后代可以进一步引申出有效后代数，即有效后代的个数。图 5-15 中，簇 G 有两个有效后代。

（5）汇聚数

定义 5-11 汇聚数

簇 N 如果在祖先层中有 n 有效祖先，那么 n 称作簇 N 的汇聚数。

公式 5-11 汇聚数

$$\inf\ lux\_num(N) = Num(realAncestors(N))$$

其中

$$N \in CLUSTER\_RESULT\_LEVEL\_ME$$

从定义可以看到，汇聚数，即有效祖先数。

（6）汇聚率

定义 5-12 汇聚率

簇 N 的汇聚数与簇 N 的祖先数之比，称作簇 N 的汇聚率。

公式 5-12 汇聚率

$$\inf\ lux(N) = \frac{Num(realAncestors(N))}{Num(ancestors(N))}$$

其中

$$N \in CLUSTER\_RESULT\_LEVEL\_ME$$

（7）分解数

定义 5-13 分解数

簇 P 如果在后代层有中有 p 个有效后代，那么 p 称作簇 P 的分解数。

公式 5-13 分解数

$$branch\_num(P) = Num(realOffsprings(P))$$

其中

$$P \in CLUSTER\_RESULT\_LEVEL\_ME$$

从定义可以看到，分解数，即有效后代数。

（8）分解率

定义 5-14 分解率

簇 P 的分解数与簇 P 的后代数之比，称作簇 P 的分解率。

公式 5-14 分解率

$$branch(P) = \frac{Num(realOffsprings(P))}{Num(offsprings(P))}$$

其中

$$P \in CLUSTER\_RESULT\_LEVEL\_ME$$

（9）本质性

定义 5-15 本质性（相对于祖先）

某时间窗内的簇 N，如果该时间窗与祖先层时间窗的公共时间窗是 inter_span，那么簇 N 在公共时间窗内的成员数与簇 N 的全部成员数之比，称作簇 N 相对于祖先的本质性。

公式 5-15 本质性（相对于祖先）

$$quality(N) = \frac{num\_member\_in\_inter\_span(N)}{Num(N)}$$

其中

$$N \in CLUSTER\_RESULT\_LEVEL\_ME$$

$$\exists span\_1, span\_2, inter\_span(span\_1, span\_2)$$

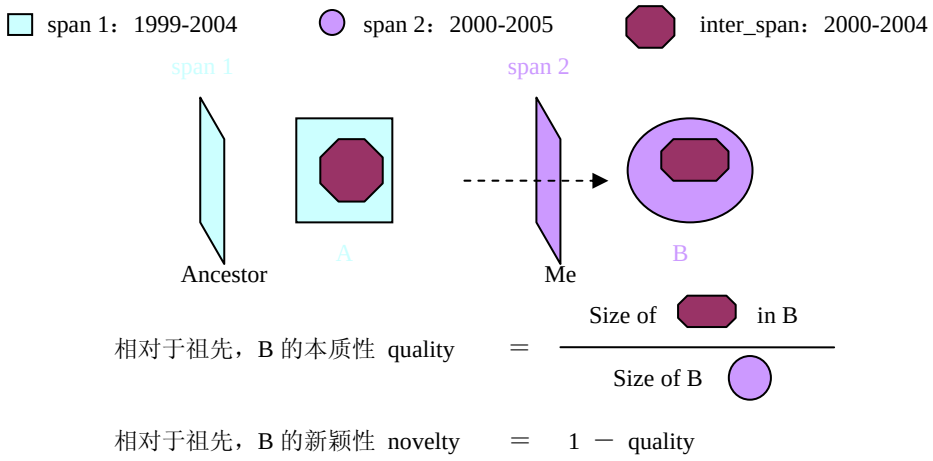


图 5-9 本质性（相对于祖先）

定义 5-16 本质性（相对于后代）

某时间窗内的簇 P，如果该时间窗与后代层时间窗的公共时间窗是 inter_span，那么簇 P 在公共时间窗内的成员数与簇 P 的全部成员数之比，称作簇 N 相对于后代的本质性。

公式 5-16 本质性（相对于后代）

$$quality(P) = \frac{num\_member\_in\_inter\_span(P)}{Num(P)}$$

其中

$$P \in CLUSTER\_RESULT\_LEVEL\_ME$$

$$\exists span\_2, span\_3, inter\_span(span\_2, span\_3)$$

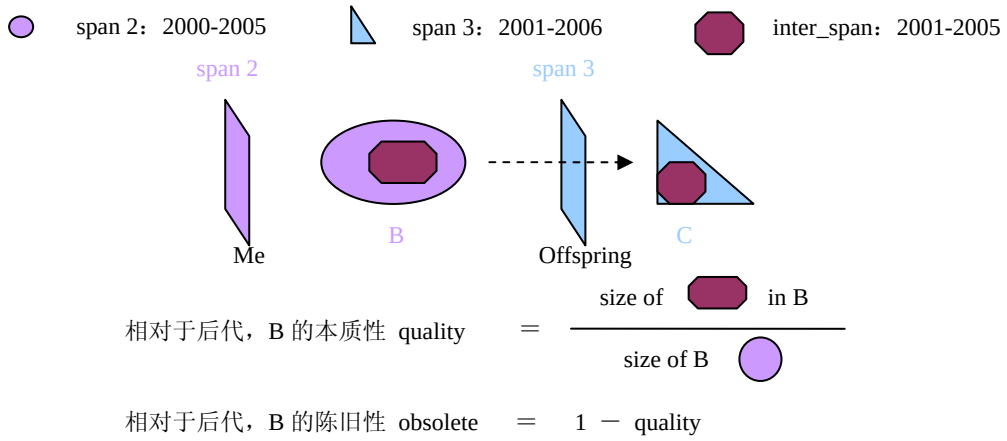


图 5-10 本质性（相对于后代）

从定义可以看出，不论是相对于祖先还是相对于后代，本质性都是反映某个簇与祖先或者后代相似的可能性。如果在公共时间窗中某个簇有较多的成员，那么该簇与祖先或者后代相似的可能性就大。相反亦然。

（10）新颖性

定义 5-17 新颖性

某时间窗内的簇 N，如果簇 N 相对于祖先的本质性是 quality，那么 1－qualit 称作簇 N 相对于祖先的新颖性。

公式 5-17 新颖性

$$novelty(N)=1-quality(N)$$

其中

$$N \in CLUSTER\_RESULT\_LEVEL\_ME$$

$$\exists span\_1, span\_2, \text{inter}\_span(span\_1, span\_2)$$

相对于祖先，某个簇如果具有较大的本质性，那么新颖性 novolty=1-quality 就会较小，表明该簇对于祖先没有太多的新成员。反之，表明该簇对于祖先有很多新成员，而且是在时间窗后期新增的新成员。

（11）陈旧性

定义 5-18 陈旧性

某时间窗内的簇 P，如果簇 P 相对于后代的本质性是 quality，那么 1－qualit 称作簇 P 相对于后代的陈旧性。

公式 5-18 陈旧性

$$obsolete(P)=1-quality(P)$$

其中

$$P \in CLUSTER\_RESULT\_LEVEL\_ME$$

$$\exists span\_2, span\_3, \text{inter}\_span(span\_2, span\_3)$$

相对于后代，某个簇如果具有较大的本质性，那么陈旧性 obsolete=1-quality 就会较小，表明该簇对于后代没有位于太多时间窗前期的成员。反之，表明该簇对于后代有很多位于时间窗前期不会遗传到后代的老成员。

5.3.1.3 6种演变簇

根据簇特性，本文将把中间时间窗聚类结果中的演变簇归纳为 6 类，包括合并簇、分化簇、融合簇、扩散簇、新增簇和消失簇。详细定义和描述如下。

（1）合并簇

定义 5-19 合并簇

某时间窗内的簇 M，如果 M 的簇内成员数大于或者等于预先设定的最小成员数，而且相对于祖先，M 的本质性大于或者等于最小本质性，汇聚数大于或者等于最小汇聚数，汇聚率大于或者等于最小汇聚率，那么簇 M 称作合并簇。

公式 5-19 合并簇

$$M\text{是合并簇，}\ if$$

$$Num(M) \geq MIN\_SIZE\_FROM\_MERGE$$

$$quality(M) \geq MIN\_QUALITY\_RATE$$

$$\inf\ lux\_num(M) \geq MIN\_INFLUX\_NUM$$

$$\inf\ lux(M) \geq MIN\_INFLUX\_RATE$$

其中

$$M \in CLUSTER\_RESULT\_LEVEL\_ME$$

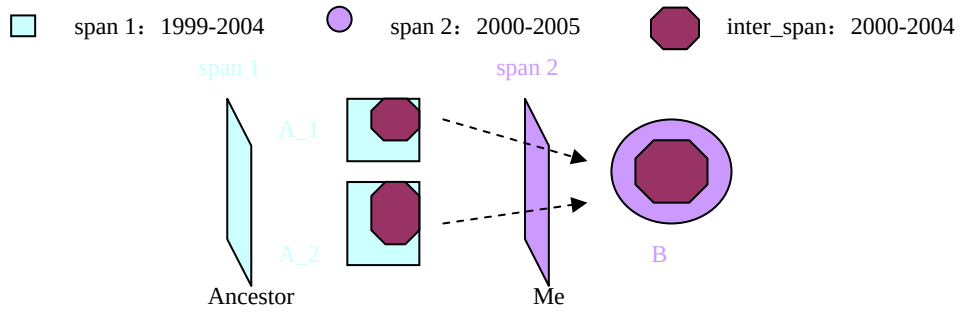


图 5-11 合并簇

（2）分化簇

定义 5-20 分化簇

某时间窗内的簇 M，如果 M 的簇内成员数大于或者等于最小成员数，而且相对于后代，M 的本质性大于或者等于最小本质性，分解数大于或者等于最小分解数，分解率大于或者等于最小分解率，那么簇 M 称作分化簇。

公式 5-20 分化簇

$$
\begin{aligned}
&M\text{是分化簇，}\ if\\
&Num(M)\geq MIN\_SIZE\_TO\_DIVIDE\\
&quality(M)\geq MIN\_QUALITY\_RATE\\
&branch\_num(M)\geq MIN\_BRANCH\_NUM\\
&branch(M)\geq MIN\_BRANCH\_RATE\\
&\text{其中}\\
&M\in CLUSTER\_RESULT\_LEVEL\_ME
\end{aligned}
$$

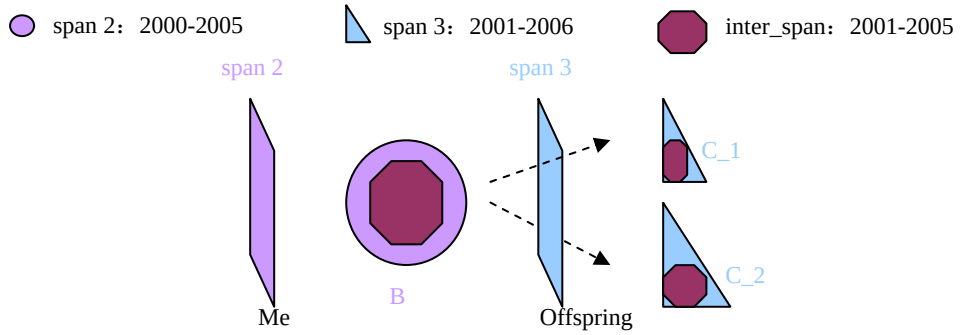


图 5-12 分化簇

（3）融合簇

定义 5-21 融合簇

某时间窗内的簇 M，如果 M 的簇内成员数大于或者等于最小成员数，而且相对于祖先，M 的本质性小于最小本质性，新颖性大于或者等于最小新颖性，汇聚数大于或者等于 1，那么簇 M 称作融合簇。

公式 5-21 融合簇

M是融合簇，if

$$Num(M) \geq MIN\_SIZE\_FROM\_FUSE$$

$$quality(M) < MIN\_QUALITY\_RATE$$

$$novelty(M) \geq MIN\_NOVELTY\_RATE$$

$$\inf\ lux\_num(M) \geq 1$$

其中

$$M \in CLUSTER\_RESULT\_LEVEL\_ME$$

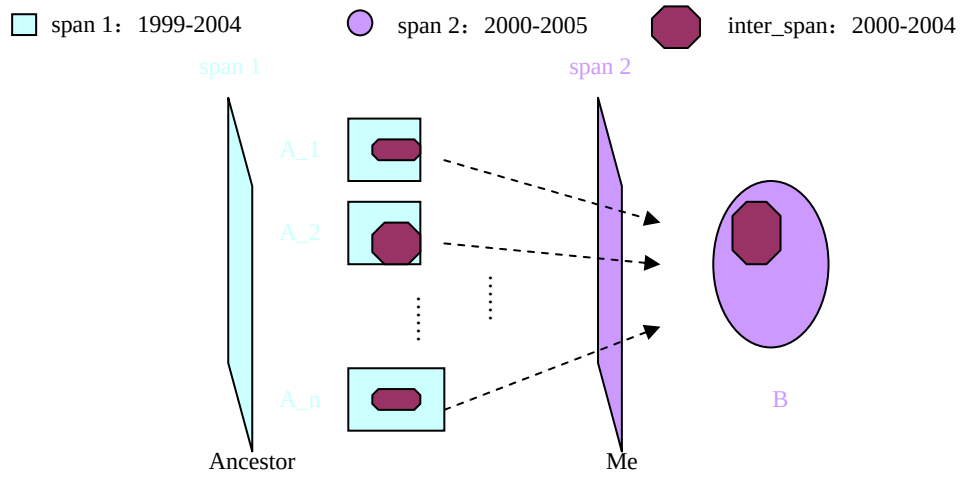


图 5-13 融合簇

（4）扩散簇

定义 5-22 扩散簇

某时间窗内的簇 M，如果 M 的簇内成员数大于或者等于最小成员数，而且相对于后代，M 的本质性小于最小本质性，陈旧性大于或者等于最小陈旧性，分解数大于或者等于 1，那么簇 M 称作扩散簇。

公式 5-22 扩散簇

M是扩散簇，if

$$Num(M) \geq MIN\_SIZE\_TO\_DIFFUSE$$

$$quality(M) < MIN\_QUALITY\_RATE$$

$$obsolete(M) \geq MIN\_OBSOLETE\_RATE$$

$$branch\_num(M) \geq 1$$

其中

$$M \in CLUSTER\_RESULT\_LEVEL\_ME$$

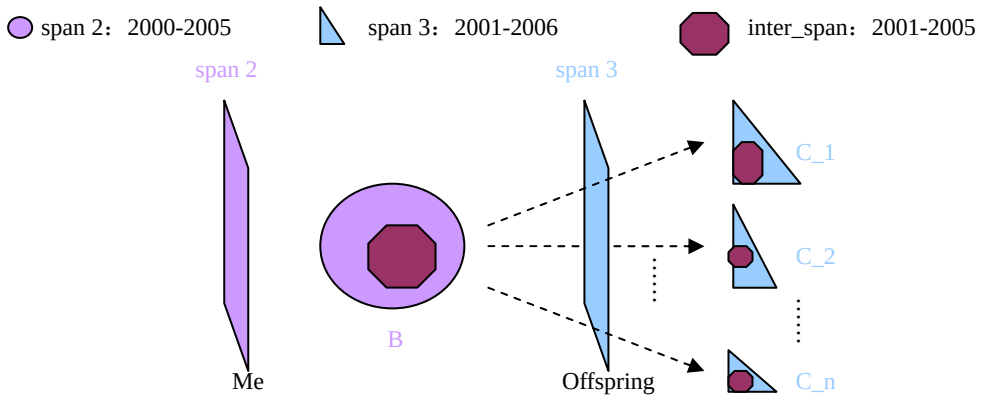


图 5-14 扩散簇

（5）新增簇

定义 5-23 新增簇

某时间窗内的簇 M，如果 M 的簇内成员数大于或者等于最小成员数，祖先数等于 0，后代数大于 0，而且相对于祖先，M 的新颖性大于或者等于最小新颖性，那么簇 M 称作新增簇。

公式 5-23 新增簇

$$
\begin{aligned}
&M\text{是新增簇，}\ if\\
&Num(M)\geq MIN\_SIZE\_FROM\_\text{EMERGE}\\
&ancestors\_num(M)=0\\
&offsprings\_num(M)>0\\
&novelty(M)\geq MIN\_NOVELTY\_RATE\\
&\text{其中}\\
&M\in CLUSTER\_RESULT\_LEVEL\_ME
\end{aligned}
$$

（6）消失簇

定义 5-24 消失簇

某时间窗内的簇 M，如果 M 的簇内成员数大于或者等于最小成员数，祖先数大于 0，后代数等于 0，而且相对于后代，M 的陈旧性大于或者等于最小陈旧性，那么簇 M 称作消失簇。

公式 5-24 消失簇

$$
\begin{aligned}
&M\text{是消失簇，}\ if\\
&Num(M)\geq MIN\_SIZE\_TO\_OBSOLETE\\
&ancestors\_num(M)>0\\
&offsprings\_num(M)=0\\
&obsolete(M)\geq MIN\_OBSOLETE\_RATE\\
&\text{其中}\\
&M\in CLUSTER\_RESULT\_LEVEL\_ME
\end{aligned}
$$

综上所述，为了实现划分时间窗寻找演变结构的目的，先归纳出簇与簇之间的两种关系——ancestor 关系和 offspring 关系。另外利用两种关系，分析出簇与

演变有关的几种特性。在此基础上，根据与祖先、与后代的关系，将演变簇划分为两组。一组是相对于祖先而言，由祖先演变而来，包括合并簇、融合簇、新增簇。另一组是相对于后代而言，将演变成后代，包括分化簇、扩散簇、消失簇。根据本文的定义，聚类结果中簇的种类划分体系如图 5-15 所示，各种簇的区别如表 5-1 所示。

