词频分析法中高频词阈值界定方法适用性的实证分析

刘奕杉，王玉琳，李明鑫
（东北师范大学信息科学与技术学院，长春 130117）

摘要：词频分析法是文献计量学的重要分析方法之一，而确定高频词阈值是进行词频分析的必要前提，高频词阈值的选取不仅决定词频分析法的分析结果，而且对整个分析研究都有着极其重要的影响。本文首先以近三年国内运用词频分析法展开研究的文献为调研基础，发现目前学界常用的高频词阈值选取方法主要有自定义选取法、高低频词界定公式选取法、普赖斯公式选取法及混合选取法四类；其次，以个人知识管理领域的文献为研究对象，对前三类高频词阈值选取方法分别进行取值计算并做领域热点聚类分析，对比验证聚类结果，同时以此结果为基础讨论高频词阈值选择对分析结果的影响及其合理性；最后，指出我国学界在高频词阈值选取方面存在主观性强、方法原理不明、改进方法适用性不明，高低频词界定公式和普赖斯公式适用性尚待研究等问题。

关键词：高频词；文献计量学；词频分析
中图分类号：G250
DOI：10.3772/j.issn.1673-2286.2017.09.007

1 引言

在科学研究中，常通过表达文献核心内容的关键词或主题词的出现频次确定该领域的研究重点和发展动向。由于一篇文献的关键词或主题词是文献核心内容的浓缩和提炼，因此，如果某一关键词或主题词在其领域文献中反复出现，则可认为该关键词或主题词所表征的研究主题即为该领域的研究热点[1]。词频分析法可以结合其他方法（如共词分析、多维尺度分析、知识图谱等），加深对研究主题的理解。虽然词频分析法的实践应用广泛，但很少有对其方法理论的深入研究，缺乏对其内涵、特征、模式、流程等内在规律的系统归纳。此外，对词频分析法与传统文献综述法在方法论基础、研究对象、应用范围等方面的探讨亦比较少见[2]。

确定领域高频词是运用词频分析等方法进行下一步工作的基础，因此如何合理界定领域高频词成为重要的研究课题。如杨建林对基于词频阈值和基于贡献强度阈值的两种选词策略进行分析，得出将这两种方法合并后得到的关键词集具有更好的共词分析效果[3]；陈果等提出基于学科背景的全局视角，对比关键词在领域内外的出现频率，提出领域度计算公式，并融合领域度和热度指标进行关键词筛选[4]；安兴茹提出基于正态分布的方法，通过实证分析，验证关键词或主题词在文献库中的分布符合正态分布，并根据正态分布的特性，提出词频分析法高频词阈值的计算方法[5]。

目前已有学者尝试提出改进高频词阈值的方法，但这些新方法是否具有广泛的适用性，是否能解决目前高频词阈值选取中存在的问题，以及使用这些新方法是否会产生新的问题，在学界尚无法达成共识，还需要继续探讨；而传统高频词阈值选取方法相对不规范，因此关于高频词阈值的选取方法未来还有很长的路要走。

2 常用高频词阈值选取方法

为反映目前我国学界关于高频词阈值选取方法的现状，本文在中国学术期刊网络出版总库中检索“研究热点”相关的文献。以摘要=“热点”and主题=“词频+共词”为检索式，选取来源类别为CSSCI，检索时间为

2015—2017年的文献，共得到229条记录，再通过人工筛选，去除不符合研究主题的文献，最终得到174篇文献。

2.1 近三年“研究热点”类文献的统计结果分析

本文通过提取174篇文献中高频词阈值的方法，并以此为代表，整理目前我国学界常用的高频词阈值选取方法，结果见表1。

表 1 高频词阈值选取方法

方 法	数量/篇	占比/%
频次选取法（自定义选取法）	80	45.98
前N位选取法（自定义选取法）	44	25.29
中心度选取法（自定义选取法）	5	2.87
高低频词界定公式选取法	11	6.32
普赖斯公式选取法	5	2.87
频次选取法+前N位选取法（混合方法）	5	2.87
高低频词界定公式选取法+频次选取法（混合方法）	2	1.15
普赖斯公式选取法+前N位选取法（混合方法）	4	2.30
普赖斯公式选取法+频次选取法（混合方法）	2	1.15
高低频词界定公式选取法+普赖斯公式选取法（混合方法）	1	0.57

