共现聚类分析结果的判读

概念与关系的逻辑思维

供中国医科大学 2009 级研究生《文本挖掘在科研选题中的应用》课程使用

崔雷

2009-10-14

表现文献内容的标识（如关键词、分类号）可以用根据它们共现的情况进行聚类分析，一些表现文献外部特征的标识，如作者、引文等等，也可以进行共现分析，如作者的合著分析、引文的同被引分析、作者的同被引分析，这些都可以为展示某一学科领域里科学研究获得的结构和特点提供手段。这些分析的方法原理都是大同小异的，都属于基于共现的聚类分析。比较常用的方法是通过对特定领域或者学科的**高频主题词**以及**高被引论文**进行共现聚类分析，通过这些分析，可以把主题词和论文分类为不同的群组，客观地反映出这些概念（主题词）或者概念群（论文）之间的亲疏关系。但是，如何解读这些聚类结果，由此反映出当前该领域研究的结构和热点，成为共现聚类分析的一大难点。本文就是针对这个问题，总结多年来共现聚类分析实践所积累的经验，提出一些判读聚类分析结果的基本原则和体会，共同道在使用中参考。

一．词共现聚类分析的结果判读

以使用SPSS进行聚类分析为例，可以选择Analysis-Classify-Hierarchical过程，经过设置相应的参数后，对胃癌治疗的高频主题词共现矩阵进行分析，最后获得该研究领域高频主题词的共现聚类分析树图（如图1）。

1.聚类树图的结构分析

首先从宏观上观察聚类树图的结构。聚类树图中的最左边的一列标号(Label)和数字（Num）代表着高频主题词，由于采用的是系统聚类法的凝聚聚类算法，因此，最初每一个主题词都是单独的一个类，通过计算每一对主题词之间的相似性，发现2号和3号主题词的相似性在所有主题词词对之间是最小的，因此，它们首先聚集成为一个类，然后它们又和7号主题词合成为一个类。图中最上方的带有数字的标尺表示分类对象之间的距离（在SPSS中是重新量化计算的）。随着被分类的对象（主题词）之间的距离越来越大，最终所有的主题词都成为一个类，我们可以根据需要在不同的距离水平上分割整个聚类树图。

通过树图的结构我们可以看到，所有的主题词从整体上可以分为三个部分：由2、3、7号词组成的一个类别（A），由1、4、8、5、9号主题词组成的一个类别（B），和由6号词单独组成的一个类别（C）。

2.各类的内容分析

主要是通过各个类别主题词之间语义关系的分析。基于凝聚聚类算法的原理，对聚类分析结果的语义分析也采用了“自下而上”的步骤。即首先获取各个小类的含义，然后将各个小类的含义在语义上叠加而组合成为大类的含义。具体而言，就是首先从每个小类中关系最近的两个主题词着手，分析二者之间的语义关系，获得该类的“种子”概念，在“种子”概念的基础上，根据同类别中其他主题词与该“种子”的距离，逐次加入主题词，丰富该类别的内容，一般而言，距离比较远的主题词往往是该核心的相关因素，如核心概念的具体的应用或者影响因素。

本例中，对于3个高频主题词的类别中的主题词进行具体的语义分析，可以发现：

在A类中，“Stomach Neoplasms/drug therapy，胃肿瘤/药物治疗”（2）与 “Antineoplastic Combined Chemotherapy Protocols/therapeutic use，抗肿瘤联合化疗方案/治疗应用”（3）组合在一起表明的是对胃肿瘤采用联合化疗，加上“Adenocarcinoma/drug therapy，腺癌/药物治疗”（7）表明这一类主要是关于胃腺癌的联合化疗的主题。