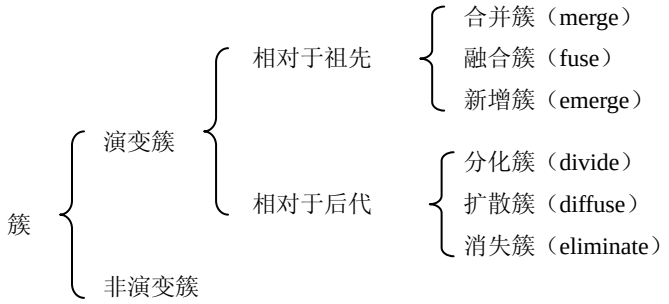


图 5-15 簇及演变簇的种类划分

表 5-1 时间策略中各种簇之间的区别

簇种类	区别	
演变簇	是否属于合并、融合、新增、分化、扩散、消失中的一类	
非演变簇		
合并簇	本质性是否足够大 新颖性是否足够大 汇聚率是否足够大 汇聚数是否有限制条件	是否有祖先
融合簇		
新增簇		
分化簇	本质性是否足够大 陈旧性是否足够大 分解率是否足够大 分解数是否有限制条件	是否有后代
扩散簇		
消失簇		

5.3.2 演变结构的发现

通过前面的分析，从簇的 ancestor 和 offspring 关系出发，定义了时间策略最需要关注的 6 类演变簇，以及演变簇的演变特性。应用时间策略发现演变结构的主要思路就是在任意三个相邻的时间窗聚类结果之间构建 ancestor 和 offspring 关

系，将三者联系在一起，在此基础上，发现中间时间窗中蕴涵的演变结构。这个过程的核心问题就是相对于前一时间窗和后一时间窗在中间时间窗中寻找到可能存在的演变簇及其演变关系。

5.3.2.1 发现演变结构的主要步骤

时间策略实际上是通过划分时间窗，从同一数据集中挑选时间属性位于时间窗的待分析对象进行多次聚类计算。然后从所有聚类结果中依次挑选出相邻的三个时间窗，为中间时间窗构建与前一时间窗的 ancestor 关系、与后一时间窗的 offspring 关系。在此基础上，以中间时间窗聚类生成的簇为考察的主体对象，根据演变簇的判断标准寻找中间时间窗相对于前一时间窗的合并簇、融合簇、新增簇，以及相对于后一时间窗的分化簇、扩散簇、消失簇。图 5-16 描述的是应用时间策略，在三个相邻时间窗的聚类结果中发现演变结构的主要步骤。

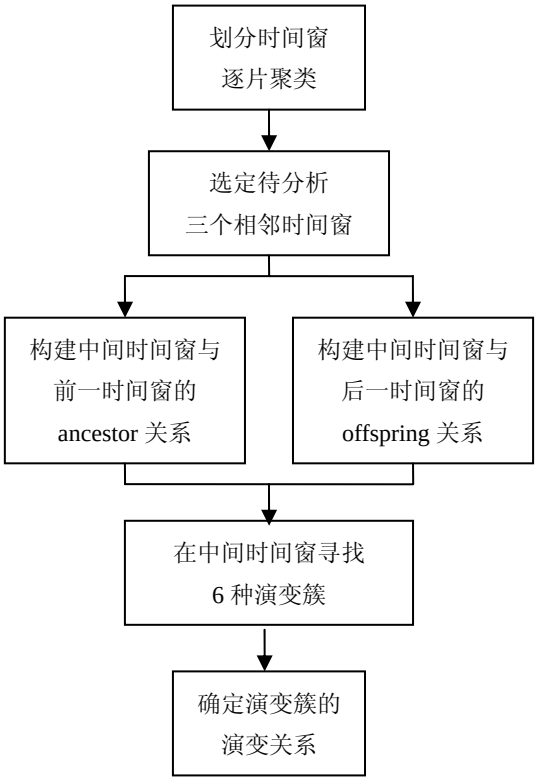


图 5-16 发现演变结构的主要步骤

这里，值得一提的是，虽然与阈值策略类似，时间策略也是在多个聚类结果中寻找特殊的簇，但是阈值策略需要在寻找潜在结构之前确定“突变阈值区间”，从该区间中选定差异明显的两个阈值层，而时间策略则不需要这个步骤。分析原因，有下面几点：

（1）通常划分时间窗的时候，为了使前后时间窗有各自独有的特点，一般会使公共时间窗不占前后时间窗太大的比例。如果公共时间窗占的比例过大，那么因为公共成员太多，导致前后两个时间窗的聚类结果基本没有变化。如果公共

时间窗占的比例适当，那么因为非公共成员占有相当大的比例，自然会使前后两个时间窗的聚类结果差异明显。这样一来，从差异明显的两个时间窗中发现演变结构的可能性大大增加。

（2）通常划分的时间窗数量不会太多，一般十到二十个，不像阈值策略那样有可能产生数十个甚至上百个阈值层。这样，应用时间策略发现演变结构的过程不会太漫长，计算量也不如阈值策略那么巨大。因此，可以不用考虑如何提高计算效率的问题。

5.3.2.2 寻找演变簇及其演变关系的算法流程

（1）构建 ancestor 关系和 offspring 关系。构建 ancestor 关系，详细算法流程如图 5-17 所示。构建 offspring 关系的详细算法流程与此类似。

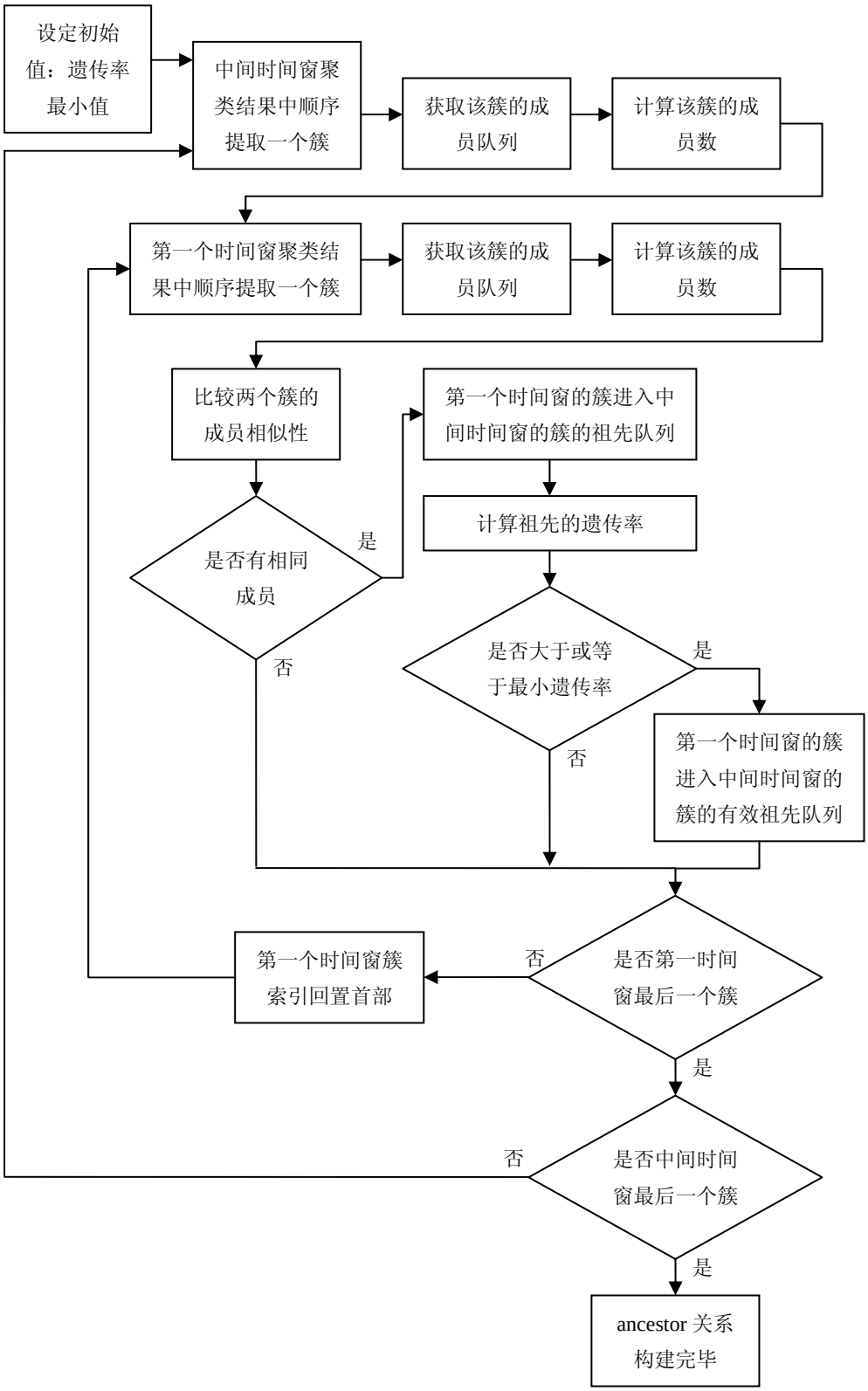


图 5-17 构建 ancestor 关系的算法流程

（2）在中间时间窗寻找演变簇。详细算法流程如图 5-18 所示。

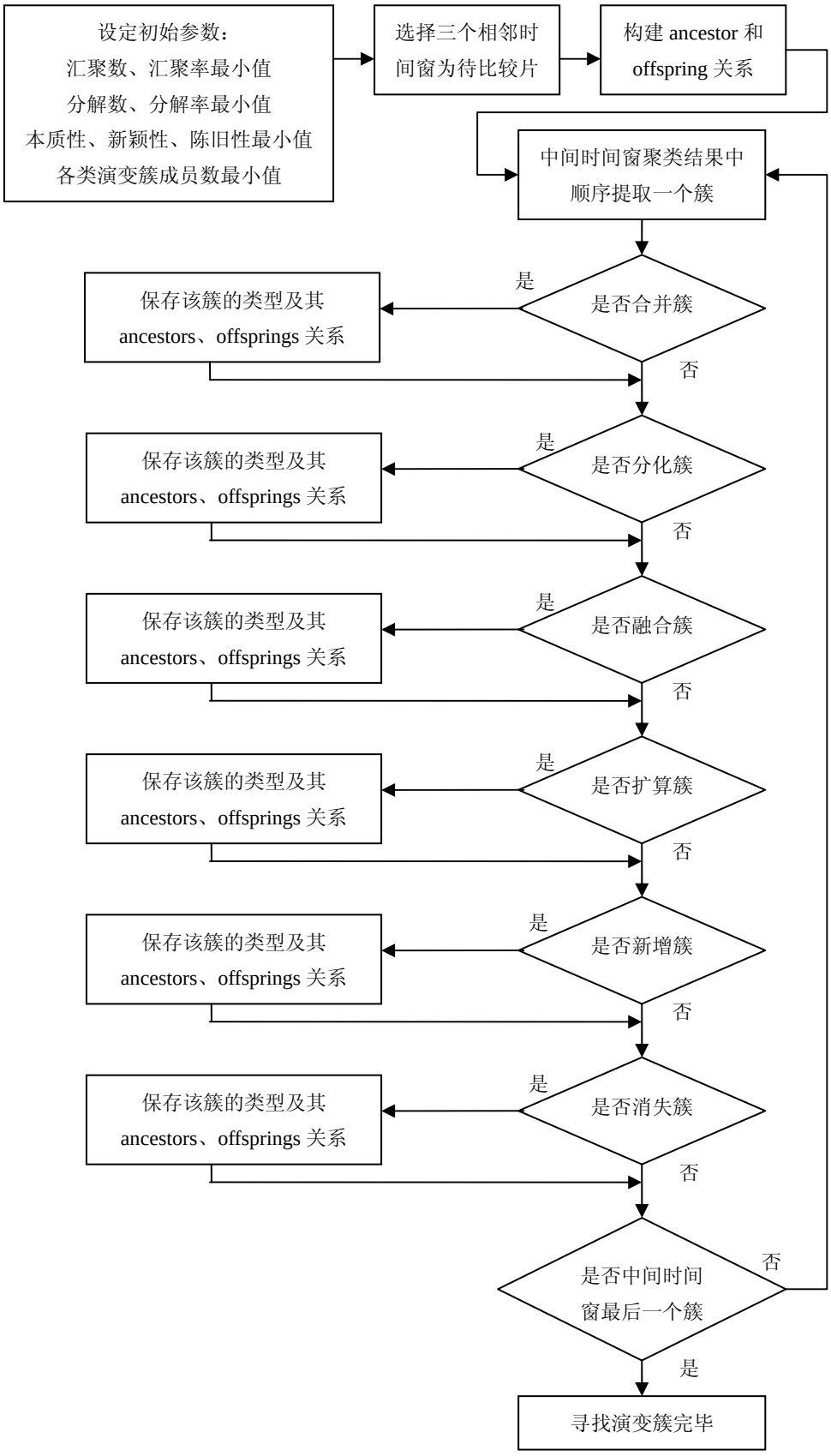


图 5-18 寻找演变簇的算法流程

在寻找演变簇的过程中，核心步骤是判断一个簇属于哪一种类型的演变簇。

以判断合并簇为例，详细算法流程如图 5-19 所示。其它演变簇的判断类似。

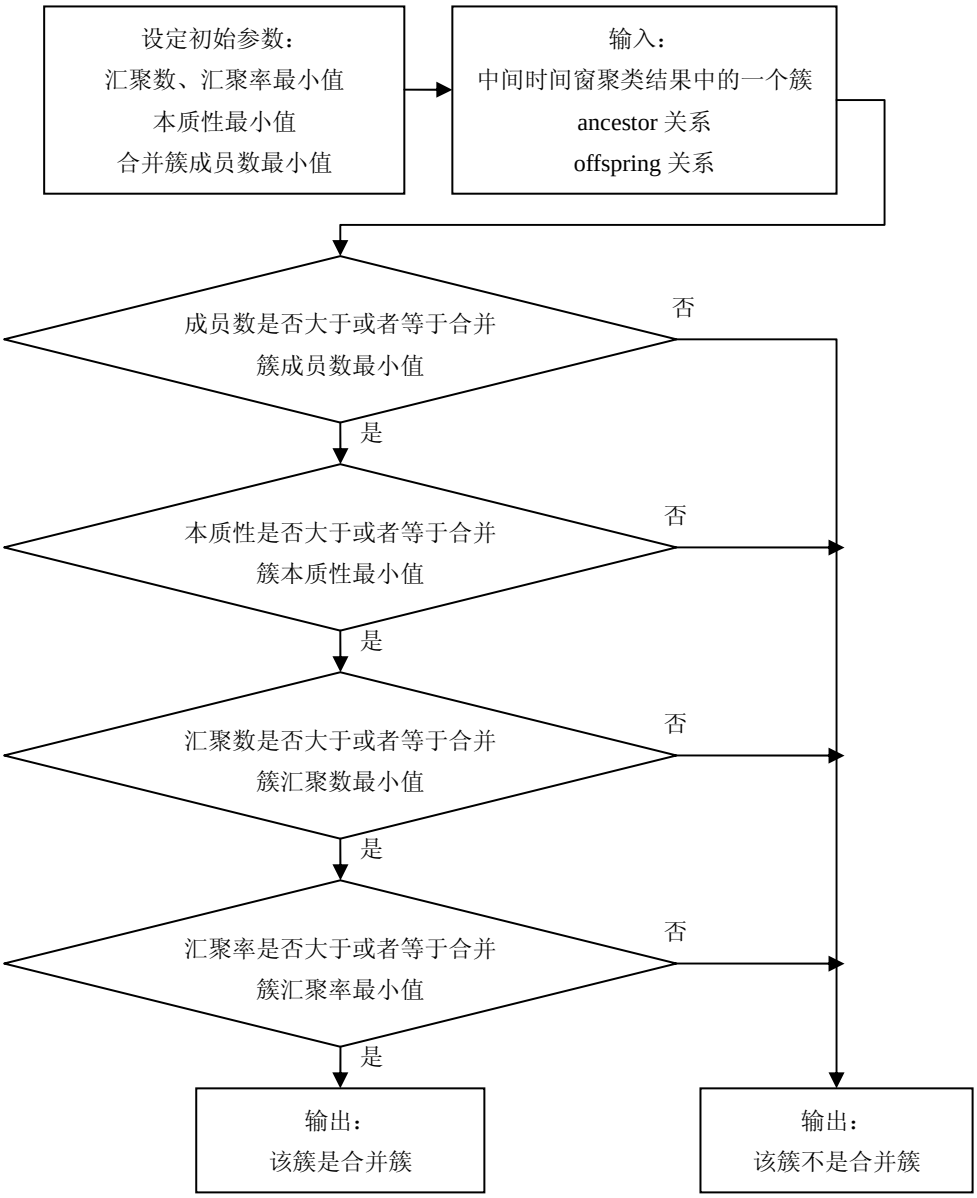


图 5-19 判断合并簇的算法流程

（3）同阈值策略一样，在中间时间窗寻找相对于前后两个时间窗的演变结构，有两个主要任务。一个是在中间时间窗寻找 6 种演变簇，另一个任务是确定这些演变簇同前后两个时间窗簇的演变关系。演变关系将由两个部分组成：a、对祖先的继承关系。相对于前一个时间窗，该簇继承了祖先的某些成员，因此可以将该簇对有效祖先的继承率视作该簇对祖先的演变关系。b、对后代的遗传关系。相对于后一个时间窗，该簇遗传某些成员给后代，因此可以将该簇对有效后代的遗传率视作该簇对后代的演变关系。构建演化关系的主要目的是通过演化关系将中间时间窗中的演化簇与前后时间窗中的簇联系起来，用演化链描述三个相邻时间窗的演化过程，用遗传率、继承率描述演化强度。如图 5-20 所示。