注：有15篇文献未提供具体研究方法，故不在本文研究之列。

2.2 自定义选取法

从表1可以看出，目前我国学界在研究领域热点问题时，常用的高频词选取办法是自定义选取法，合计129篇，占比74.14%。自定义选取法，作者可根据研究需要自行规定高频词的选取方法和高频词的阈值，这种选择方法主观性强，在阈值的选择上较随意。通过本文所得到的174篇文献的研究数据发现，样本文献数据量从58—25 990篇，频次的选择从2—300次，跨度比较大。对这些具有一定随意性的高频词选取方法所选出的高频词进行分析，其分析结果的准确性和科学性值得商榷。即使是同一领域的研究，也存在不同研究者有不同取值标准的现象，从而导致研究结果不一致。

2.2.1 频次选取法

从调研结果来看，最常用的自定义方法是频次选取法，即作者自行规定高频词的阈值，这类文献占比45.98%。这种高频词选取方法主要依据研究者在研究过程中遇到的具体情况和自身经验，选取合适的阈值来确定高频词。这种方法的优点是操作简便，可节省大量时间和精力，使研究者把更多注意力放在后续分析研究上。但由于此种方法的全部操作步骤均为研究者自定义，其可信度和科学性无法保证，尤其高频词阈值的确定是后续分析研究的基础。

在现有样本数据中，有11篇文献的研究者在使用频次选取法时，按照高频词累计频次达到总频次40%左右的取词标准进行取词，占频次选取法文献的13.75%，全部样本文献的6.32%。由此也可以看出，在频次选取法的实际应用中，研究者的主观意愿在一定程度上占据主导地位。

2.2.2 前N位选取法

前N位选取法即按照词频由高到低进行排序，作者自选前N位词为高频词；这类文献共44篇，占比25.29%。这种方法与频次选取法类似，也是以研究者主观意志为主的一种高频词选取方法。

不同的是，这种方法的随意性更大。前N位选取法中N的阈值如何界定，目前没有标准。从本文样本统计结果来看，高频词阈值选取标准从前5—100位不等，其所选第N位高频词的出现频次也从2—100次不等。由于这种方法是将具体频次数据抽象为排名形式，因此不可避免地丢失部分具体频次信息。这种更抽象的前N位选取法，通常使研究者更易忽略其截取频次的合理性，而更关注所选高频词个数是否更易构造相异矩阵，是否能够为研究带来更多的方便。

2.2.3 中心度选取法

目前，由于词频分析软件的普及，在进行词频分析时，大量文献选择把原始数据直接导入词频分析软件中（如CiteSpace、Ucinet等），以关键词中心度为排序依据选取高频词的样本数据共5篇，占比2.87%。实际上，CiteSpace等词频分析软件的工作原理是根据词频多少来确定相应的节点中心度，因此这种以中心度确定高频词的方法其实质与前N位选取法的原理一致。

2.3 高低频词界定公式选取法

第二大类方法是用高低频词界定公式确定高频词

阈值。高低频词界定公式由Donohue在1973年提出，源于齐普夫第二定律[6]。高低频词界定公式作为文献计量学里的一项重要内容，本应是用于高频词阈值界定的一种普遍方法，但从本文样本调研结果来看，实际上使用此高低频词界定公式法进行高频词选取的文献只有11篇，仅占比6.32%。

以齐普夫第二定律为基础的高低频词界定公式：$T=\frac{1}{2}(-1+\sqrt{1+8\times I_1})$。其中$I_1$指数据中只出现过1次的关键词数量。使用此高低频词界定公式出现取值不理想问题可能有两种原因。

（1）依赖I_1。从此公式可知，词频阈值完全由I_1决定，计算出的高频词个数一般过多或过少，这可能是计算结果不理想的外在原因[5]。

（2）理论基础和适用性。高低频词界定公式是由齐普夫第二定律得来，同齐普夫第二定律一样都是针对某一具体文献得出的词频分界公式，而非针对文献库得出，所以高低频词界定公式在基于文献库的高频主题词和关键词阈值计算上，缺乏理论基础和适用条件[5]。