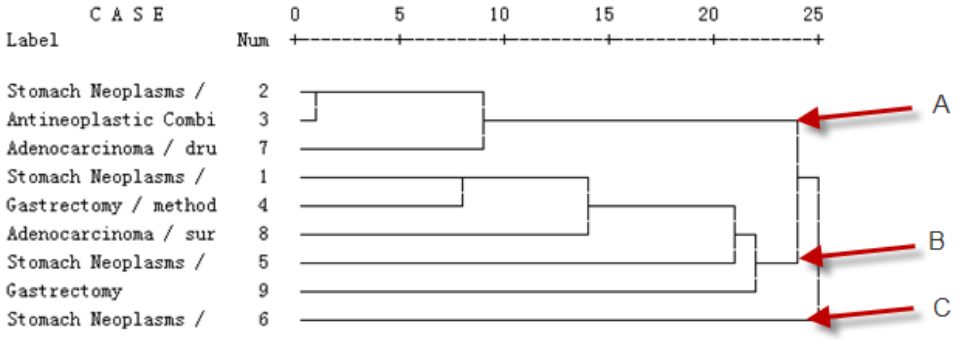

图1 胃癌治疗高频主题词聚类分析结果（部分）

在B类中，“Stomach Neoplasms/surgery，胃肿瘤/手术”(1)与“Gastrectomy / methods，胃切除/方法”（4）表明是关于胃肿瘤外科手术切除治疗的主题，同样，“Adenocarcinoma/surgery，腺癌/手术”（8）也是说明主要是针对胃腺癌的手术切除治疗，“Stomach Neoplasms/pathology，胃肿瘤/病理”则是说明这种疗法会受到不同的病理类型的影响。这一类别最边缘的主题词“Gastrectomy，胃切除”（9）没有副主题词，则是表示对胃切除的概述，是对上述主题的强化，由于与其他主题词距离比较大，也可以忽略不计。

C类则是单独一个单词，“Stomach Neoplasms/therapy，胃肿瘤/治疗”（6）则表示的是对各种治疗方法的概述性研究主题。

综上，我们可以总结出目前胃癌治疗研究的主要方向有三个：（1）胃癌胃腺癌的联合化疗，（2）胃腺癌的手术切除治疗（包括病理因素的影响），（3）胃肿瘤各种治疗方法。

如果在对聚类结果的分析中还有无法解释说明的主题词组合，可以采用下面的辅助方法：（1）到相应的文献数据库中检索含有共现词对的文献，通过文献的相应内容分析，寻找文献中共现词对的关系。（2）请教对本专业研究状况比较熟悉的专业人员，请教他们对专业词汇组合的含义。

二．同被引聚类分析结果的判读

对于高被引论文的同被引聚类分析，也应该遵循结构分析和内容分析两个步骤。

1.聚类结构的分析

与词共现分析一样，遵循聚类分析树状图的整体结构。不同之处在于这里的聚类对象由主题词变成了高被引的论文，根据它们被后来发表的论文同时引用次数的多寡，这些高被引论文被聚集到不同的类别之中。有研究成果表明，论文之间的同被引强度越大，它们在主题上相关性就越大。因此，同被引聚类分析的树图表明了高被引论文在主题上的相关程度。以下图为例，所有的高被引论文被分成了4个大类，其中A、C、D三个大类还可以进一步分为两个小类。需要结合语义内容的分析，决定类别的数目和各个大小。

2.聚类内容的分析

对每个类别的内容分析，也遵循“自下而上”的原则，以类内最小距离的两篇论文为核心，逐渐扩展丰富类别的内容描述。人工阅读同类论文的标题、文摘

甚至全文，分析这些文献的共同之处；同时，还要站在全局的角度，分析这一类的论文在主题上与其他类别的不同之处。

处于各个类的边缘的论文，一般代表了该核心的某个方面，或者为影响因素或者为主要应用领域。应注意揣摩各个论文的内容，分析论文与同类论文和不同类论文之间的关系，即：对属于同类的论文，寻找它们的共性，以此形成该类别的类标签；同时，对不同类的论文，分析该论文与不同类之间的差异。由于聚类分析是一种经验性很强的非监督学习方法，聚类的效果也会有不尽如人意的地方，对于某些类别中距离比较大的，远离中心的个别论文可以舍弃。

例如，图2是对近年来健康管理研究领域高被引论文的同被引聚类分析结果。其中第一类的论文标题如表1。第一大类包括两个小类，其中第一个小类的核心是论文12和34，其标题中都含有“健康素养”这个概念，其中12号论文是侧重健康素养的测量方法，而34号论文则侧重对具体人群测量健康素养的实践。二者相辅相成，都是对健康素养进行定量的测量，针对某种服务（管理医疗）或者某些人群（慢性病/糖尿病）进行比较分析。第二小类的核心则是17号和35号两篇论文，从图2中可以看出两篇论文在很远的距离上聚集成为一类（与12号和34号聚集的距离相比较），说明二者关系不似第一小类那样紧密；从标题上分析基本上是关于健康促进项目的实施，而且对高胆固醇进行预防是这一类研究的重点。这两个小类合在一起所形成的第一类的标签还需要与其他类的内容进行比较后最后形成。

表1 健康管理类高被引论文标题及其类别标签

序号	论文标题	类标签
12	功能性健康素养的测试方法的设计与实现	健康素养：具体应用于对保险参加者以及糖尿病患者的疗效等。
34	在管理医疗组织中医疗保险参与者的健康素养	
46	健康素养与糖尿病治疗结果的关系	
41	对 2 型糖尿病患者的自我管理进行培训的效果：一个随机对照试验的系统综述	
13	对慢性病患者实施管理医疗服务	
37	纵向研究中对预兆性地同患多种疾病进行分类的新方法：设计与实现	
2	MOS SF-36 健康状况评估量表：概念框架和项目选择	
36	医疗服务项目的经济学评价方法	
17	国家胆固醇教育计划关于第三次成人高胆固醇的检测、评估和治疗调查报告	健康促进项目：主要应用领域为高危人群（高胆固醇）预防服务。
35	MRC/BHF心脏保护研究之20536名高危个体服用辛伐他汀降低胆固醇的试验	
20	临床预防服务指南	
16	健康促进项目的经济学影响：文献综述	