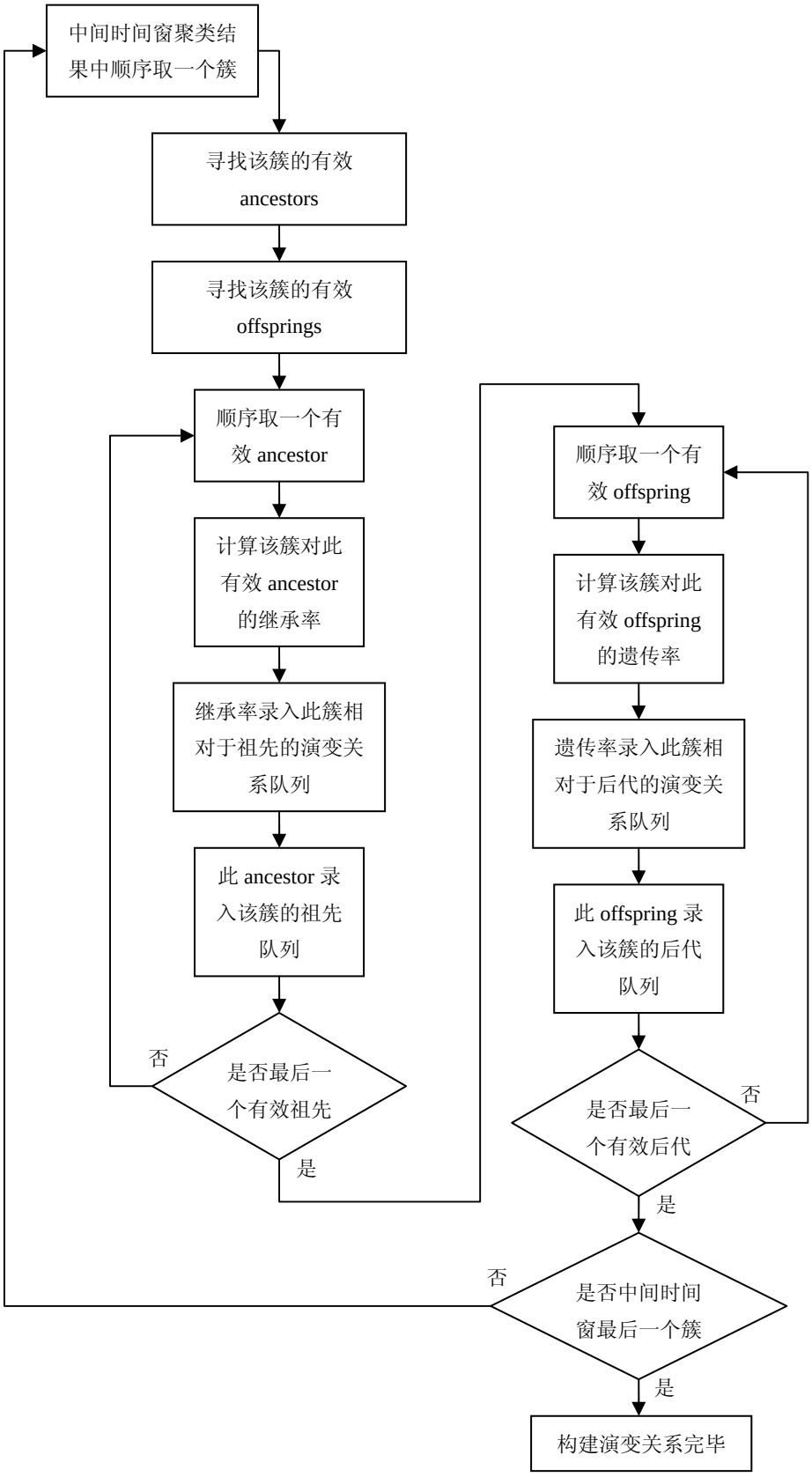


图 5-20 构建演化关系的算法流程

5.3.3 演变结构的展示

与阈值策略不同的是，时间策略寻找演变结构的过程虽然只有一种方式，但是时间策略演变结构的展示也有两种方式。一种方式是以一种简单的形式分别将三个时间窗的所有簇排列起来，然后用连线将具有演变关系的簇联系起来。另一种方式是图形上只展现中间时间窗的簇结构，另外用文字注释哪些簇是演变簇，这些演变簇有怎样的演变关系。

针对第一种方式，可视化展示的内容是：

（1）展示三个时间窗生成的簇。

（2）展示从中间时间窗中寻找的演变簇。

（3）展示中间时间窗的演变簇同前后两个时间窗簇的演变关系。

针对第二种方式，可视化展示的内容是：

（1）展示中间时间窗生成的簇及其关系。

（2）标示中间时间窗的演变簇同前后两个时间窗簇的演变关系。

两种展示方式各有特点。第一种方式能清晰地展示出三个相邻时间窗演变关系的全貌，但是不能反映每一个时间窗的簇结构，而簇结构的变化也是多个时间窗之间演变的一个具体体现。第二种方式可以展示某个时间窗的簇结构，在此基础上用文字标识该时间窗的演变簇和演变关系，但是不能直观反映相邻时间窗演变关系的全貌。

不论采用哪一种方式，寻找演变结构的主体过程没有变化。通过上面的分析，归纳演变结构展示的具体步骤为：

第一步：按顺序依次获取三个相邻时间窗及其聚类结果。

第二步：获取中间时间窗的演变簇和演变关系。

第三步：选择展示方式。

第四步：判断是哪一种展示方式。

第五步：如果是第一种方式，将三个时间窗的所有簇分别排成三列，用不同形式的直线连接三个时间窗中有演变关系的簇；如果是第二种方式，图形中只展示中间时间窗的簇及其簇结构，而不展示前后两个时间窗的簇及其簇结构，而且演变关系只用文字加以标示。

第六步：返回第一步，遍历所有时间窗，直至达到整个时间段的终点为止。

5.4 小结

本章的主要目标是通过时间策略，在某一时间窗上发现演变结构，侧重定量计算演变关系以及演变关系的强弱。时间策略的主要内容是：

（1）构建时间窗之间簇与簇的关系。站在中间时间窗的位置上，如果向前

看，有 ancestor 关系，如果向后看，有 offspring 关系。ancestor 关系和 offspring 关系将是后面发现演变结构的基础。另外，通过对 ancestor 关系和 offspring 关系的定量分析，可以确定时间窗之间存在强烈的“遗传继承”性，这正是阈值策略的研究起点。

（2）探讨演变簇的种类。利用 ancestor 关系和 offspring 关系，先定义了簇的一些特性，如遗传率、继承率、有效祖先数、有效后代数、本质性、新颖性、陈旧性、汇聚率、分解率等。利用这些演变特性，进而将演变簇划分为 6 个种类并给出定义，它们分别是：合并簇、分化簇、融合簇、扩散簇、新增簇、消失。

（3）设计时间策略实施的步骤。主体步骤包括时间窗之间 ancestor 关系和 offspring 关系的构建、6 种演变簇的寻找、演变关系的确定。本文一般取三个相邻的时间窗作为分析对象，而且中间时间窗是主体分析对象。相对于前一时间窗，在中间时间窗寻找合并簇、融合簇、新增簇，相对于后一时间窗，在中间时间窗寻找分化簇、扩散簇、消失簇。

（4）最后一部分内容是演变结构的可视化展示。同潜在结构展示一样，演变结构展示也会采用两种不同的方式，一个是只显示一个时间窗，另一个是显示三个时间窗。

6 方法实现和结果展示

本文第四章、第五章分别详细设计了深度探寻知识演化结构的两种策略。本论文的另一个重要任务是设计一个试验系统，实现这两种方法策略。通过构建一个集成应用平台，形成一个完整的应用流程。本章节将在两个方法策略设计的基础上，主要讨论两种策略作为两个主模块以及主模块中的若干子模块在试验系统中的实现。最后展示试验系统的运行结果。

6.1 阈值策略的实现

6.1.1 阈值策略主模块的实现

阈值策略要实现的主要功能是从同一数据集中按逐步降低的关系阈值挑选待聚类的对象，并多个阈值层次的对象集合分别进行聚类计算。然后对不同阈值层面的聚类结果进行比较，找到聚类结果存在较明显差异的阈值层次，即“突变阈值区间”。从突变区间中选择一个高阈值层和一个低阈值层，应用判断标准计算低阈值层中簇与簇的 relative 关系，低阈值层与高阈值层之间簇与簇的 twin 关系。在此基础上，再根据潜在簇的判断标准寻找低阈值层相对于高阈值层的各种潜在簇。最后用可视化方法展示寻找到的潜在结构。根据阈值策略的功能要求，阈值策略主模块的主要实现方法如图 6-1 所示。

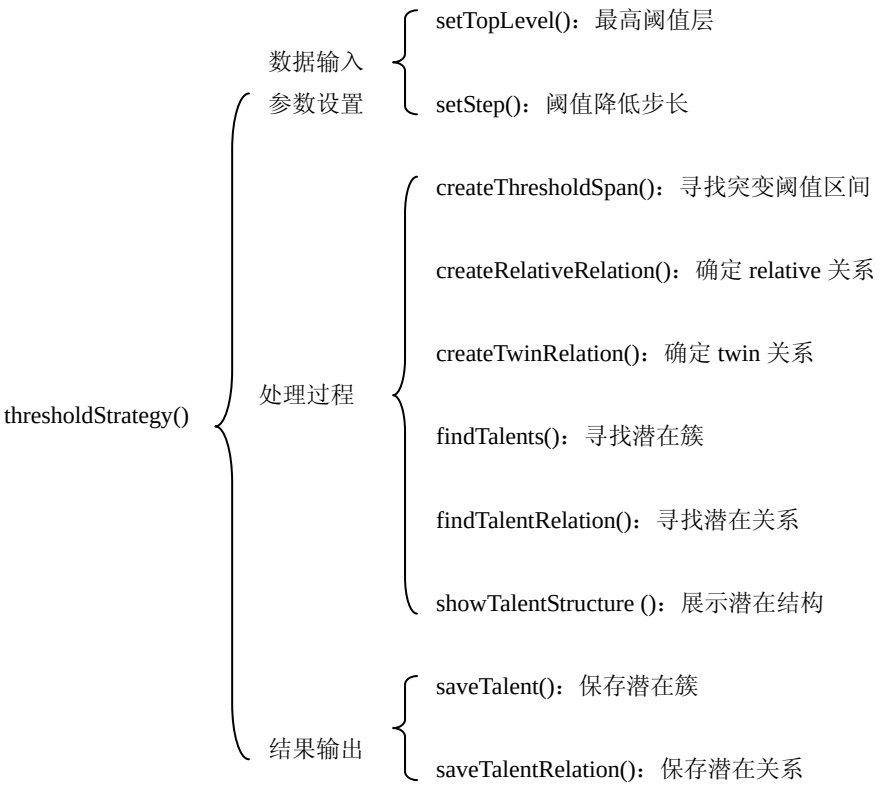
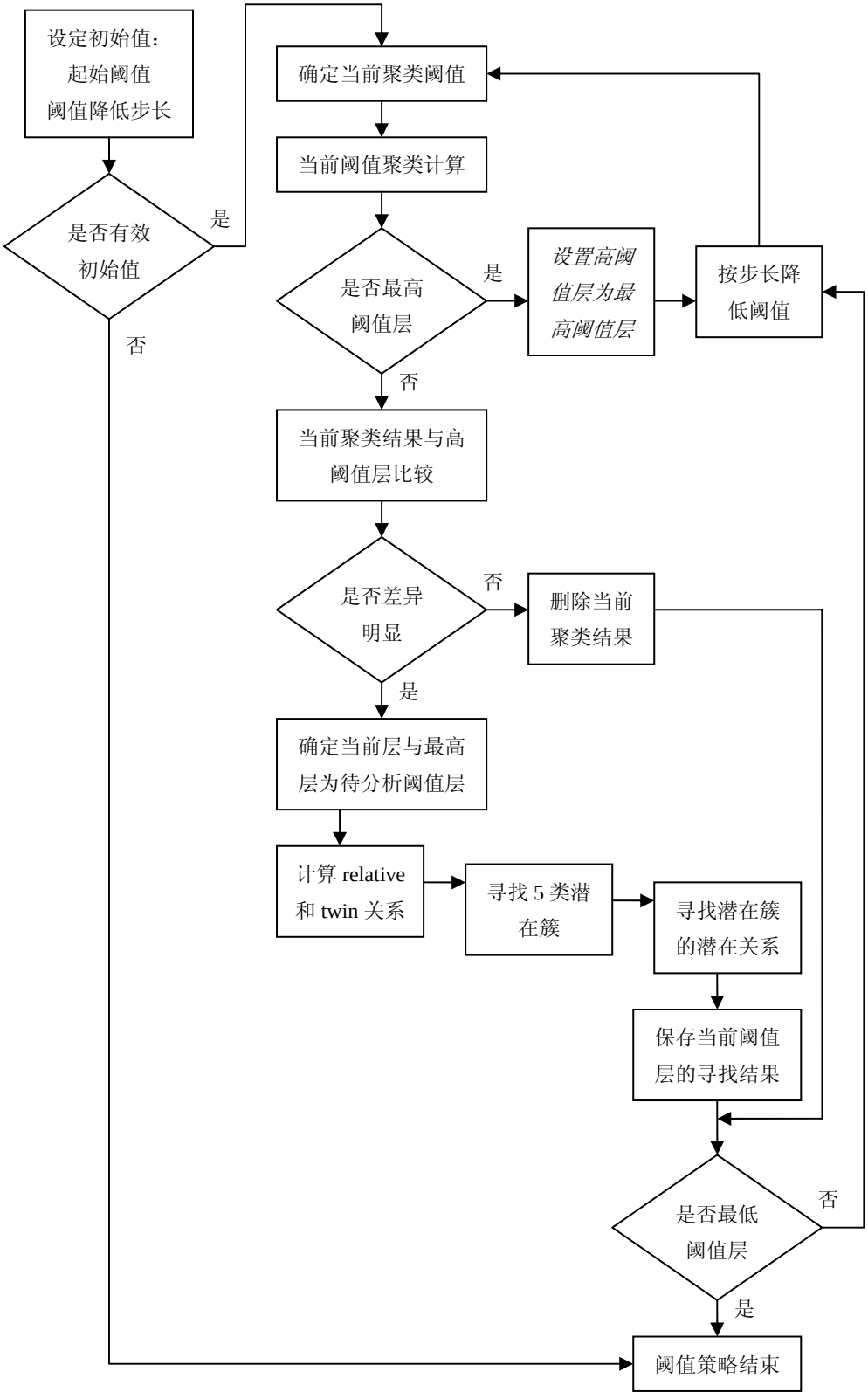
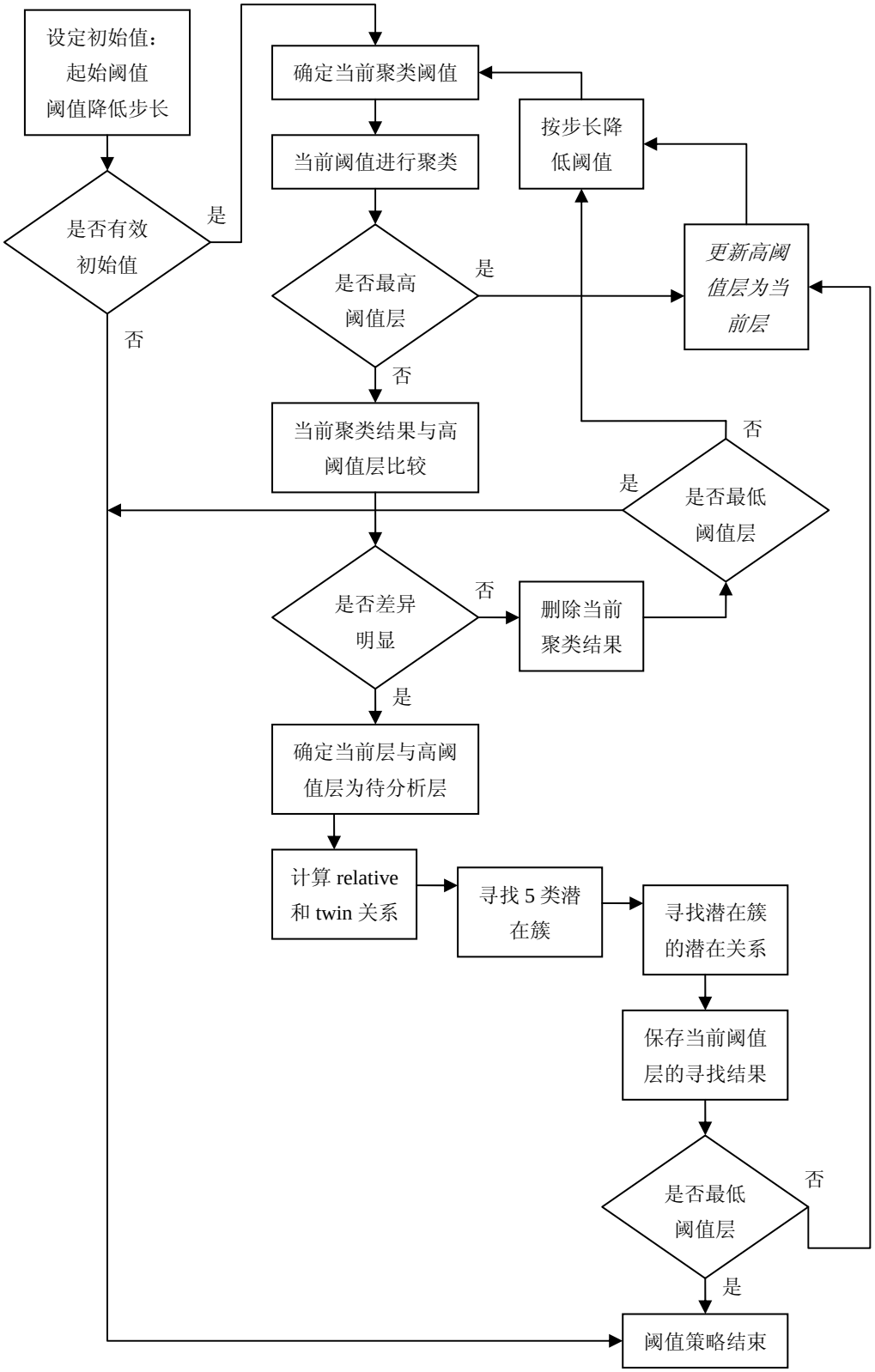


图 6-1 阈值策略主模块的主要实现方法

图 6-2 展现了应用阈值策略，发现潜在结构的详细实现流程。值得一提的是，高、低两个阈值层因为突变阈值区间的两种计算方式，使阈值策略有两种不同的实现方式。图（a）描述的是高阈值层固定设置为最高阈值层，低阈值层逐层降低。图（b）描述的是高阈值层变动设置为差异明显的当前低阈值层，即高、低两个阈值层按顺序依次被设置为两个相邻的差异明显的阈值层。这两种方式主要影响了高、低两个阈值层的选择，而不影响后续潜在簇的寻找过程。