相较于自定义选取法，用高低频词界定公式计算高频词比较复杂，虽然孙清兰对其进行过改良，提供相对简便的算法[6]，但与自定义方法相比，高频词的选取仍然比较复杂。因此，由于上述两方面原因，学界较少使用高低频词界定公式方法界定领域高频词。

2.4 普赖斯公式选取法

虽然在选择领域高频词方面有许多学者提出多种方法，如熊回香等使用大数据搜索与挖掘共现平台提取特征词[7]，巴志超等用LDA和word2vec模型提取高频词[8]，姚小娇用词频g指数方法界定高频词等[9]。但就目前统计来看，除作者自定义和用高低频词界定公式界定高频词这两种方法外，还主要借用普赖斯公式来确定领域高频词（共计5篇，占比2.87%）。

普赖斯公式最早被用于确定高被引文献，进而确定某研究领域内的核心作者。因方法相较于用高低频词界定公式更简单，比自定义选取法更科学，逐渐被T学者接受并应用于不同领域的研究中。其高频词阈值根据普赖斯公式确定，计算公式：$M=0.749\sqrt{N_{max}}$，其中M为高频词阈值，N_{max}表示区间学术论文被引频次最高值[10]。

普赖斯公式可以用于确定领域核心文献，因此在一定程度上利用此公式确定领域核心关键词也是可行的。但在具体应用时，还需进一步改进，如将自变量N_{max}表示为关键词的频次最高值，而不是被引次数的最高值，这样用此公式得出的结果才更符合“领域核心词”（即领域高频词）。

为更清晰地表现上述我国学界常用的三类（5种）高频词阈值选取方法，在此将这5种方法归纳、对比，整理见表2。

2.5 混合选取法

混合选取法指将表2中两种或两种以上方法合并使用的情况。最常用的方法为先用高低词频界定公式或普赖斯公式计算得出一个高频词阈值，由于该阈值与实际情况存在一定差异，不能满足研究需要或为研究带来不必要的麻烦。对此，研究者通常根据实际情况进行人工选词，即在公式计算结果的基础上，人工扩大或缩小范围，自定义阈值。

表 2 5种常用高频词阈值选取方法对比

方 法	定 义	特 征	文献来源
频次选取法	自选频次大于x次的词为高频词	操作简便，但无法保证其可信度和科学性	文献[11]
前N位选取法	依据频次由多到少排序，自选前N位词为高频词	将具体频次数据抽象成排名，丢失部分具体频次信息，更易忽略其截取频次的合理性	文献[12]
中心度选取法	以关键词的中心度排序，自选前m个词为高频词	将具体中心度数值和频次数值抽象为排名，丢失部分中心度及频次信息，更易忽视其合理性	文献[13]
高低频词界定公式选取法	依据高低频词界定公式计算高频词阈值	高低频词界定公式计算阈值可以保证科学合理性，但不易操作，作者自选关键词造成大量频次为1的关键词出现，影响计算结果	文献[14]
普赖斯公式选取法	依据普赖斯公式计算高频词阈值	以最高被引次数确定高频阈值，无法保证科学性和合理性	文献[10]

3 高频词阈值选取的实证分析

本文以上述总结出的我国学界目前常用的三大类高频词选取方法为基础，对个人知识管理领域的研究文献进行实证研究，并将所得结果进行对比验证，以此说明不同高频词阈值选取方法对选词结果的影响，以及不同选词结果对后续分析研究的影响。本文仅通过聚类分析方法来体现其影响，对于多维尺度分析、网络节点分析等定量方法，以及领域热点、发展趋势分析等定性方法暂不予以说明。

本文通过中国知识基础设施工程的期刊数据库，检索得到“个人知识管理”领域的文献，共得1 241条记录。用Bicomb软件提取所有文献的关键词[15]，经过数据清洗后共得2 412个关键词，将词频出现频次按由高到低排序，部分结果（Top 20）如表3所示。