如果采用上述方法分析每一个类别的主要内容仍然存在困难，可以采用如下辅助方法：

（1）统计同类论文的词频：有条件的话，分析同类别论文的词频，高频词表示该类的主要内容，甚至采用向量空间模型的算法来区分各个类别在主题内容上的特色。

（2）请教专业人员：对难于理解的主题领域，适时请教资深的阅读文献比

较多的专业人员，尤其要注意说明本研究的意图和聚类的意义。

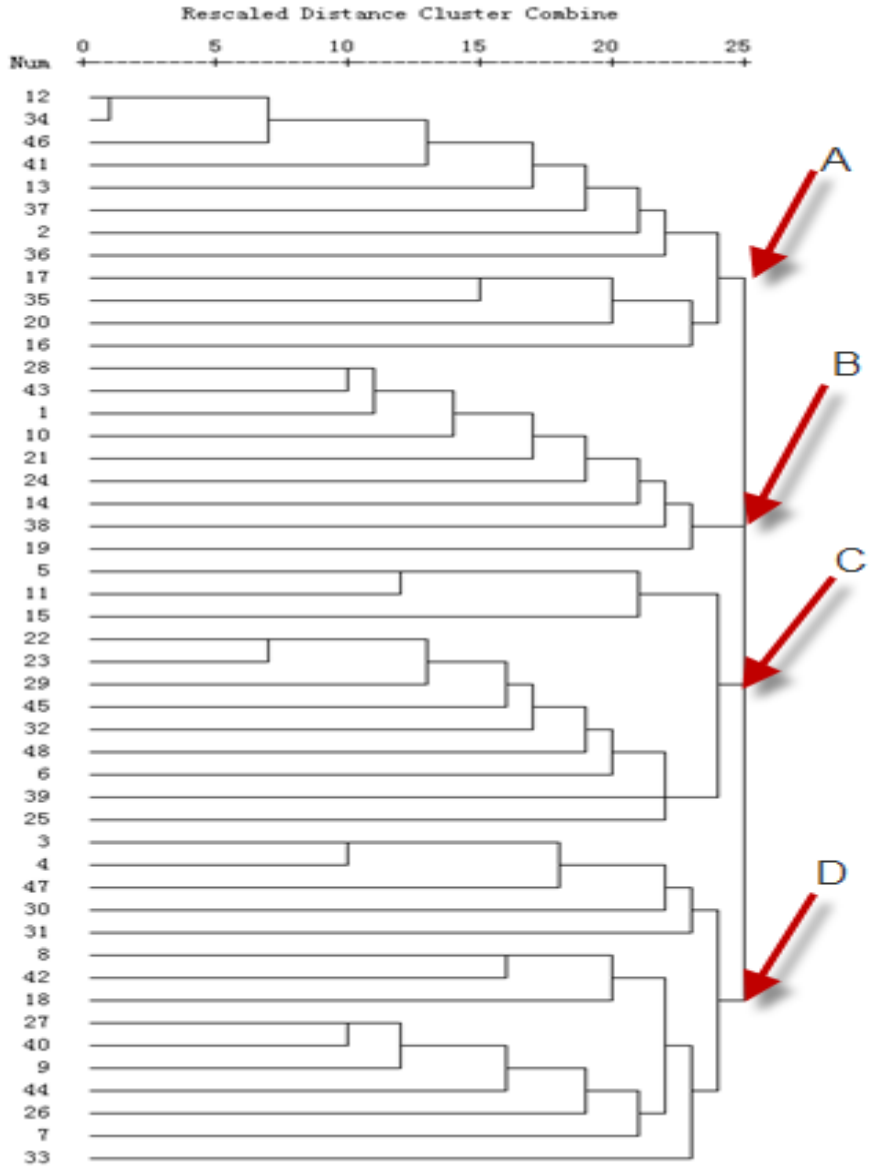

图2 健康管理领域高被引论文聚类分析树图

综上所述，对于共现聚类分析结果的判读分析，主要遵循的原则就是在掌握所分析课题的背景知识的基础上，依照系统聚类方法的原理，划分出大类，然后对每一大类则采取自下而上的语义叠加的方法，从每一类最小距离的核心开始逐步拓展丰富该类别的内容；同时，要兼顾整个分析课题所在领域的总体结构和与其他类别的区别和联系，对该类别赋予合适的类别标签。

这是一项需要心平气和与足够智慧的脑力体操，希望你能从中得到发现的快乐。