（a）高阈值层固定设置为最高阈值层



（b）高阈值层变动设置为差异明显的当前低阈值层

图 6-2 应用阈值策略，发现潜在结构的实现流程

6.1.2 阈值策略子模块的实现

（1）两个阈值层面聚类结果差异比较子模块的实现

两个阈值层面聚类结果差异比较的主要功能是想通过两个层次聚类结果的比较，判断两个阈值层的差异是否明显，以此为依据，确定此次阈值策略的突变阈值空间。根据阈值策略的功能要求，两个阈值层面聚类结果差异比较子模块的主要实现方法如图 6-3 所示。

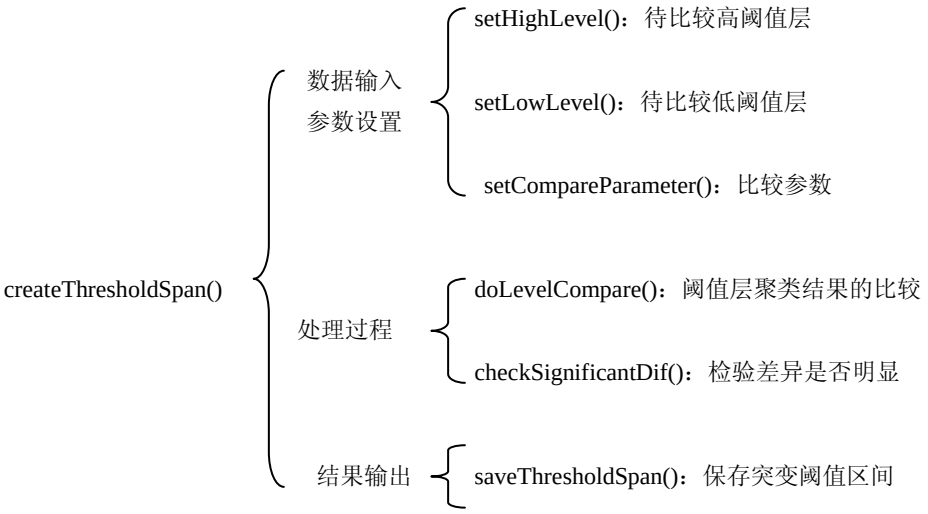


图 6-3 两个阈值层面聚类结果差异比较子模块的主要实现方法

按照要求，chechSignificant()方法必须要求六个比较项目都满足，才能确定两阈值层的聚类结果差异是明显的。

- checkSignificantDif()方法实现

```
    //检验两次聚类差异是否显著
     private void checkSignificantDif(){
    //计算差异度
            totalClusterNumDif = Math.abs(numClusterComp1-numClusterComp2);
            percentTatalClusterNumDif =
Integer.parseInt(String.valueOf(Math.round(((double)totalClusterNumDif/(double)numClusterComp2)*100)/1));
            sameClusterNum = numClusterSame;
            percentSameClusterNum =
Integer.parseInt(String.valueOf(Math.round(((double)sameClusterNum/(double)Math.max(numClusterComp1,numClusterComp2))*100)/1));
            containClusterNum = numClusterContain1;
            percentContainClusterNum =
Integer.parseInt(String.valueOf(Math.round(((double)numClusterContain1/(double)Math.max(numClusterComp1,numClusterComp2))*100)/1));
```

```
//比较差异度，必须全部满足
    if(totalClusterNumDif>=minTatalClusterNumDif
            && percentTatalClusterNumDif>=minPercentTatalClusterNumDif
            && sameClusterNum<=maxSameClusterNum
            && percentSameClusterNum<=maxPercentSameClusterNum
            && containClusterNum<=maxContainClusterNum
            && percentContainClusterNum<=maxPercentContainClusterNum
            ){
        bSignificantDif = true;//返回差异明显
    }else{
        bSignificantDif = false;
    }
}
```

（2）构建 relative 关系、twin 关系子模块的实现

构建 relative 关系、twin 关系的主要功能是为阈值层中的簇构建两种相互联系的关系，此关系将是寻找潜在结构的基础。根据阈值策略的功能要求，构建 relative 关系、twin 关系子模块的主要实现方法如图 6-4 所示。

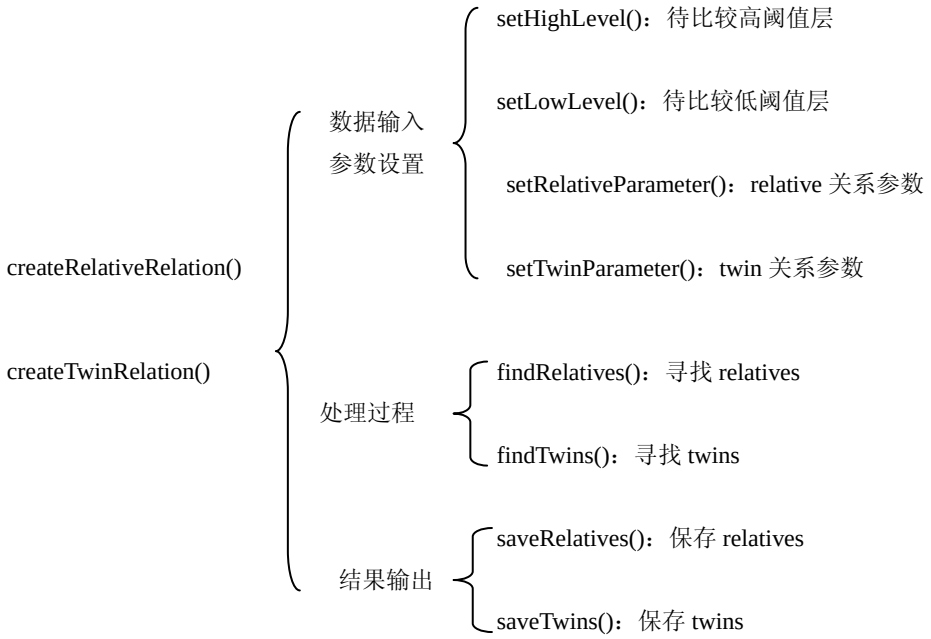


图 6-4 构建 relative 关系、twin 关系子模块的主要实现方法

该子模块的重要方法是 findTwins（），要涉及高低两个阈值层。

- findTwins()方法实现

```
//为 mapTable1 在 mapTable2 中寻找最相似的簇
 //mapTable1：低阈值层
 //mapTable2：高阈值层
 private void findTwin(){
//获取低阈值层簇及其簇内成员特征
```

```
sql = "select count(*) as cc,"+ colName+" from "+ mapTable1+" group by " +colName+" order by "+colName;
//获取与低阈值层簇相似的高阈值层簇及其簇内成员特征
sql = "select count(*) as cc,"+colName+" from "+mapTable2+" where id in "+"("+"select id from "+mapTable1+" where "+colName+" = ?"+")"+" group by "+colName+" order by cc desc"+", "+ colName;
//获取高阈值层簇及其簇内成员特征
sql = "select count(*) as cc,"+colName+" from "+mapTable2+" where "+colName+" = ? group by " +colName;
while(rs1.next()){
    similarity = MIN_SIMILARITY_TWIN;//预先设定 twin 相似性最小值 0.4;
    numSimil = 0;
    twin = 0;
    size_twin = 0;
    intersection_twin = 0;
    c_id1 = rs1.getInt(colName);
    size_c_id1 = rs1.getInt("cc");
    twins[n][0] = c_id1;
    twins[n][1] = size_c_id1;
    pstmt2.setInt(1,c_id1);
    rs2 = pstmt2.executeQuery();
    if(rs2.last()){
        numSimil = rs2.getRow();
        rs2.beforeFirst();
    }
    //找到所有可能 twin
    while(rs2.next()){//遍历高阈值层
        c_id2 = rs2.getInt(colName);
        intersection = rs2.getInt("cc");
        pstmt3.setInt(1,c_id2);
        rs3 = pstmt3.executeQuery();
        if(rs3.next()){
            size_c_id2 = rs3.getInt("cc");
        }//if(rs3.next())
        rs3.close();
        //计算相似性
        double tmp = 1*((double)intersection/(double)size_c_id1);
        //比较相似性，寻找相似性最大者
        if(tmp>similarity){
            similarity = tmp;
            twin = c_id2;
            size_twin = size_c_id2;
            intersection_twin = intersection;
        }
    }//遍历高阈值层完毕
```

```
            rs2.close();
            twins[n][2] = twin;//记录有效的 twin
            twins[n][3] = size_twin;
            if(twin==0){
                  similarity = 0.0;
            }
            n++;//继续遍历
            if(bGUI){
                  insertLog(toStringArray(c_id1,size_c_id1,twin,size_twin,intersection_twin,similarity));
            }
      }//低阈值层遍历完毕
       rs1.close();
       }
```

（3）寻找潜在簇子模块的实现

寻找潜在簇的主要功能是在低阈值层中通过 raletive、twin 关系，寻找潜在簇。根据阈值策略的功能要求，寻找潜在簇子模块的主要实现方法如图 6-5 所示。

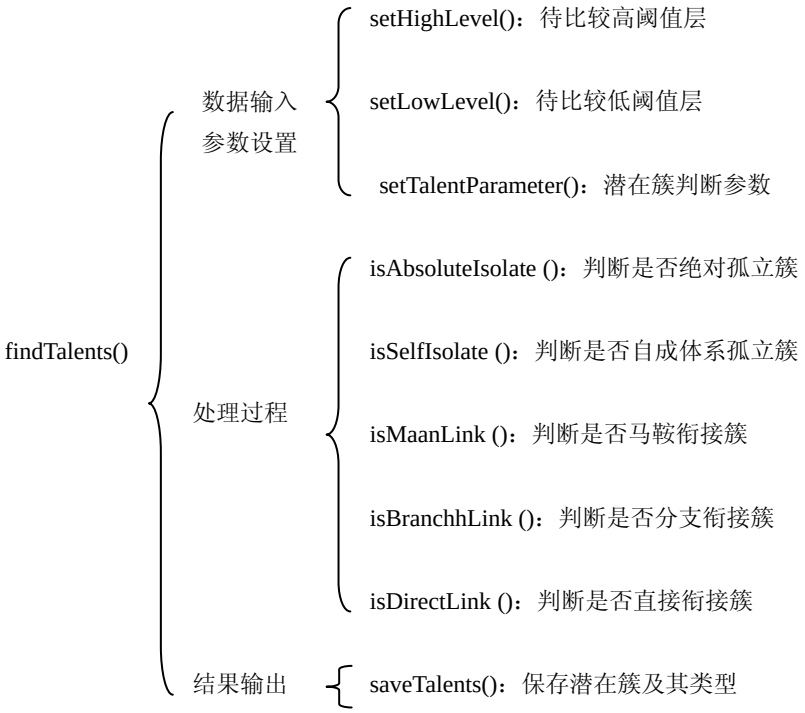


图 6-5 寻找潜在簇子模块的主要实现方法

由潜在簇的定义可知，任意一个簇只可能是 5 种潜在簇中的一种，只要当该簇满足其中一个潜在簇的条件时，就可以保存该簇的类型，进而转到下一个簇进行判断。以判断是否直接衔接簇为例，描述方法的实现。

- isDirectLink()方法实现

```
//判断是否是直接衔接簇
 //是，即新簇（相对于高阈值层），非孤立点，与高阈值层至少两个主要簇有关系
```

```
private boolean isDirectLink(int c_id){
    boolean b = false;
    try{
        int[] relative = getRelative(c_id,matrixTable1);//获取 relatives
        int[] twin_of_relative = getTwinOfRelative(relative);//获取 relatives 的 twins
        //是否是新簇、孤立簇
        if(!isNewCluster(c_id)||isAbsoluteIsolate(c_id)||isSelfIsolate(c_id)){
            b = false;
            return b;
        }
        //判断 relatives 是否新簇
        for(int i=0;i<relative.length;i++){
            if(!isNewCluster(relative[i])){
                n++;//relatives 中非新簇数
            }
        }
        if(n>1){
            //判断是否满足直接衔接簇的特性
            size = getSize(c_id);
            if(size>=MIN_SIZE_LINK){
                relation = getMinRelationWithNotNewCluster(c_id,matrixTable1);
                if(relation>=MIN_RELATION_LINK){
                    b = true;
                }else if(n>=MIN_RELATION_NUM_LINK){
                    b = true;
                }
            }
        }
    return b;
}
```

（4）构建潜在关系子模块的实现

构建潜在关系的主要功能是通过潜在关系将低阈值层中的潜在簇与高阈值层中的簇联系起来，用潜在关系近似描述两者之间潜在可能性。根据阈值策略的功能要求，构建潜在关系子模块的主要实现方法如图 6-6 所示。

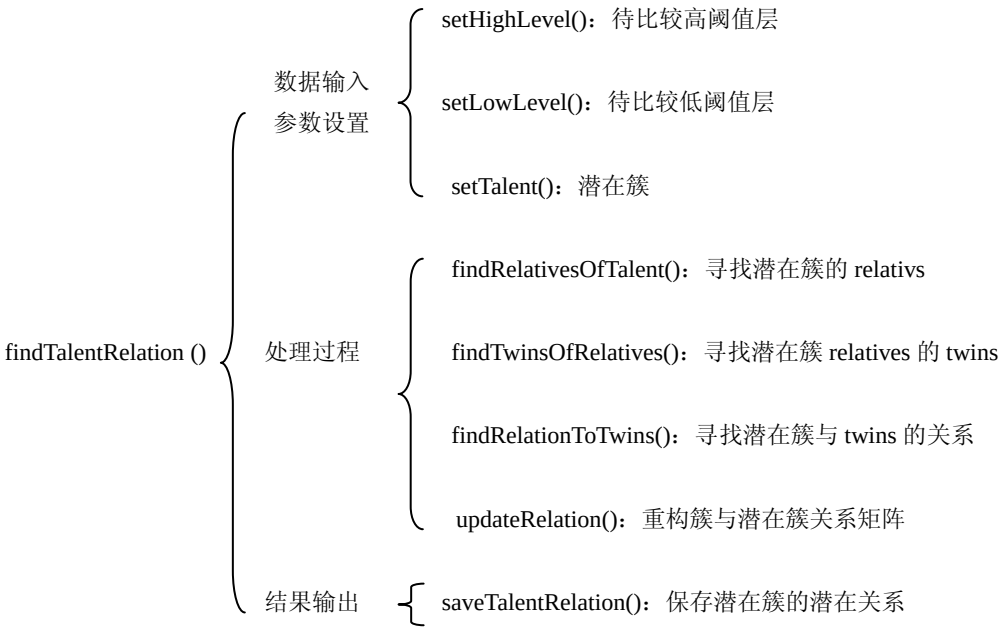
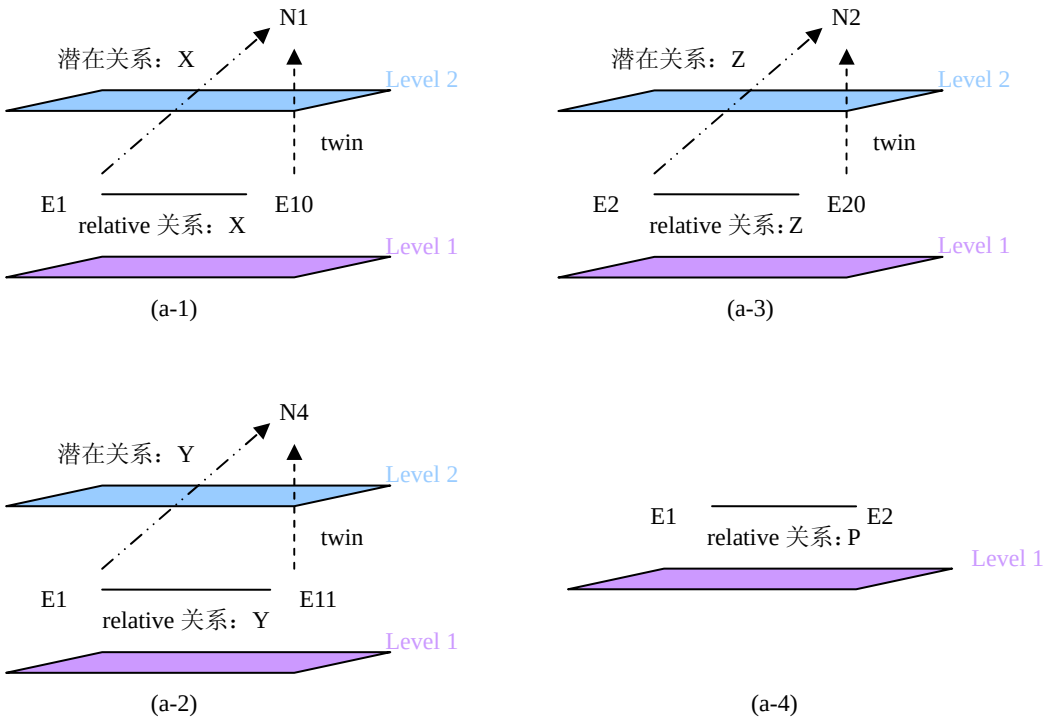


图 6-6 构建潜在关系子模块的主要实现方法

构建潜在关系的核心步骤是寻找潜在簇的 relatives 及其 relatives 的 twins，进一步，该潜在簇与 twin 的潜在关系近似表示为该潜在簇与 twin 相对应的 relative 的关系。

构建潜在关系过程中，有一个重要步骤是重构关系矩阵，如图 6-7 所示。图（a）描述了低阈值层 Leve_1 和高阈值层 Leve_2 的三种关系。左上角反映了潜在簇 E1 与 E10 由 relative 关系 X，E10 在上层有 twin 簇 N1，此时可以认为潜在簇 E1 与上层簇 N1 有潜在关系 X。左下角表明簇 E1 还与簇 E11、上层簇 N4 这种三种关系，其中潜在关系为 Y。同理，右上角的簇 E2 与上层簇 N2 有潜在关系 Z。而右下角反映两个潜在簇 E1 和 E2 之间还有 relative 关系 P。图（b）描述了高阈值层簇与簇的关系矩阵。图（c）则是在高阈值层关系矩阵中附加上潜在簇 E1、E2，而且增加了潜在簇与高阈值层簇的潜在关系（X、Y、Z）、潜在簇之间的相互关系（P）。此时，新矩阵完成了对原有矩阵的重构，包含了高阈值层簇，以及潜在簇、潜在关系。