表 3 个人知识管理领域关键词词频统计表（Top 20）

序 号	关键词	出现频次	占比/%	累计占比/%
1	博客	74	1.63	1.63
2	隐性知识	71	1.56	3.19
3	个人知识	66	1.45	4.64
4	图书馆	52	1.14	5.78
5	Web 2.0	51	1.12	6.90
6	知识共享	48	1.05	7.95
7	显性知识	44	0.97	8.92
8	Blog	43	0.94	9.86
9	大学生	39	0.86	10.72
10	知识	31	0.68	11.40
11	知识管理系统	29	0.64	12.04
12	社会性软件	29	0.64	12.68
13	教师专业发展	28	0.61	13.29
14	教师	28	0.61	13.90
15	个人学习	21	0.46	14.36
16	教师知识管理	20	0.44	14.80
17	信息技术	20	0.44	15.24
18	图书馆员	20	0.44	15.68
19	组织知识	19	0.42	16.10
20	知识经济时代	19	0.42	16.52

3.1 二八定律验证自定义选取法

从本文第一部分分析来看，绝大多数自定义选取法凭研究者意愿进行。但根据统计，自定义选取法的词频截取比例通常在8%—40%。因此，为重现自定义选取方法并同时体现一定的取词依据，本文以二八定律为基础，对自定义选取法进行实证验证，选取个人知识管理领域的高频词。依照表3中的统计结果，从高到低抽取累计占比达20%的词为该领域的高频词。

按照此种方法抽取高频词，应抽取的高频词范围为所有频次大于或等于15的词，即位于前30位的词为个人知识管理领域的高频词，累计占比20.14%。用SPSS软件对此30个高频词进行聚类分析，以倒数第二大聚类层次作为划分标准，统计聚类类别。

在选取前30个词为高频词的条件下，个人知识管理领域的关键词大致可以分为三类，即“显性知识”“隐性知识”与“图书馆”为第一类；“个人知识”“组织知识”“个人学习”“组织学习”“知识经济时代”与“知识结构”为第二类；其余如“博客”“应用”与“策略”等为第三类。

3.2 高低频词界定公式选取法的验证

据统计，关键词词频为1的共有1 860个词，将其代入高低频词界定公式，计算得出高频词阈值为60，即由高低频词界定公式确定的个人知识管理领域的高频词为所有出现频次大于60次的词汇。因此，如果按照高低频词界定公式方法取词，那么个人知识管理领域的高频词为表3中的前3个词，即“博客”“隐性知识”与“个人知识”。由于此方法只提取到3个高频词，不便进行聚类分析。

3.3 普赖斯公式选取法的验证

根据对普赖斯公式选取法的论述，可知普赖斯公式确定高频词主要取决于区间关键词出现的频次。由表3可见，个人知识管理领域文献的关键词最高频次为74次。因此，根据普赖斯公式计算得出的高频词阈值6.4。即频次大于或等于6的词均为个人知识管理领域的高频词，共计103个。同样用SPSS得出这103个高频词的聚类分析结果。

将此聚类结果大致分为三类：“显性知识”“隐性知识”与“图书馆”为第一类；“知识获取”“知识利用”与“知识管理工具”等为第二类；“学习型组织”“个人学习”与“组织学习”等为第三类。对比前30个词的聚类结果，虽然二八定律取值后的聚类划分结果也是三类，但两种方式的聚类结果差别较大。以“个人知识”“组织知识”“知识经济时代”“个人学习”与

“组织学习”这5个词为例来说明，在频次大于或等于15（Top 30）的词为领域高频词时，这5个词是被划分为同一类；而在频次大于或等于6（Top 103）的词为领域高频词时，这5个词则被划分到两大类中，即“个人知识”“组织知识”“知识经济时代”与“知识获取”“知识利用”等归为第二类，而“个人学习”“组织学习”则与“学习型组织”“企业”“知识管理能力”等归为第三类，即相同的5个词在聚类类别上发生了明显变化。由此可以说明，即使使用同一组数据，由于截频方法或截取阈值不同，也会产生明显不同的聚类分析结果，从而导致后续分析结论发生偏差。