（a）relative关系、twin关系、潜在关系

	N1	**N2**	**N3**	**N4**
N1	1	0.4	0.2	0.38
N2	0.4	1	0	0
N3	0.2	0	1	0
N4	0.38	0	0	1

（b）高阈值层关系矩阵

	N1	**N2**	**N3**	**N4**	**E1**	**E2**
N1	1	0.4	0.2	0.38	*X*	*0*
N2	0.4	1	0	0	*0*	*Z*
N3	0.2	0	1	0	*0*	*0*
N4	0.38	0	0	1	*Y*	*0*
E1	*X*	*0*	*0*	*Y*	*1*	*P*
E2	*0*	*Z*	*0*	*0*	*P*	*1*

（c）重构后，高阈值层关系矩阵附加上潜在关系

图 6-7 关系矩阵的重构

（5）绘制潜在结构展示图子模块的实现

绘制潜在结构展示图的主要功能是通过可视化方法将潜在簇和潜在关系在高阈值层结构图中显现出来，以直观、生动的方式体现出潜在性。根据阈值策略的功能要求，绘制潜在结构展示图子模块的主要实现方法如图 6-8 所示。

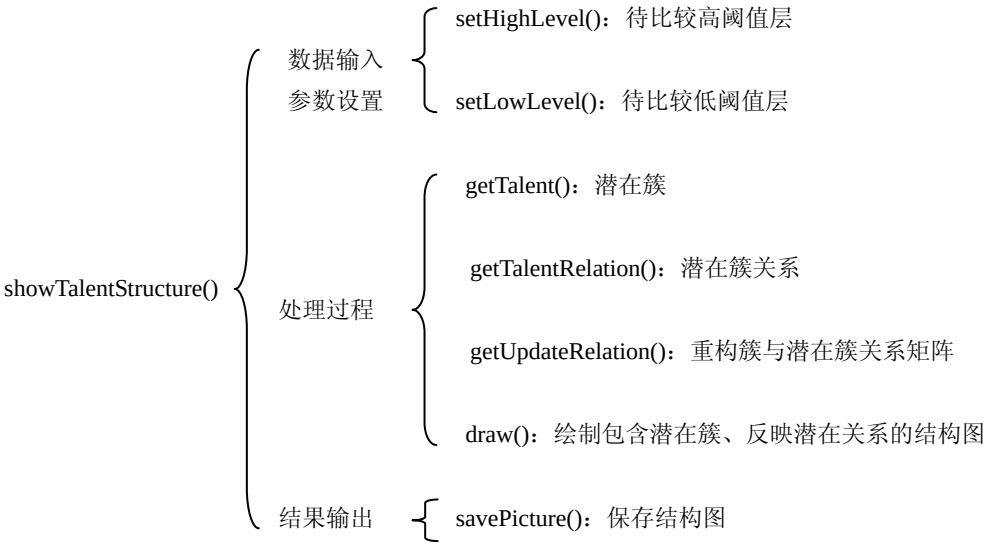


图 6-8 绘制潜在结构展示图子模块的主要实现方法

6.2 时间策略的实现

6.2.1 时间策略主模块的实现

时间策略要实现的主要功能是通过划分时间窗，从同一数据集中挑选时间属性位于时间窗的待分析对象进行多次聚类计算。然后从所有聚类结果中依次挑选出相邻的三个时间窗，为中间时间窗构建与前一时间窗的 ancestor 关系、与后一时间窗的 offspring 关系。在此基础上，以中间时间窗聚类生成的簇为考察的主体对象，根据演变簇的判断标准寻找中间时间窗相对于前一时间窗的合并簇、融合簇、新增簇，以及相对于后一时间窗的分化簇、扩散簇、消失簇。根据时间策略的功能要求，时间策略主模块的主要实现方法如图 6-9 所示。

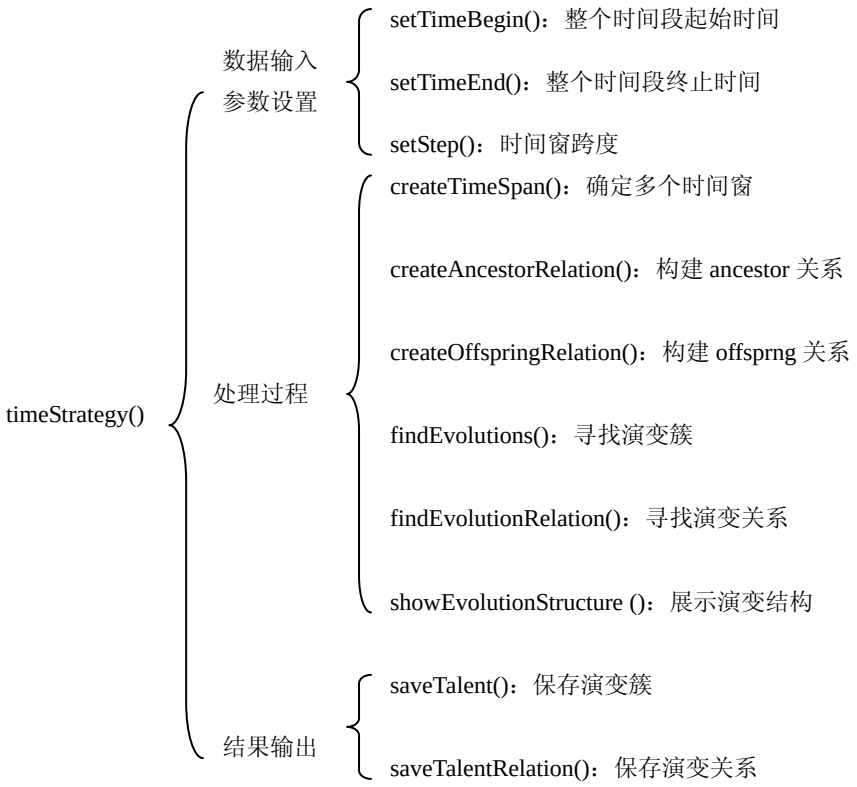


图 6-9 时间策略主模块的主要实现方法

图 6-10 展现了应用时间策略，发现演变结构的详细实现流程。值得一提的是，时间策略虽然也是在多个据类结果中寻找某种关系，但是没有对前、后两个阈值层突变区间的计算。这是与阈值策略最大的不同。

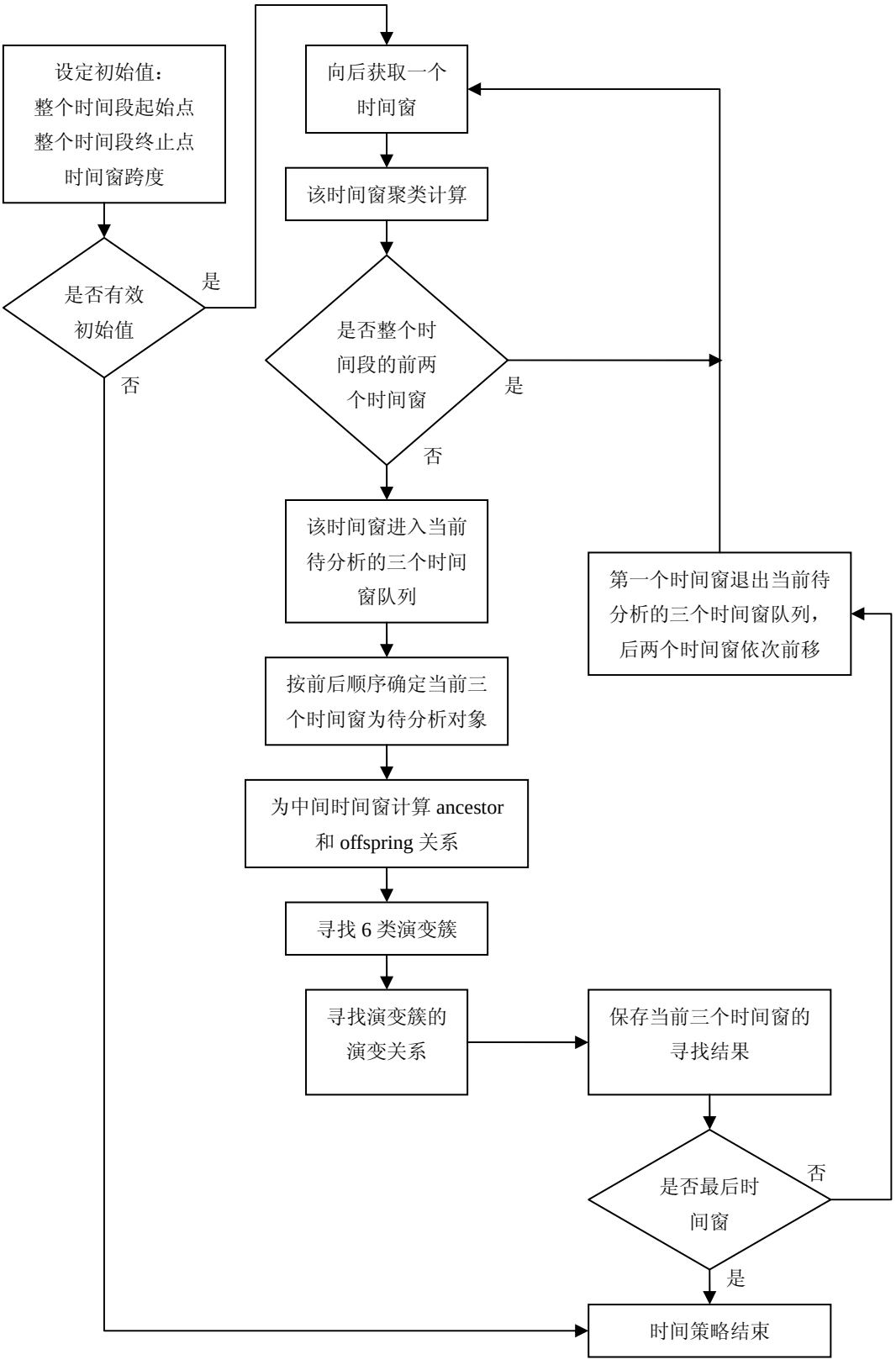


图 6-10 应用时间策略，发现演变结构的实现流程

6.2.2 时间策略子模块的实现

（1）构建 ancestor 关系、offspring 关系子模块的实现

构建 ancestor 关系、offspring 关系的主要功能是为三个相邻时间窗中的簇构建两种相互联系的关系，此关系将是寻找演变结构的基础。根据时间策略的功能要求，构建 ancestor 关系、offspring 关系子模块的主要实现方法如图 6-11 所示。

图 6-11 构建 ancestor 关系、offspring 关系子模块的主要实现方法

构建三个时间窗的 ancestor 关系、offspring 关系在实现中合并在一个方法 findEvolution（）中，因为它们的计算方法是一样的。

- findEvolution()方法实现

```
//为 mapTable1、mapTable2、mapTable3 三个相邻时间窗寻找 evolution 关系
    private void findEvolution(){
    int[] evolve = null;//保存演变关系的数组
    try{
        //先清空祖先关系和后代关系临时队列
        evolutions1To2.clear();
        evolutions2To3.clear();
        //为 mapTable1 在 mapTable2 中寻找演变
        sql = "select count(*) as cc,"+colName+ " from "+mapTable1+ " group by "+colName+" order by "+colName;
        sql = "select count(*) as cc,"+colName+" from "+mapTable2+" where id in "+"("+ "select id from "+mapTable1+" where "+colName+" = ?"+")"+" group by "+colName+" order by cc desc"+", "+colName;
        sql = "select count(*) as cc,"+colName+ " from "+mapTable2+" where "+ colName+" = ? group by " +colName;
        ……
        while(rs1.next()){//遍历第一个时间窗
            c_id1 = rs1.getInt(colName);
            size_c_id1 = rs1.getInt("cc");
            pstmt2.setInt(1,c_id1);
            rs2 = pstmt2.executeQuery();
            while(rs2.next()){
```

```
                    c_id2 = rs2.getInt(colName); //在第二个时间窗中寻找后代
                    intersection = rs2.getInt("cc");
                    pstmt3.setInt(1,c_id2);
                    rs3 = pstmt3.executeQuery();
                    if(rs3.next()){
                        size_c_id2 = rs3.getInt("cc");
                        //记录演变关系，演变簇推入队列
                        evolve = new int[5];
                        evolve[0] = c_id1;
                        evolve[1] = size_c_id1;
                        evolve[2] = c_id2;
                        evolve[3] = size_c_id2;
                        evolve[4] = intersection;
                        evolutions1To2.add(evolve);
                    }
                    rs3.close();
                }//第二个时间窗遍历完毕
                rs2.close();
            }//第一个时间窗遍历完毕
            rs1.close();

            //为 mapTable2 在 mapTable3 中寻找 evolution
            sql = "select count(*) as cc,"+colName+" from "+mapTable2+" group by "+colName+"
order by "+colName;
            sql = "select count(*) as cc,"+colName+" from "+mapTable3+" where id in "+"("+"select id
from "+mapTable2+" where "+colName+" = ?"+ ")" +" group by "+ colName+" order by cc desc"+ ",
"+colName;
            sql = "select count(*) as cc,"+colName+" from "+mapTable3+" where "+colName+" = ?
group by "+colName;
            ……
            while(rs1.next()){//遍历第二个时间窗
                c_id1 = rs1.getInt(colName);
                size_c_id1 = rs1.getInt("cc");
                pstmt2.setInt(1,c_id1);
                rs2 = pstmt2.executeQuery();
                while(rs2.next()){
                    c_id2 = rs2.getInt(colName); //在第三个时间窗中寻找后代
                    intersection = rs2.getInt("cc");
                    pstmt3.setInt(1,c_id2);
                    rs3 = pstmt3.executeQuery();
                    if(rs3.next()){
                        size_c_id2 = rs3.getInt("cc");
                        //记录演变关系，演变簇推入队列
                        evolve = new int[5];
```

```
                                evolve[0] = c_id1;
                                evolve[1] = size_c_id1;
                                evolve[2] = c_id2;
                                evolve[3] = size_c_id2;
                                evolve[4] = intersection;
                                evolutions2To3.add(evolve);
                            }
                            rs3.close();
                        }//第三个时间窗遍历完毕
                        rs2.close();
                    }//第二个时间窗遍历完毕
                    rs1.close();
                }
            }
```

（2）寻找演变簇子模块的实现

寻找演变簇的主要功能是在中间时间窗中通过 ancestor、offspring 关系，寻找演变簇。根据时间策略的功能要求，寻找演变簇子模块的主要实现方法如图 6-12 所示。

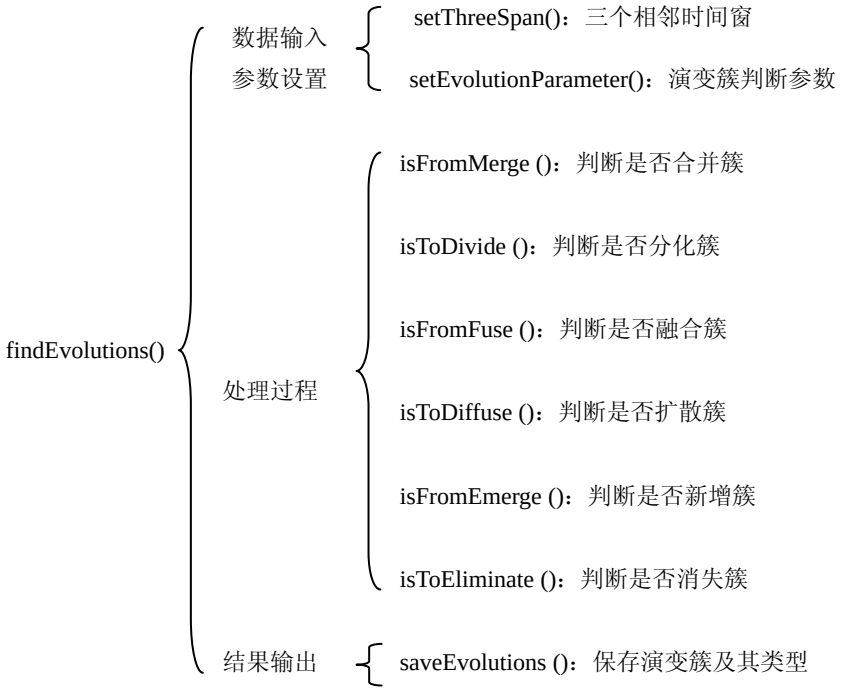


图 6-12 寻找潜在簇子模块的主要实现方法

由演变簇的定义，任意一个簇可能是 6 种演变簇中的一种或几种，因此，必须为每个簇进行 6 次判断，当满足一个条件时就要保存该簇的类型，待六次判断全部结束后再转到下一个簇进行判断。以判断是否融合簇为例，描述方法的实现。

- isFromFuse()方法实现

```
//      判断是否是来自融合
```

```
    //是，即成员数>=MIN_SIZE_FROM_FUSE
    //本质性<MIN_QUALITY_RATE
    //新颖性>=MIN_NOVELTY_RATE
    //汇聚数>=1
    private boolean isFromFuse(int c_id){
        boolean b = false;
        try{
            if(getSizeOfSelf(c_id)>=MIN_SIZE_FROM_FUSE//成员数
                    && (getQualityRate(c_id,ANCESTOR)<MIN_QUALITY_RATE//本质性
                    && 1-getQualityRate(c_id,ANCESTOR)>=MIN_NOVELTY_RATE)//新颖性
                    && getNumOfInflux(c_id)>=1//MIN_INFLUX_NUM//汇聚数
                    ){
                b = true;
            }
        }
        return b;
    }
```

（3）构建演化关系子模块的实现

构建演化关系的主要功能是通过演化关系将中间时间窗中的演化簇与前后时间窗中的簇联系起来，用演化链描述三个相邻时间窗的演化过程，用遗传率、继承率描述演化强度。根据时间策略的功能要求，构建演化关系子模块的主要实现方法如图 6-13 所示。