据此可以推测，在面对不同高频词截取结果时，同样是以高频词分析为基础的多维尺度图、节点网络图等多种分析方法的呈现结果不一样，而目前我国学者对于领域发展、热点分析、趋势预测等的把握基本上是由聚类分析图、多维尺度图、节点网络图等综合得出，即不同的呈现结果必然会导致研究者分析结果的差异，由此可以证明不同的高频词选取方法导致不同的截取结果，确实会对后续的分析结果产生不同影响。

3.4 三种方法验证结果对比

从上述验证结果可以看出，即使是在同一研究领域，由于高频词阈值选择的方法不同，所确定的高频词也是明显不同，甚至差异巨大。在个人知识管理领域中，用二八定律确定的高频词阈值为15，包含前30个高频词；用高低频词界定公式方法确定的阈值为60，包含前3个高频词；用普赖斯公式方法确定的阈值为6，包含前103个高频词。三种方法确定的高频词数量最高相差100，相比之下，选用二八定律截取出的高频词阈值更合理。

目前我国学界在应用普赖斯公式计算高频词阈值时，绝大多数存在适用性问题。即将普赖斯公式计算得出的M值（实际代表高被引文献的阈值）用做高频词阈值。为说明普赖斯公式在高频词阈值界定上的不合理应用，及其对聚类分析结果的影响，本文特将这种情况列出，并与前文中所取阈值较合理的二八定律方法（阈值15）的聚类结果进行对比分析。

通过检索个人知识管理领域文献得到的最高被引次数为430，普赖斯公式计算结果约为16。以16作为高频词阈值，则可获取个人知识管理领域的前27个关键词高频词。通过对前27个词的聚类，分析发现个人知识管理领域的关键词可以聚为四类，明显不同于前30个词的三类划分，并且同样出现了同一关键词被划分在不同聚类的情况，如“博客”在前30个词的聚类中应划归第三类，而在前27个词的聚类中应划归在第二类，与“教师”“大学生”“知识管理系统”和“知识创新”等词划成一类。由此可见，即使截取的高频词阈值差异很小，对于聚类分析结果的影响也是很大的，仍然会对研究者的分析结论造成较大的影响，进而影响其对当前领域发展的认识和对未来发展趋势的判断。

通过上述实证研究，再次证实不同高频词选取方法对结果的巨大影响。在高频词取值差异的影响下，后续分析研究的可信度和科学价值值得商榷。因此，若想保证后续分析中的研究价值和意义，规范、科学地确定领域高频词是一个必不可少且十分重要的前提条件。但仅从目前我国学界的研究现状来看，绝大多数研究者尚未意识到高频词阈值的选取会对后续分析结论带来严重影响，更没有意识到现有高频词选取方式中的不足，以及其对分析研究的不利影响。

3.5 验证研究的结论

从三种选取方法的结果对比来看，用二八定律方法来确定领域高频词阈值是较合理的。一是以二八定律代替完全凭借研究者主观意愿的自定义选取方式相对客观；二是二八定律作为较成熟的理论，已成功应用在图书情报领域的诸多研究主题中，将其应用于高频词界定是有其理论依据的，以20%累计词频覆盖率作为该领域的核心关键词是合理的；三是从验证结果来看，二八定律所选取的高频词阈值较合理，与高低频词界定公式取词偏少、普赖斯公式取词偏多的情况相比，二八定律截取的高频词数量更适中，更符合研究者的需求；四是二八定律取词法在操作上更简便，即使对高低频词界定公式、普赖斯公式等方法运用不够熟练，也可以按照此定律取得合适的结果；五是二八定律是一个定值，既不需要经过公式计算，也不需要考虑公式中由于自变量取值不准确而对阈值计算造成的影响。因此，相较于其他的高频词界定方法，二八定律更适用。

4 高频词阈值选取方面存在的问题

4.1 主观性强

目前，学界进行的大部分词频分析或以词频分析

为基础的研究，对于高频词的截取数量和选取方法没有明确概念；且大多以作者关键词作为选词标准，而作者关键词本身就是文献作者的主观性选取结果；又因高频词的截取也是研究者的自主选择，不同研究者有不同态度，最终可能会导致结果不同，整个研究的主观性过强。