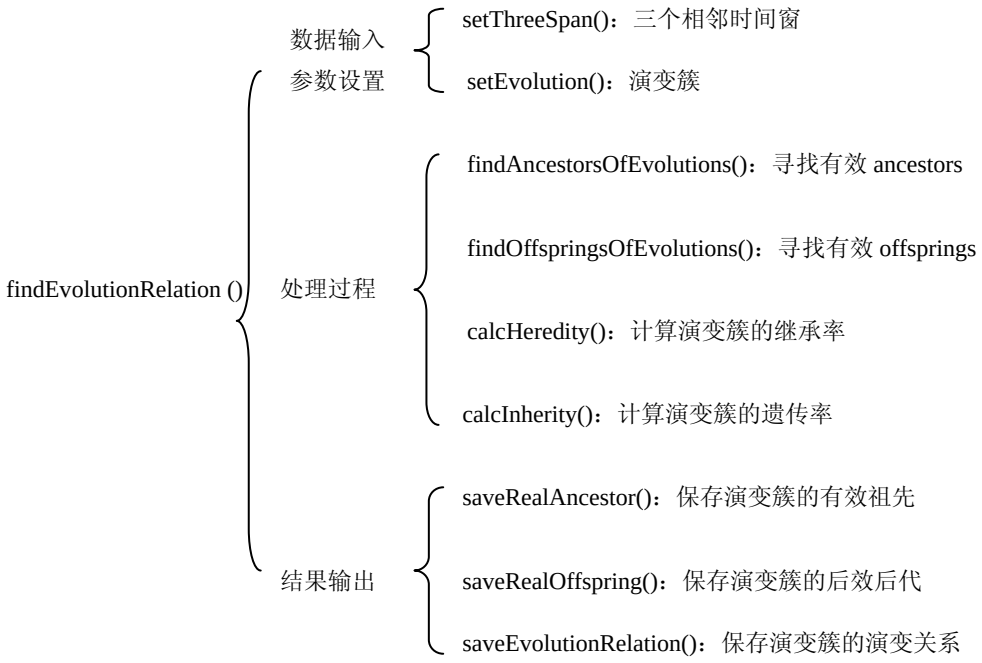


图 6-13 构建演化关系子模块的主要实现方法

演化关系的核心是演化簇有哪些祖先，相对于祖先的继承率分别是多少；演化簇有哪些后代，相对于后代的遗传率分别是多少。

演变关系保存在数据库的表中。因为演变簇可能是 6 种类型中的一种或者几种，因此在保存演变关系的时候，应该考虑保存演变簇的有效祖先和有效后代，即使这个演变簇可能没有有效祖先或者没有有效后代。基于这样的考虑，保存演变关系的数据库表结构设计如表 6-1 所示。

表 6-1 演变关系的数据库表结构

字段名	数据类型	含义
ID	int	演变簇 ID 号
TYPE	char	演变簇类型（6 中类型中的一种）
EVOLUTION_ID	int	祖先或后代 ID 号
RELATION	float	相对于祖先的继承率或相对于后代的遗传率

例如，某个演变簇 ID 号是 10，属于合并簇，其中一个有效祖先 ID 号是 55，继承率是 0.87，那么此时演变关系的一条记录是[10；“合并簇”；55；0.87]。同样还是这个演变簇，属于扩散簇，其中一个有效后代 ID 号是 88，遗传率是 0.68，那么此时该簇的另一条演变关系记录是[10；“扩散簇”；88；0.68]。这样的表结构，可以通过“TYPE”字段，辩明“EVOLUTION_ID”指的是祖先，还是后代，进一步辩明“RELATION”指的是继承率，还是遗传率。

（4）绘制演变结构展示图子模块的实现

绘制演变结构展示图的主要功能是通过可视化方法将演变簇和演变关系三个时间窗中显现出来，以直观、生动的方式体现出演变性。根据时间策略的功能要求，绘制演变结构展示图子模块的主要实现方法如图 6-14 所示。

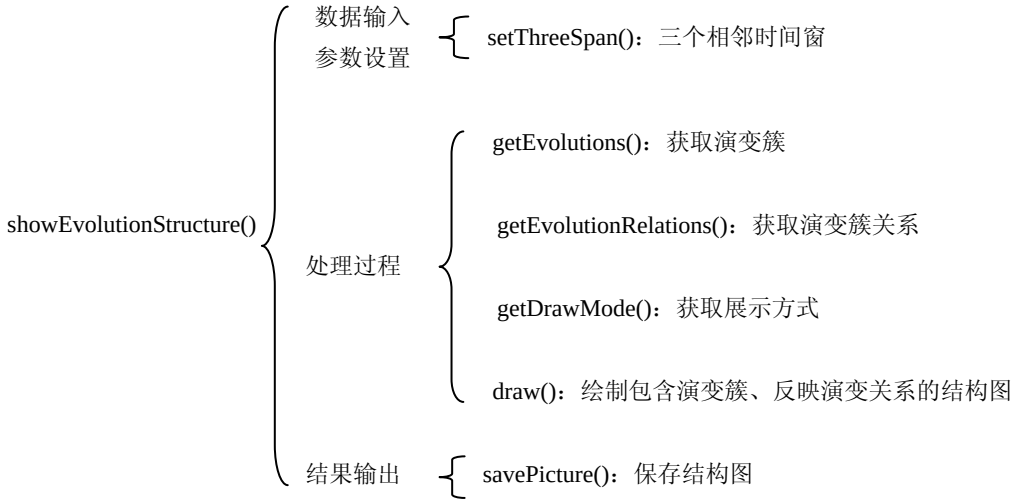


图 6-14 绘制演变结构展示图子模块的主要实现方法

6.3 系统运行结果

两种策略集成在一个系统中，在共引聚类的基础上立即自动启动寻找潜在结构或者演变结构的计算，聚类结果以及潜在结构或演变结构的结果保存在数据库中，然后可以开始可视化显示。

（1）数据

本文采用 ISI 的网络资源库 ESI（2007 年更新），选择其中 COMPUTER 学科的高被引文献作为分析数据。总数据量为 2226 篇文献，发表年份分布在 1996－2006 年。

（2）方法

共引分析，单链接聚类。

（3）软件环境

操作系统：Microsoft Windows 2003 Server Enterprise Edition Service Pack 2

数据库服务器：Microsoft SQL Server 2000

Java 运行环境：Jrockit-R27.3.0-jre1.5.0_11

（4）硬件环境

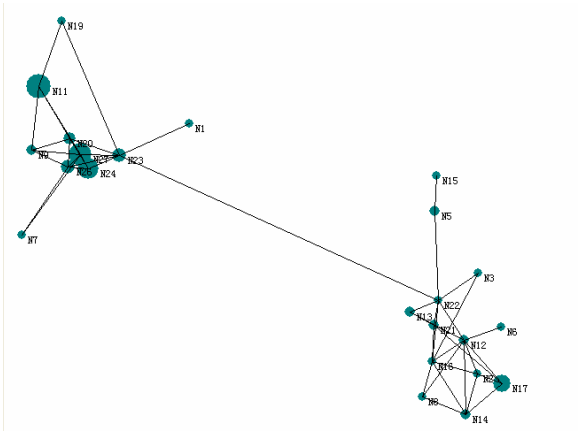
CPU：Intel(R) Xeon(R) E5335 2.00GHz

RAM：8.00G

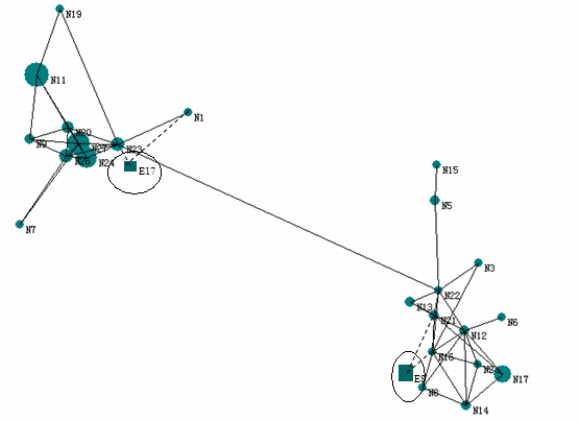
6.3.1 阈值策略结果

针对持续的突变阈值区间，绘制的展示图如图 6-15 所示。此时应用阈值策略，最高阈值设定为 0.2，降低阈值的步长设定为 0.02。为了与主体结构以示区别，潜在簇用方形表示，潜在关系用虚线表示。观察三幅图可以看到，（a）是最高阈值层的聚类结果展示图，阈值还没有开始下将，因此没有潜在簇。（b）是从突变阈值区间中首次差异显著的阈值层找到了 2 个潜在簇，用圆圈表示，此时的阈值是 0.16，已经降低了 2 步。由于采用的是持续的突变阈值区间，因此后面阈值层相对于最高阈值层都是差异显著的，故而从阈值 0.16 开始，下面的每一个低阈值层都能找到潜在簇。（c）展示了下一个低阈值层（阈值为 0.14）相对于最高阈值层的潜在簇，用圆圈表示，其中有两个潜在簇与上一个低阈值层是相同的，另外两个是新增的。

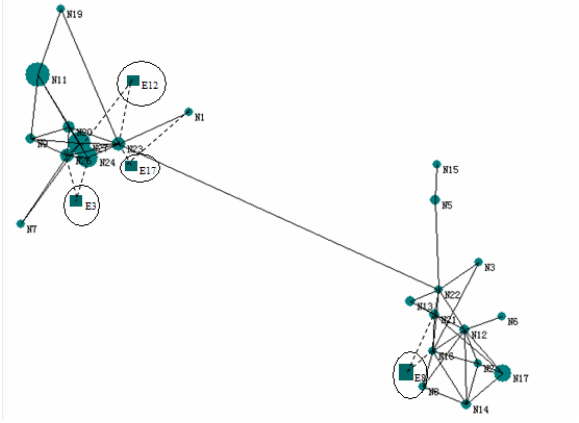
采用持续的突变阈值区间绘制潜在结构，优点是能保持主体结构稳定，随着阈值逐步降低，潜在结构逐渐附加在主体结构之上，使每一次展示能清晰看到潜在结构的逐渐复杂、丰富的过程。但是，这种方式并不能保证每降低一次阈值就能发现潜在结构，因此展示效率会比较低。



（a） 阈值0.2



（b） 阈值0.16



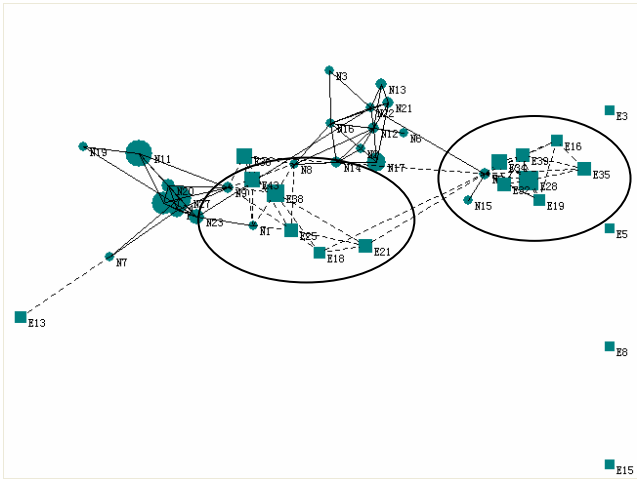
（c） 阈值0.14

图 6-15 采用持续的突变阈值区间，寻找潜在结构的展示图

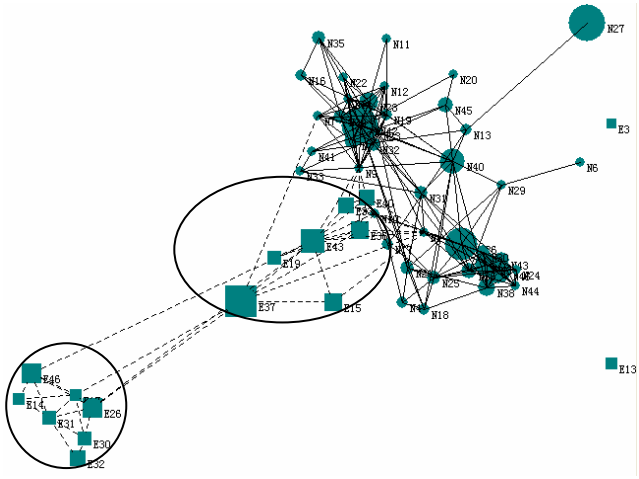
针对跳跃的突变阈值区间，绘制的展示图如图 6-16 所示。此时应用阈值策略，最高阈值设定为 0.4，降低阈值的步长设定为 0.02。为了与主体结构以示区别，潜在簇用方形表示，潜在关系用虚线表示。为相邻两个突变阈值区间绘制两幅图，观察这两幅图可以看到，（a）是高阈值层 0.2，低阈值层 0.1 时寻找潜在结构的展示图。（b）是高阈值层 0.1，低阈值层 0 时寻找潜在结构的展示图。由

于采用的是跳跃的突变阈值区间，因此两幅图的主体结构是不同的，前者以阈值层 0.2 的聚类结果为主体结构，后者以阈值层 0.1 的聚类结果为主体结构。而且前者反映的是低阈值层 0.1 相对于高阈值层 0.2 的潜在结构，后者反映的是低阈值层 0 相对于高阈值层 0.1 的潜在结构。所以，两幅图的差异很大。

采用持续的突变阈值区间绘制潜在结构，优点是能较快定位到一个最有可能蕴涵潜在结构的低阈值层，而且高阈值层与低阈值层在每次差异显著后，低阈值层将变为高阈值层，将为它在更低阈值层中寻找潜在结构。缺点是不能保持主体结构稳定，随着阈值逐步降低，潜在结构的逐渐变化无法体现。



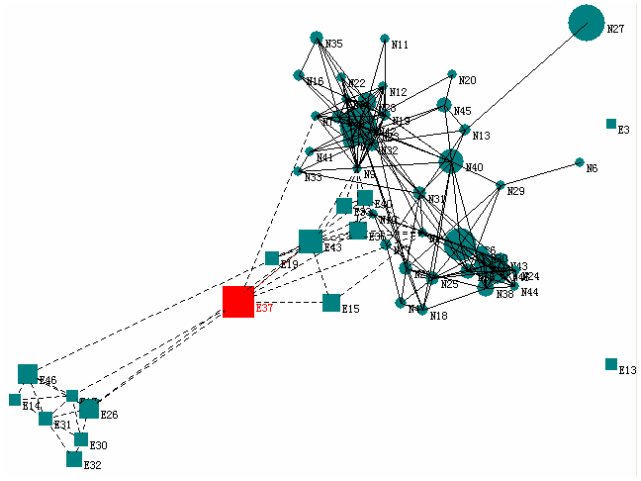
（a） 高阈值层0.2，低阈值层0.1



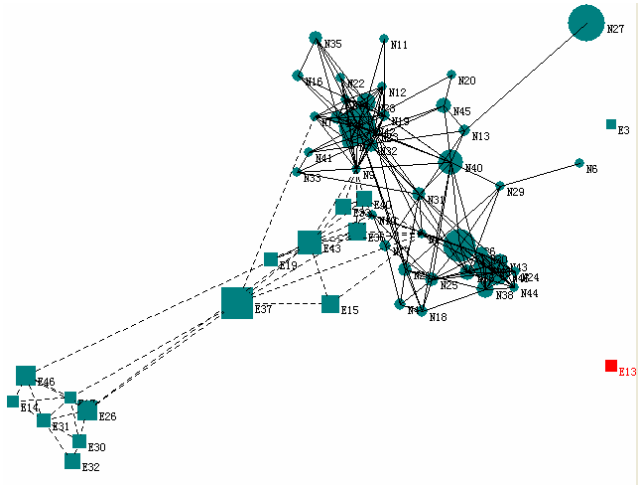
（b） 高阈值层0.1，低阈值层0

图 6-16 采用跳跃的突变阈值区间，寻找潜在结构的展示图

寻找潜在簇的结果示例如图 6-17 所示。（a）中红色方形 E37 是个直接衔接簇，与高阈值层的簇 N5、N1、N7 有潜在关系。（b）中红色方形 E13 是个自成体系孤立簇，与高阈值层的所有簇都没有潜在关系。此处寻找潜在簇的详细的结果可以参见附录。



（a） 直接衔接簇E37



（b）自成体系孤立簇E13

图 6-17 寻找潜在簇的结果示例

6.3.2 时间策略结果

显示三个时间窗，绘制的展示图如图 6-18 所示（截取部分）。此时应用时间策略，采用发表年份位于 1996 年至 2005 年的数据，时间窗跨度设置为 3 年。图中显示了三个时间窗的演变关系，能清晰看到中间时间窗由哪些祖先经过什么方式演变而来，又经过什么方式演变成哪些后代。