由于一些研究的领域较新，已有文献数据量不大，导致用公式计算得出的结果不准确，阈值界定范围过小，无法进行下一步分析。如张丛昱等虽使用普赖斯公式，但其实际计算结果与预期结果存在较大差异，最终只能根据作者对领域的理解，并结合公式计算数据确定高频词阈值的范围[16]，这也是混合选取法出现的根本原因之一。

4.2 方法原理不明

目前我国学者对某一领域的现状、趋势、热点的研究较多，但大部分文献在进行分析前，对如何准确地选择合适的高频词，以及高频词阈值选择标准等问题并未给出明确说明。大部分研究者只是在更换研究领域后，机械性照搬前人文献和写作模式。如依靠普赖斯公式计算得出词频大于6的词为领域高频词，但是对普赖斯公式的原理、优缺点、所取阈值是否合理等问题并未详加考量。

4.3 改进方法适用性不明

目前，高频词阈值的选取方法并没有形成统一概念，因此有人对当前高频词阈值选取方法提出异议并给予相应改进方法。即使有学者提出改进此问题的方法，且在某一领域内检测其适用性，但这种新方法也可能存在问题。巴志超等认为，文献中构建的语义网络度分布并不符合幂律分布特性，但没有过多讨论是否由于模型的参数设置、Top N的关键词选择以及语义度量指标等因素的影响和相互关系，也未检验构建的网络结构中的其他分布，如节点权值分布、中间中心性分布等是否符合幂律分布特性[8]。也就是说这一类文献虽然对提出的问题进行改善，但也可能只对提出的问题进行改善，并不排除可能会并发其他影响。而这些新方法本身也具有局限性，是研究者对词频截取中出现的某一问题或某几个问题做出的改进，而研究者对新方法的验证也仅是采用了某一领域的某一组数据。因此这种方法是否真正适用于该领域或其他领域，以及使用这种方法是否会产生其他并发性问题还需要进一步讨论。

4.4 高低频词界定公式存在适用性问题

从已有研究的情况来看，高低频词界定公式的取值偏大，导致截取到的高频词过少。造成这种情况的原因有两个：一是研究领域的相关主题本身比较分散，因此关键词重复率不高，仅出现1次的关键词数量较多。二是我国期刊文献的关键词多为作者关键词，即文献作者自定义的关键词，这种作者关键词的规范性不足，对同一事物可能存在多种不同说法。因此，大量不规范的作者自定义关键词就成为仅出现1次的关键词的主要组成部分，从而导致高低频词界定公式取值结果偏大，无法为领域高频词的确定提供合理参考。

4.5 普赖斯公式适用性不明

目前我国大多数学者将普赖斯公式的计算结果直接作为确定高频词的方法，这种做法虽然简单易行，在实际科学研究中有其独特优势和实用性。但此公式毕竟是为确定高被引文献而设计的，将其直接应用于领域高频词提取，实际上是不适合的。公式中自变量N_{max}表示区间学术论文被引频次最高值，即被引次数的最高值，计算得出的M值应该是“被引量”（即高被引文献的阈值）而不应是“关键词频次”（高频词阈值）。因此，将普赖斯公式直接用于确定高频词阈值值得商榷，目前仅有少数研究者意识到该问题。如胡利勇虽然在界定高频词时借用普赖斯公式[17]，但同时对该公式究竟是否适用于界定高频词提出质疑。

5 关于高频词阈值界定方法改进的思考

5.1 普赖斯公式法的改进

除上文中提到的将现有普赖斯公式中的自变量变为“最高关键词频次”来增加其应用于高频词界定的合理性之外，也可以在普赖斯公式确定领域核心文献的基础上，尝试将这些核心文献中的关键词作为领域核心关键词。普通计算关键词词频的方法是单纯将关键词累加，并没有考虑到核心文献中的关键词应该具有更大的影响。如高影响力作者的一篇高被引文献中的关