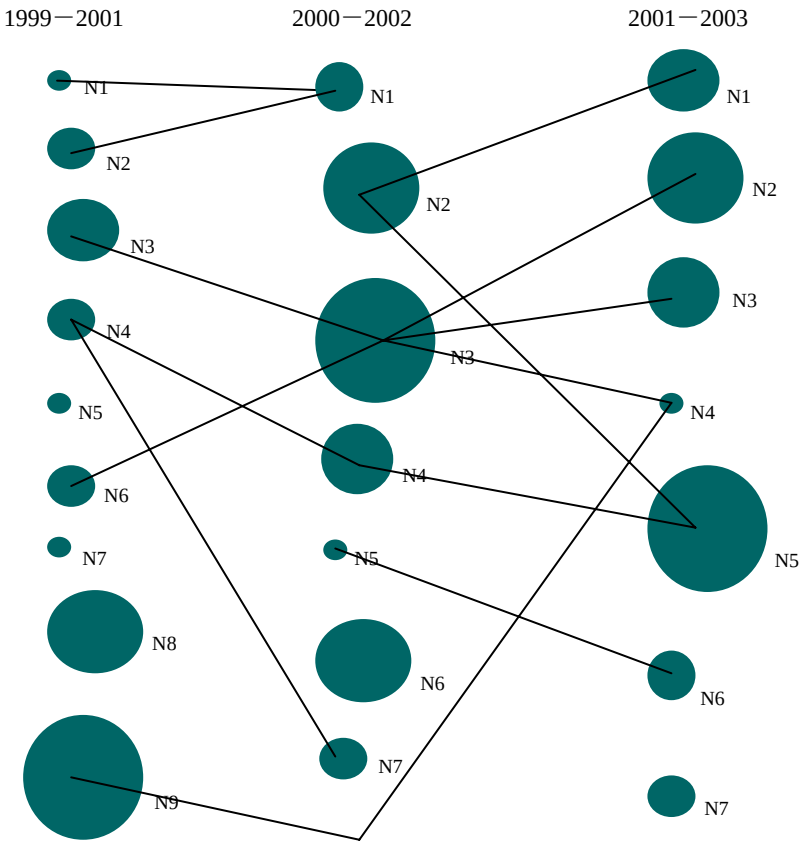


图 6-18 显示三个时间窗

只显示以一时间窗，绘制的展示图如图 6-19 所示。此时应用时间策略，采用发表年份位于 1996 年至 2005 年的数据，时间窗跨度设置为 5 年。为了与主体结构以示区别，演变簇用方形表示。图中显示的是 2000－2004 时间窗聚类结果的簇结构，方形簇代表这个时间窗中相对于前、后时间窗的演变簇，而圆形簇代表这个时间窗中相对于前、后时间窗的非演变簇。这种展示方式虽然只能另外用文字标示演变簇的演变关系，而不能像前面一个方式那样直接看到，但是可以显示演变簇在本时间窗的位置和结构。

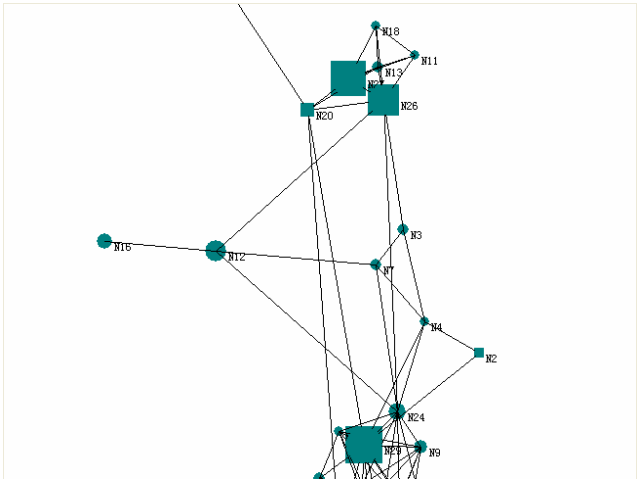
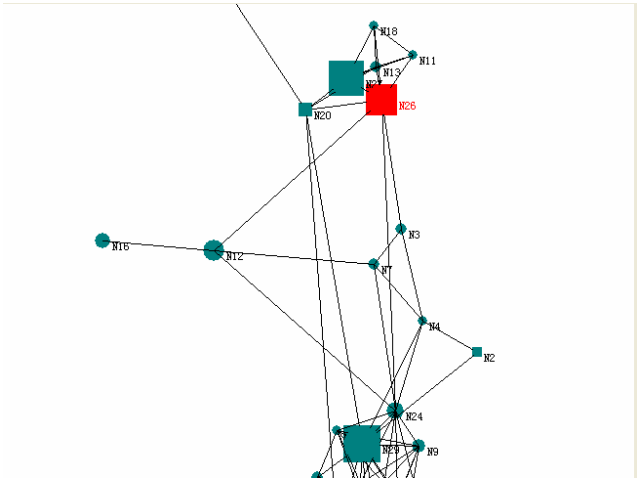
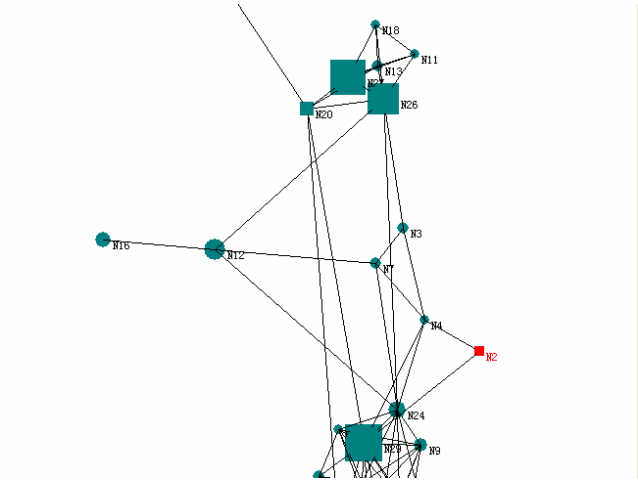


图 6-19 只显示一个时间窗

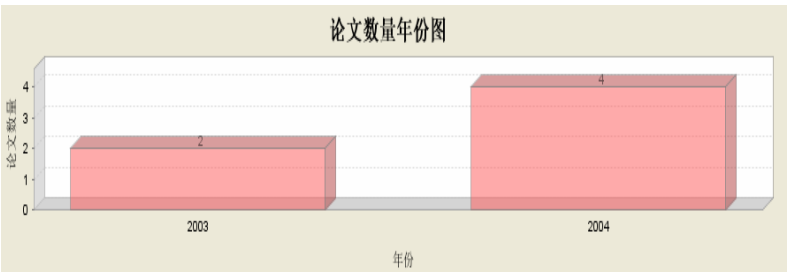
寻找演变簇的结果示例如图 6-20 所示。（a）中红色方形 E26 是个合并簇，是祖先簇 N3、N8、N9 合并而来，同时，它又是分化簇，将分解成簇 N7、N14、N2 等后代。（b）中红色方形 N2 是个新增簇，它没有自己的祖先。而且相对于祖先层，它新颖性很强，这一点可以从（c）中簇 N2 的成员年份分布图看到，它的成员基本上位于本时间窗的后部，因而与祖先层的簇没有公共成员。此处寻找演变簇的详细的结果可以参考附录。



（a）合并而来、走向分化的簇N26



（b）新增簇N2



（c）新增簇N2的成员年份分布图

图 6-20 寻找演变簇的结果示例

6.4 小结

本章基于前面设计出的阈值策略和时间策略，分别对两个策略中的方法进行了实现。各种主要方法的详细实现流程体现了这些集成在系统中的方法的实现过程。从得到的结果看，阈值策略能实现寻找潜在结构的目标，时间策略能实现寻找演变结构的目标。

7 总结与展望

本文利用阈值和时间两个特性，从潜在和演变两个角度，对知识结构演化的分析方法进行拓展，以期实现深度分析，补充现有不足，更好满足需求。

7.1 工作总结

在对探寻知识结构演化相关方法的分析基础上，本文的主要工作集中在阈值策略、时间策略的设计和实现。具体表现在以下几个方面：

（1）分析了与本论文相关的其他研究，在了解现状的前提下，分析其主要特点，指出其存在的不足之处，进而明确本文要解决的主要问题。

（2）确定了整体研究框架。同时，根据研究框架，为试验系统设计了整体体系结构，整体运行流程，核心功能模块。

（3）分析和设计了阈值策略。首先通过对不同阈值层面聚类结果的比较，用定量的方法证明了以阈值逐步降低为基础的阈值策略用于发现潜在结构是可行的。接着先定义了同一阈值层中簇与簇的 relative 关系和不同阈值层之间的 twin 关系。然后定义了潜在簇，进而划分并定义 5 种潜在簇：绝对孤立簇、自成体系孤立簇、马鞍衔接簇、分支衔接簇、直接衔接簇。进一步，设计出阈值策略实施的步骤，包括突变阈值区间的确定、relative 关系和 twin 关系的构建、5 种潜在簇的寻找、潜在关系的确定。因为持续的突变阈值区间和跳跃的突变阈值区间，潜在结构展示采用不同方式，但都将以高阈值层作为主体结构，低阈值层的潜在簇作为次要结构，并通过潜在关系将次要结构附属在主体结构上。

（4）分析和设计了时间策略。首先构建时间窗之间簇与簇的关系：ancestor 关系和 offspring 关系。通过对两种关系的定量分析，可以确定时间窗之间存在强烈的“遗传继承”性。紧接定义了簇关于时间演变的特性，如遗传率、继承率、有效祖先数、有效后代数、本质性、新颖性、陈旧性、汇聚率、分解率等。利用这些演变特性，进而将演变簇划分为 6 个种类并给出定义：合并簇、分化簇、融合簇、扩散簇、新增簇、消失。然后设计时间策略实施的步骤，包括 ancestor 关系和 offspring 关系的构建、6 种演变簇的寻找、演变关系的确定。最后演变结构可视化展示采用两种不同方式，一种只显示一个时间窗，另一种显示三个时间窗。

（5）分别对两个策略中的主要方法设计了实现流程。运行系统，两个策略付诸实施，得到相应结果，证明了阈值策略能实现寻找潜在结构的目标，时间策略能实现寻找演变结构的目标。

7.2 创新之处

本论文的创新之处主要体现在以下几个方面：

（1）提出并初步实现分别从纵向和横向分析知识结构演化的两种方法。

本文提出了基于不同阈值层聚类结构中聚类簇的相互关系、对不同阈值层聚类结构间差异性进行自动检测分析、自动发现宏观结构下的潜在结构的方法，并提出了确定突变阈值区间、分析潜在结构、可视化表示潜在结构的方法。还提出了不同时间窗聚类结构主题簇关系自动检测分析、自动发现不同时间下聚类结构的演变过程、并进行可视化表现的方法。对这两种方法进行了初步验证。

（2）提出簇与簇的关系模型。

通过构建一个簇与本阈值层簇的 relative 关系，以及 relatives 与高阈值层簇的 twin 关系，建立低阈值层与高阈值层的关联，为寻找潜在结构打下基础。

通过构建 ancestor 关系研究簇的产生与由来，通过构建 offspring 研究簇的走向和发展，以建立三个相邻时间窗的关联，为寻找演变结构打下基础。

（3）提出簇的描述特性及其种类划分定义。

提炼出潜在簇和演化簇的特性，对各种潜在簇和演化簇给出了详细的定义和计算公式，为系统实现潜在簇和演变簇的寻找打下了基础，为辩明、描述、理解潜在关系和演变关系提供了帮助。

（4）提出并初步实现知识结构演化深度分析的研究框架和系统模型。

本文以阈值策略和时间策略为核心，对知识结构演化深度分析的方法进行了深入研究，并为两个策略设计了详细的实现流程。系统以阈值策略、时间策略为主要功能模块，在纵向和横向两个维度初步实现了对知识结构演化的深度分析。

7.3 研究不足及后续工作

7.3.1 本研究局限与不足

本文提出以阈值层略和时间策略为核心的知识结构演变的深度分析方法，构建并实现一个试验系统，初步证实方法的有效性，但是由于研究能力、实验条件、时间等方面的限制，使得研究中仍然有许多不足之处。主要体现在：

（1）在有效性方面，虽然本文提出方法在试验系统中得以实现，但是方法的有效性、普适性，需要大量的数据和来自不同学科领域的数据进行验证，结果的可靠性、健壮性与稳定性需要在长期的实践应用过程中加以检验。

（2）在时间窗划分方面，本文提出的时间策略能处理的时间窗要求必须有公共部分。这是由于时间策略的两个关键关系——ancestor 关系、offspring 关系，

构建在簇内有相同成员的基础之上。但是，绝对时间窗，即没有公共部分的时间窗，同样反映出演变关系。因此，本文的局限使得系统对于没有公共部分的时间窗无法进行处理。

（3）在可视化方面，对潜在结构、演变结构的展示效果不太理想。本文利用实验室已有的可视化工具并做了相应改进，用于本论文的结果展示工具。由于论文得到结果的特殊性，可视化工具在显示结果关系时表达不够充分，不够灵活。

7.3.2 后续研究工作

本论文有许多不足之处，因此可以在后续的研究工作中继续改进和完善。主要有以下几个方面：

（1）对系统进行测试，对得到的结果进行检验。一方面选取大数据量、多学科领域的数据测试系统的健壮性和普适性，尤其对于不同的学科领域系统内设的一些参数要进行修正，以适用于对不同学科的分析应用。另一方面，对于得到的结果，要请相关专家进行判读，用事实进行检验。

（2）改善可视化效果。进一步改进可视化效果，更充分、灵活地体现潜在性、和演变性，提高系统的直观性和可理解性，使结果更易于用户理解和接受。

（3）深入内容层面进行主题分析，构建深层次的簇间关系。不论阈值策略的 twin 关系，还是时间策略的 ancestor、offspring 关系，都是建立在簇内有相同成员的基础上，显然在这个基础上构建的关系略显片面、薄弱。后续工作最好能深入簇的内容，利用簇内成员的主题相似度，不仅能构建更全面、更坚实的簇关系，而且能处理绝对时间窗，把多个时间窗的演变结构接续起来形成一串长时间的演变链，增强系统的处理能力。

（4）扩大分析对象的范畴。本文的研究对象都是单个簇，讨论的都是单个簇的特性和关系。后续工作可以将分析对象扩大到簇群这个层面，对簇群特性、簇群之间的关系进行分析。比如，如果簇群代表一个学科领域，那么对簇群交叉融合的分析研究，就能反映出不同领域交叉融合的特征。

参考文献

[1] 江文年, 杨建梅. 基于多视角知识演化的企业知识管理体系研究[J]. 科技管理研究, 2004, 24(4): 65-67.

[2] 何云峰. 关于建构知识科学的问题[J]. 上海师范大学学报: 哲学社会科学版, 2003, 32(1): 8-12.

[3] 梁战平. 情报学若干问题辨析[J]. 情报理论与实践, 2003, 26(3): 193-198.

[4] 刘植惠. 知识基因理论的由来,基本内容及发展[J]. 情报理论与实践, 1998, 21(2): 71-76.

[5] Kuhn. The structure of scientific revolutions[M].Chicago: University of Chicago Press,1996.

[6] Small. H. Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents[J]. Journal of the American Society for information Science, 1973, 24(4):28-31.

[7] Small. H. Structural Dynamics of Scientific Literature[J]. International Classification, 1976, 3(2): 67-74.

[8] Small. H. A co-citation model of scientific specialty:a longitudinal study of collagen research[J]. Social studies of science, 1977(7): 139-166.

[9] Small, H. G. Co-citation context analysis and the structure of paradigms[J]. Journal of Documentation, 1998, 36(3): 183-196.

[10] Small, H. G. A general framework for creating large-scale maps of science in two or three dimensions: The sciviz system[J]. Scientometrics, 1998, 41(1): 125-133.

[11] Small, H. A sci-map case study: building a map of AIDS research[J]. Scientometrics, 1994, 30 (1): 229-41.

[12] Small, H. Macro-level changes in the structure of co-citation clusters: 1983-1989[J]. Scientometrics, 1993, 26 (1): 5-20.

[13] Small. H. Clustering the science citation index using co-citations[J]. Scientometrics, 1985, 7(3): 391-409.

[14] Small, H. The synthesis of speciality narratives form co-citation clusters[J]. Journal of the American Society for Information Science, 1986, 37(3): 97-110.

[15] Small, H., Greenlee, E. Collagen research in the 1970s[J]. Scientometrics, 1986, 10(1-2): 95-117.

[16] http://portal.isiknowledge.com/portal.cgi?DestApp=ESI&Func=Frame

[17] Small，H．Macro-level changes in the structure of co-citation clusters：1983-1989[J]．Scientometrics，1993,26 (1): 5-20.

[18] Small, H. Visualizing Science by Citation Mapping[J]. Journal of the American Society for Information Science, 1999, 50(9): 799-813.

[19] Small, H．A passage through science: crossing disciplinary boundaries[J]．Library Trends，1999,48(1) : 72-108.

[20] E. Garfield. Scientography: Mapping the tracks of science[J]. Current Contents: Social Behavioral Sci.,1994, 7: 5-10.

[21] Morris S A, Yen G, Wu Z, et al. Time Line Visualization of Research Fronts[J]. Journal of the American society for information science and technology. 2003, 54(5): 413-422.

[22] van den Besselaar P, Heimeriks G. Mapping research topics using word-reference co-occurrences: A method and an exploratory case study[J]. Scientometrics. 2006, 68(3): 377-393.

[23] 王曰芬，宋爽，苗露．共现分析在知识服务中的应用研究[J]. 现代图书情报技术, 2006(4)：29-34.

[24] Egghe, L., Rousseau, R. Co-citation, bibliographic coupling and a characterization of lattice citation networks[J]. Scientometrics, 2002, 55(3): 349-361.

[25] Kessler, M. M. Bibliographic coupling between scientific papers[J]. American Documentation, 1996, 14: 10-25.

[26] Glanzel, W., Czerwon, H. J., A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level[J]. Scientometrics, 1996, 37 (2): 195-221.

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

[28] Mubeen, M. A. Bibliographic coupling: an empirical study of economics[J]. Annals of Library. Science and Documentation, 1995, 42 (2): 41-53.

[29] MuHsuan, Huang. Constructing a patent citation map using bibliographic coupling: A study of Taiwan's high-tech companies[J]. Scientometrics; 2003, 58(3): 195-221.

[30] Glanzel, W., Czerwon, H. J. A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level[J]. Scientometrics, 1996, 37(2): 195-221.

[31] Kostoff R N, Eberhart H J, Toothman D R. Database Tomography for Technical Intelligence: a Roadmap of the Near-earth Space science and Technoloy Literature[J]. Information Pmcessing & Management. 1998, 34(1): 69-85.

[32] D. R. Swanson. On the fragmentation of knowledge, the connection explosion, and assembling other people’s ideas[J]. Bull. Amer. Soc. Inform. Sci. Technol., 2001, 27: 12-14.

[33] D. R. Swanson, N. R. Smalheiser. An interactive system for finding complementary literatures: A stimulus to scientific discovery[J]. Artif. Intell., 1997, 91: 183-203.

[34] D. R. Swanson. Migraine and magnesium—Eleven neglected connections[J]. Pers. Biol. Med., 1988, 31: 526-557.

[35] D. R. Swanson. Somatomedin-C and arginine—Implicit connections between mutually isolated literatures[J]. Pers. Biol. Med., 1990, 33: 157-186.