键词与普通文献的关键词权重完全相同。为显示出高被引文献的影响力，可以将被引次数作为权重参数加入到关键词词频的计算中。被引次数越多，经过加权后的关键词累加值也就越高，其相对应的核心关键词的频次就越高，这种方法的优点是可以突出核心文献对所在领域的影响。现有高频词取值方法是将所有文献中的关键词无差别计数，即无视核心文献的被引量和重要程度，与其他影响力一般的普通文献采用同样的关键词频次计数方式，这对于领域热点问题的分析和未来研究趋势的把握是非常不利的。被引次数高的核心文献对于同一领域热点研究趋势的影响明显高于普通文献，因此在进行领域热点或趋势分析时，应考虑核心文献的领域影响力并在研究方法中体现出来。现有领域高频词提取方法均未将该问题考虑在内。

5.2 高低频词界定公式法的改进

上文研究可进一步发现，目前造成高低频词界定公式界定高频词不理想的原因是领域内关键词分布较分散，虽然在具体计算前都有数据清洗流程，但这种清洗只能达到降噪的效果，无法解决关键词分散现象，分散现象的直观表现就是存在大量仅出现1次的关键词。使用普通清洗方法无法降低仅出现1次的关键词数量，所以只能借助其他方法来处理，从而降低人为标注关键词而产生的不规范行为对高低频词界定公式取值结果的影响。如当文献数量与关键词数量为1∶1.5时，定义仅出现1次的关键词在全部关键词集中的比例为x；当文献数量与关键词数量为1∶2时，定义仅出现1次的关键词在全部关键词集中的比例为y；在不同的文献与关键词数量比例区间下，仅出现1次的关键词数量在全部关键词数量中的占比应是不一样的。将此经过处理后的仅出现1次关键词数代入高低频词界定公式，这样可以在一定程度上避免高低频词界定公式计算结果过大而截取到的高频词过少或取不到高频词的情况。对于文献数量与关键词数量比例区间的划分方法，各区间仅出现1次的关键词所占比例等具体量值的确定，以及如何区分由于研究主题分散和作者关键词不规范这两种情况导致的关键词集分散等问题，尚有待进一步研究。

6 结语

高频词的阈值选取是词频分析的重要基础，而我国学界对于词频的阈值选取存在严重的不规范现象。在总结目前常用的三种高频词界定方式之后，引入个人知识管理领域样本进行实证检验，说明高频词截取的不同结果对后续分析的影响，总结出二八定律方法更适用于截取领域高频词。同时指出目前我国高频词界定方面存在主观性强、方法原理不明、改进方法适用性不明等问题。针对我国目前常用的高频词界定方法的不足，提出关于高频词界定方法的改良设想；但改良后的具体数值、应用条件等一系列问题未能明确，期待后续研究能够解决这些问题。总体来说，在高频词界定领域存在一种重实践轻理论的现象：依靠选取高频词进行的分析研究众多，但多数只是机械地仿照前人关于领域热点的研究模式进行，而对于高频词界定方法本身进行研究的论文并不多。总之，高频词界定方法中还存在许多问题，未来需要学者继续关注此问题，更加注重高频词界定方法的内在理论研究并提出有效且权威的界定方法，以减轻这种方法的乱用现象。

参考文献

[1] 马费成,张勤.国内外知识管理研究热点——基于词频的统计分析[J].情报学报,2006,25(2):163-171.

[2] 张勤.词频分析法在学科发展动态研究中的应用综述[J].图书情报知识,2011(2):95-98.

[3] 杨建林.关键词选择策略及其对共词分析的影响[J].情报学报,2014,33(10):1083-1090.

[4] 陈果,肖璐,赵雪芹.领域知识分析中的关键词选择方法研究——一种以学科为背景的全局视角[J].情报学报,2014,33(9):959-968.

[5] 安兴茹.基于正态分布的词频分析法高频词阈值研究[J].情报杂志,2014(10):129-136.

[6] 孙清兰.高频词与低频词的界分及词频估算法[J].中国图书馆学报,1992(2):78-81.

[7] 熊回香,杨雪萍.社会化标注系统中的个性化信息推荐研究[J]. 情报学报,2016,35(5):549-560.

[8] 巴志超,李纲,朱世伟.共现分析中的关键词选择与语义度量方法研究[J].情报学报,2016,35(2):197-207.