[36] N. R. Smalheiser, D. R. Swanson. Assessing a gap in the biomedical literature—Magnesium-deficiency and neurologic disease[J]. Neurosci. Res. Commun.,1994, 15: 1-9.

[37] D. R. Swanson. Indomethacin and Alzheimer's disease[J]. Neurol., 1996, 46: 583.

[38] D. R. Swanson. Computer-assisted search for novel implicit connections in text databases[J]. Abstr. Papers Amer. Chem. Soc.,1999, 217.

[39] Chen C, Kuljis J, Paul R J. Visualizing Latent Domain Knowledge[J]. Transactions on Systems, Man, and Cybernetics -- Part C: Applications and Reviews. 2001, 31(4): 518-529.

[40] C. Chen, Information Visualization and Virtual Environments[M]. London, U.K.: Springer-Verlag London, 1999.

[41] C. Chen, R. J. Paul. Visualizing a knowledge domain's intellectual structure[J]. IEEE Computer, 2001, 34: 65-71.

[42] C. Chen. Visualization of knowledge structures[J]. Handbook of SoftwareEngineering and Knowledge Engineering, S. K. Chang, Ed, Singapore: World Scientific, 2002, 2: 700.

[43] R.W. Schvaneveldt. Pathfinder Associative Networks: Studies in Knowledge Organization[M]. Ablex Series in Computational Sciences, D. Partridge,Ed. Norwood, NJ: Ablex, 1990.

[44] R. W. Schvaneveldt, F. T. Durso, and D. W. Dearholt. Network structures in proximity data[J]. The Psychology of Learning and Motivation, 24, G. Bower, Ed. New York: Academic, 1989: 249–284.

[45] PRICE, D. J. DE SOLLA. Science Since Babylon[M]. New Haven, Conn: Yale University Press, 1961

[46] CRANE, D. Invisible Colleges: Diffusion of Knowledge in Scientific Communities[M]. Chicago: TheUniversity of Chicago Press, 1972

[47] MEADOWS, A. J., O'CONNOR, J. G. Bibliographical statistics as a guide to growth points in science[J]. Science Studies, 1971, 1: 95-99.

[48] GOFFMAN, W., HARMON, G. Mathematical approach to prediction of scientific discovery. Nature, 1971, 229(5280) : 103-104

[49] WAGNER-DÖBLER, R. William Goffman's "Mathematical approach to the prediction of scientificdiscovery" and its application to logic, revisited. Scientometrics, 1999, 46 (3) : 635-645.

[50] TABAH, A. N. Nonlinear dynamics and the growth of literature[J]. Information Processing and Management, 1992, 28 (1) : 61-73

[51] GARFIELD, E. Historiographic mapping of knowledge domains literature[J]. Journal of Information Science, 2004, 30 (2) : 119-145

[52] LEYDESDORFF, L. Top-down decomposition of the journal citation report of the social science citationindex: graph- and factor-analytical approaches[J]. Scientometrics, 2004, 60 (2) : 159-180.

[53] DIANA LUCIO-ARIAS, LOET LEYDESDORFF. Knowledge emergence in scientific communication from fullerenes to nanotubes[J]. Scientometrics, Scientometrics, 2007, 70(3): 603-632

[54] LEYDESDORFF, L., COZZENS, S. E. The delineations of specialities in terms of journals

using the dynamic journal set of the SCI[J]. Scientometrics, 1993, 26 : 133-154.

[55] FUJIGAKI, Y. Filling the gap between the discussion on science and scientist's everyday's activity: applying the autopoiesis system theory to scientific knowledge[J]. Social Science Information, 1998, 37 (1) : 5-22

[56] LEYDESDORFF, L., HELLSTEN, I. Metaphors and diaphors in science communication: Mapping the case of 'stem-cell research'[J]. Science Communication, 2005, 27 (1) : 64-99

[57] HESSE, M. Revolutions and Reconstructions in the Philosophy of Science[M]. Harvester Press, London, 1980.

[58] Chen C, Lin X, Zhu W. 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) : 359-377

[59] Small H. Tracking and predicting growth areas in science[J]. Scientometrics. 2006, 68(3): 595-610.

[60] J. Allan, J.G. Carbonell, G. Doddington, et al. Topic Detection and Tracking Pilot Study: Final Report[C]. Proc. DARPA Broadcast News Transcription and Understanding Workshop, Feb. 1998

[61] J. Allan, R. Papka, V. Lavrenko. On-line new event detection and tracking[C]. Proc.SIGIR Intl. Conf. Information Retrieval, 1998.

[62] D. Beeferman, A. Berger, J. La_erty. Statistical Models for Text Segmentation[J]. Ma-chine Learning, 1999(34): 177-210.

[63] Y. Yang, T. Ault, T. Pierce, et al. Improving text categorization methods for event tracking[C]. Proc. SIGIR Intl. Conf. Information Retrieval, 2000.

[64] Y. Yang, T. Pierce, J.G. Carbonell. A Study on Retrospective and On-line Event Detection Proc. SIGIR Intl. Conf. Information Retrieval, 1998

[65] V. Lavrenko, M. Schmill, D. Lawrie, et al. Mining of Con-current Text and Time-Series[C]. KDD-2000 Workshop on Text Mining, 2000.

[66] R. Swan, J. Allan. Extracting signi_cant time-varying features from text[C]. Proc. 8th Intl. Conf. on Information Knowledge Management, 1999.

[67] R. Swan, J. Allan, Automatic generation of overview timelines[C]. Proc. SIGIR Intl. Conf. Information Retrieval, 2000.

[68] R. Swan, D. Jensen. TimeMines: Constructing Timelines with Statistical Models of Word Usage[C]. KDD-2000 Workshop on Text Mining, 2000.

[69] S. Havre, B. Hetzler, L. Nowell. ThemeRiver: Visualizing Theme Changes over Time[C]. Proc. IEEE Symposium on Information Visualization, 2000.

[70] N. Miller, P. Wong, M. Brewster, et al. Topic Islands: A Wavelet-Based Text Visualization System[C]. Proc. IEEE Visualization, 1998.

[71] P. Wong, W. Cowley, H. Foote, et al. Visualizing sequential patterns for text mining[C]. Proc. IEEE Information Visualization, 2000

[72] Kleinberg J. Bursty and Hierarchical Structure in Streams[J]. Data Mining and Knowledge

Discovery. 2003, 7(4): 373-397.

[73] Chen C. Detecting and Mapping Thematic Changes in Transient Networks[C]. London: IEEE Computer Society Press, 2004.

[74] Noyons E C M, van Raan A F J. Bibliometric cartography of scientific and technological developments of an R & D field[J]. Scientometrics. 1994, 30(1): 157-173.

[75] Cassiman B, Glenisson P, van Looy B. Measuring Industry-Science Links through Inventor-Author relations: A Profiling Methodology[R]. Catholic University of Leuven (KUL) - Department of Economics, 2006.

[76] Ed J.Rinia, Thed N.Van Leeuwen. Measuring knowledge transfer between fields of science[J]. Scientometrics, 2002, 54(3): 347-362

[77] STEELE, T. W., J. C. STIER. The impact of interdisciplinary research in the environmental sciences: a forestry case study[J]. Journal of the American Society for Information Science, 2000, 51 (5): 478-484.

[78] MORILLO, F., M. BORDONS, I. GOMEZ. An approach to interdisciplinarity through bibliometric indicators[J]. Scientometrics, 2001, 51 (1): 203-222.

[79] PIERCE, S. J. Boundary crossing in research literatures as a means of information transfer[J]. Journal of the American Society for Information Science, 1999, 50 (3): 271-279.

[80] NATIONAL SCIENCE BOARD. Science & Engineering Indicators – 2000[R]. Arlington, VA, National Science Foundation, 2000: 6-45.

[81] KOSTOFF, R. N., J. A. DEL RIO. The impact of physics research[J]. Physics World, 2001, 14 (6): 47-51.

[82] Leot Leydesdorff. “betweenness Centrality” as an indicator of the interdisciplinarity of scientific journals[C]. The 9th international conf. on science & technology indicators, Leuven, Belgium, 2006.

[83] Meyer M. Tracing knowledge flows in innovation systems[J]. Scientometrics. 2002, 54(2): 193-212.

[84] Chen C, Hicks D. Tracing knowledge diffusion[J]. Scientometrics. 2004, 59(2): 199-211.

[85] Callon M, Courtial J P, Laville F. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry[J]. Scientometrics. 1991, 22(1): 155-205.

[86] WASSERMAN, S., K. FAUST. Social Network Analysis: Methods and Applications.[M] Cambridge University Press, Cambridge, 1994.

[87] Otte, E., & Rousseau, R. Social network analysis: a powerful strategy, also for the information sciences[J]. Journal of Information Science, 2002, 28(6): 443-455.

[88] Freeman, L. C. A Set of Measures of Centrality Based on Betweenness[J]. Sociometry, 1977, 40(1): 35-41

[89] Freeman, L. C. Centrality in Social Networks. Conceptual Clarification[J]. Social Networks, 1978, 1: 215-239.

[90] De Nooy, W., Mrvar, A., Batagelj, V. Exploratory Social Network Analysis with Pajek[M].

New York: Cambridge University Press, 2005.

[91] Hanneman, R. A., & Riddle, M.. Introduction to social network methods. Riverside, CA: University of California, Riverside, 2005.

[92] Van den Besselaar, P., & Heimeriks, G. Disciplinary, Multidisciplinary, Interdisciplinary: Concepts and Indicators[C]. Proceedings of the 8th International Conference on Scientometrics and Informetrics - ISSI2001, M. Davis & C. S. Wilson (Eds.). Sydney: University of New South-Wales, 2001: 705-716.

[93] 同52

[94] 同82

[95] Katarina Larsen. Knowledge network hubs and measures of research impact, science structure, and publication output in nanostructured solar cell research[J]. Scientometrics, 2007.

[96] Chen, C.. Searching for intellectual turning points: Progressive knowledge domain visualization[C]. Proceedings of the National Academy of Sciences, 2004, 101(Suppl. 1): 5303-5310

[97] Chen C. Visualization and presentation: The centrality of pivotal points in the evolution of scientific networks[C]. IUI'05, San Diego,California, USA, Junuary 9-12, 2005: 98-105.

[98] Morris, S. A., & Yen, G. Crossmaps: Visualization of overlapping relationships in collections of journal papers[C]. Proceedings of the National Academy of Sciences of the United States, 101(suppl. 1): 5291-5296.

[99] Morris, S. A., Yen, G., Wu, Z & Asnake, B. Time line visualization of research fronts[J]. Journal of the American Society for Information Science and Technology, 2003, 54(5): 413-422.

[100] Morris, S. A., & Boyack, K. W. Visualizing 60 years of anthrax research[C]. 10th International Conference of the International Society for Scientometrics and Informetrics. Stockholm, Sweden: Karolinska University Press: 45-55

[101] 同99

[102] 杨立英.科技论文共现理论研究与应用[D].北京:中国科学院文献情报中心,2007.

[103] Chaomei Chen, Steven Morris. Visualizing Evolving Networks: Minimum Spanning Trees versus Pathfinder Networks[C]. Proceedings of the IEEE Symposium on Information Visualization 2003 (INFOVIS'03).

附录 1 寻找潜在簇的结果

寻找潜在簇的结果（部分）：以 ESI 中 COMPUTER 领域的数据为例，高阈值层设定为 0.1，低阈值层设定为 0。

潜在簇	类型	本阈值层中的 relative	与潜在簇有潜在关系的高阈值层簇（twin of relative）
E3	绝对孤立簇		
E13	自成体系孤立簇		
E14	分支衔接簇	N46	
		N31	
		N38	N5
		N17	
E15	分支衔接簇	N11	
		N23	N1
		N4	
		N43	
		N39	
		N37	
E40	直接衔接簇	N45	N10
		N35	N9
		N41	
		N28	N8
		N36	
		N33	
		N22	
		N23	N1
		N29	
		N43	

附录 2 寻找演变簇的结果

寻找演变簇的结果：ESI 中 COMPUTER 领域的数据，起始时间 1996，终止时间 2005，时间窗跨度为 5 年。当前时间窗是 2000－2004，祖先时间窗是 1999－2003，后代时间窗是 2001－2005。

演变簇	类型	祖先	后代
N2	新增簇		
N20	分化簇		N24
			N2
N26	合并簇	N3	
		N8	
		N9	
	分化簇		N7
			N14
			N2
N27	合并簇	N18	
		N22	
N29	分化簇		N20
			N26
			N23
			N17
N30	新增簇		

博士在读期间发表论文和参与科研课题情况

（一）发表论文

1. 韩涛.WSRF 标准规范体系研究,现代图书情报技术,2007,5
2. 韩涛.WSRF 和 WSMF 两种 Web 服务技术框架的比较研究,情报科学,2007,25(2)
3. 韩涛.Web 服务及其质量的本体描述,情报理论与实践,2007,30(3)
4. 韩涛.科学数据与科学文献相关关系研究—以生物信息学为例,图书情报知识,2008,3
5. 孙志茹,韩涛,杨文. 生物信息学科学数据与科学文献的关联关系分析,图书情报工作,2008,52(363)
6. 杨文,韩涛,孙志茹. 生物信息学序列库与文献库的融合模式浅析,情报理论与实践,2008,31(1)
7. 吴清强,韩涛.数字图书馆评价研究综述,现代图书情报技术,2006,6
8. 王小梅,吴清强,韩涛.情报分析平台的集成化实践.现代图书情报技术,2007,7

（二）参与课题

1. 中国科学院战略情报研究平台（2007.1 – 至今）
2. 情报研究中调研数据资源的信息化建设（2006.7－2006.10）
3. 科技战略情报基础分析平台（2006.3－2007.6）
4. 中国数字图书馆标准与规范建设项目（项目编号：005DKA43503；2006.9 －2006.12）
5. 储氢材料发展态势研究（2005.12－2006.3）：
6. 科学数据与科学文献相关关系研究（2006.10－2007.6）

致谢

时间飞逝，在中国科学院文献情报中心三年的博士研究生的学习和生活即将结束。这三年系统的学习和工作使我在科学研究和项目开发方面的能力得到了锻炼和提高，从中获取了很多宝贵的知识和经验。

首先，我要感谢我的导师张晓林教授，他活跃的思维，严谨求实的作风，敏锐的洞察力，精湛的学术造诣，创新的科研精神深深地影响着我。张老师忘我的工作热情，乐观的生活态度也为我树立了很好的人生榜样。自课题立项到研究，直至论文的撰写修改，无不得到了张老师的悉心指导。张老师的谆谆教导使我能顺利地完成研究工作，在此我向张老师表示衷心的谢意！

我还要感谢冷伏海教授、金碧辉教授、李广建教授、孟连生教授等各位从事图书馆学、情报学、信息管理、信息系统建设等工作的老师们，他们务实的科研态度和丰硕的科研成果都对我有很大的启发和鼓舞，同时还要感谢他们在我论文开题、中期考核、预答辩等过程中提出的许多宝贵意见和中肯的批评，感谢他们给予我的悉心指导。感谢匿名评审的几位专家提出的宝贵意见和建议。

感谢情报研究部的张薇老师、赵亚娟老师、董瑜老师，指引我步入情报研究这个陌生的学科，为我提供良好的工作环境和丰富的实践机会。感谢情报研究部的阳宁辉老师、王俊老师给我在工作和生活上的关怀和帮助。感谢情报研究部这个集体，她团结奋进的氛围令我在精神上一次次受到洗涤和鼓舞。

感谢信息系统部的王小梅老师、情报研究部的谭宗颖老师、资源建设部的刘筱敏老师，在三年的项目开发实践中，我多次向她们打扰请教，她们每次都耐心地给予我指导。她们具有丰富的工作经验，扎实的工作能力，向她们学习请教使我积累了实践经验，增长了工作能力，这是我三年获得的又一宝贵财富。

感谢人力资源处的张彦秋老师、张章老师，还有退休的王静珠老师。三年来，他们在学习、生活等方面都给了我莫大的帮助和照顾。在我犯错误的时候给了我及时的敦促和支持，还有最大的包容和理解。他们细心的管理工作，耐心的谈话教导，为研究生营造了一个轻松、活泼、和谐的学习和生活环境，使我度过了三年美好时光。

感谢课题组的各位同学，是他们共同创造了一个良好的学习氛围。特别要感谢同门吴清强、李书宁、张枚、郭文丽、陈颖、陈仕吉等同学，能和这些志同道合的同学愉快共事，对我而言是一段非常宝贵而且难忘的人生经历。感谢来自台湾的顾立平同学，平时与他探讨大陆与台湾、中国与外国的文化差异，休闲时一起打篮球锻炼，为我的生活增添了几分色彩。

感谢中心的其他同学，一起参与课题的孙志茹、杨文，热爱运动的张树良、黄国彬，乒乓过招的赵凡、杨波，还有汪丹、侯丽、李亚子，毕业的曲云鹏、彭颢舒、魏晓俊等同学，一起度过的美好时光将终身难忘。

感谢中心的工会主席唐宏瑞老师，与我一同参与工会工作的吴霞、黄飞燕同学，他们热情为他人服务的精神是我学习的榜样。

感谢我的家人对我的关心、支持和爱护。感谢多年来在我漫漫求学路上给予我默默支持的母亲、姥姥、姨妈，他们殷切的目光、信任的眼神、温暖的叮嘱都是我努力求学的巨大动力。感谢生活并不富裕的堂姐、伯父、伯母、堂哥等亲人，在我最需要帮助的时候替我承担本应由我承担的重担，让我能在外安心求学。

感谢我的爱人邓小艳对我的支持、鼓励和理解。三年来，她默默承担家庭的担子，从无怨言，在我松懈、气馁的时候给我鼓劲、鞭策。她为我营造了一个舒适、温暖、贴心的家，永远是我最坚强的后盾。感谢岳父、岳母，还有爱人的爷爷、奶奶，是他们在背后的默默奉献让我重新体会到大家庭的温馨，使我精神上得到巨大的慰藉。

最后还要感谢我所有的老师、同学、亲人、朋友，我的成长离不开他们的帮助和关心，我今天的收获也有他们的一份功劳。

向参考文献的著者们表示诚挚的谢意。

韩涛

2008 年 5 月