[9] 姚小娇.我国财经类高校近10年图书情报学研究热点分析[J].图书馆学刊,2015(2):137-140.

[10] 王佑镁,陈慧斌.近十年我国电子书包研究热点与发展趋势——基于共词矩阵的知识图谱分析[J].中国电化教育,2014(5):4-10.

[11] 李迎迎.国内“互联网+”领域研究热点及内容分析[J].情报杂志,2016(8):128-132.

[12] 赵蓉英, 魏明坤. 2010——2015年国内外情报学研究热点可视化比较[J].图书馆杂志,2016,35(8):15-22.

[13] 朱莎,杨浩,冯琳.国际“数字鸿沟”研究的现状、热点及前沿分析——兼论对教育信息化及教育均衡发展的启示[J].远程教育杂志,2017,35(1):82-93.

[14] 王米雪,张立国.我国智慧教育领域的研究热点与发展趋势分析——基于词频分析法、共词聚类法和多维尺度分析法[J].现代教育技术,2017,27(3):41-48.

[15] 崔雷,刘伟,闫雷,等.文献数据库中书目信息共现挖掘系统的开发[J].现代图书情报技术,2008(8):70-75.

[16] 张丛昱,张云中.国内近十年Folksonomy领域研究热点与趋势[J].新世纪图书馆,2016(7):91-96.

[17]胡利勇.图书情报领域高被引论文共词聚类分析[J].图书馆学刊,2016(8):132-135.

作者简介

刘奕杉，女，1992年生，硕士研究生，研究方向：数字信息资源管理，E-mail：2387161672@qq.com。
王玉琳，女，1994年生，硕士研究生，研究方向：数字信息资源管理。
李明鑫，男，1978年生，博士，讲师，研究方向：数字信息资源管理。

An Empirical Analysis for the Applicability of the Methods of Definition of High-Frequency Words in Word Frequency Analysis

LIU YiShan, WANG YuLin, LI MingXin
(School of Information Science and Technology, Northeast Normal University, Changchun 130117, China)

Abstract: Word frequency analysis method is one of the important analysis methods in bibliometrics, and the selection of high-frequency word is a necessary premise. It is to say that the selection of high-frequency word determines the results of word frequency analysis, impacts the whole analysis program. First, the paper cleared up the nearly three years papers in China by using word frequency analysis method for hot spots analysis, and found four common classes selections of high-frequency word methods mainly include: the author set the selection method, Donohue's formula selection, price formula selection and mixed selection. Secondly, we use the literature of personal knowledge management as the research object, and calculate the frond three kinds of high frequency words selections respectively, and compare the results with clustering analysis, then we discuss the effect and applicability of high-frequency word threshold selection on the analysis results. At last, the paper pointed out that there were some problems, such as the subjective is high, principle is unclear, improved methods' principle is unclear, the Donohue's formula and price formula's applicability are still unsure, in the study of high-frequency word threshold selection in our country.

Keywords: High-Frequency Word; Bibliometrics; Word Frequency Analysis

（收稿日期：2017-08-07）

书 讯

《中国高被引分析报告2016》

《中国高被引分析报告2016》按理、工、农、医、人文、社科等领域划分为50个学科，综合分析了各个学科的高影响力论文、研究热点与前沿、高影响力期刊、高影响力作者和高影响力科研机构，并以关联图谱的方式展现了多种学术关系，有助于科研人员及时发现并跟踪研究热点，有利于期刊编辑部监测本刊学术影响力，有利于科研管理机构评估科研能力，是高等院校、科研院所及期刊编辑部等相关单位和人员的参考工具书。

该书以“中国知识链接数据库”为依托，数据覆盖我国6 000余种期刊的论文及引文。书中分学科揭示了高影响力的学者、研究机构（大学、研究所、医院等）、地区（省/自治区/直辖市）、学术期刊、图书、外文期刊和会议录，并采用共词分析、共被引分析和合著分析等方法绘制出各学科的前沿主题分布以及作者、机构和期刊间关联的知识图谱。

《中国高被引分析报告2016》由中国科学技术信息研究所编制，曾建勋主编，科学技术文献出版社出版，全书约80万字，定价298.00元。