情报学报　2018 年 5 月　第 37 卷　第 5 期
Journal of the China Society for Scientific and Technical Information, May 2018, 37(5): 533-542

DOI: 10.3772/j.issn.1000-0135.2018.05.011

基于关键词重要性和近邻传播聚类的主题分析研究

李海林[1,2]，万校基[1]，林春培[1]

（1. 华侨大学工商管理学院，泉州　362021；2. 华侨大学现代应用统计与大数据研究中心，厦门　361021）

摘　要　鉴于传统科学计量方法存在共现分析缺少考虑关键词重要性和主题分析手段不能自适应地抽取核心主题等问题，本文提出一种基于关键词重要性和近邻传播聚类的主题分析方法。该方法依据大多数作者的潜在行为会按照与研究内容相关性的强弱顺序提供论文关键词，计算关键词在每个文献中的重要程度，构建主要关键词之间的相似性矩阵，结合能够反馈最优簇成员代表性结果的近邻传播聚类实现核心主题的提取与分析。本研究对图书情报类某刊物 2012-2016 年期间的文献关键词进行数据挖掘，使用新方法实现了基于重要性度量的主要关键词聚类，分析和研究了主要关键词和核心主题的演化趋势。提出的方法不仅能够考虑关键词重要性和自动识别核心主题，还可以为文献主题分析提供新的数据挖掘方法，也能有效提高期刊和学科等相关领域的主题识别效果。

关键词　主题分析；关键词重要性；近邻传播聚类；核心主题

Theme Analysis Based on Keyword Importance and Affinity Propagation Clustering

Li Hailin[1,2], Wan Xiaoji[1] and Lin Chunpei[1]

(1. *College of Business Administration, Huaqiao University, Quanzhou*　362021; 2. *Research Center of Applied Statistics and Big Data, Huaqiao University, Xiamen*　361021)

Abstract: In view of the fact that co-occurrence analysis lacks consideration of keyword importance and theme analysis in such a way that it does not adaptively extract the core themes in traditional scientific measurement methods, this paper proposes a theme analysis method based on keyword importance and affinity propagation clustering. Based on probable behavior of most authors, the method collects the keywords of theses according to the strength or weakness of the relevance to the corresponding research content, computes the importance measure of the keywords in the papers, and constructs the similarity matrix of the keywords. The extraction and analysis of the core theme is achieved through combining the method with affinity propagation clustering that can retrieve the best representative member of the cluster. In this study, the keywords in a specialized journal of literature and information during the period of 2012 to 2016 were collected, and keyword clustering based on importance measurement was implemented. The evolutionary trends of keywords and core themes were analyzed and studied. The method proposed in this study not only considers the keyword importance and automatically identifies core themes, but also provides new data

收稿日期：2017-12-08；**修回日期**：2018-01-21

基金项目：国家自然科学基金项目“高维时间序列数据聚类分析及应用研究”（71771094）；福建省社会科学规划项目“基于时间序列数据挖掘的期刊参考文献和引证文献分析研究”（FJ2017B065）。

作者简介：李海林，男，1982 年生，博士，副教授，主要研究方向为数据挖掘和文献分析，E-mail: hailin@mail.dlut.edu.cn；万校基，男，1984 年生，博士，讲师，主要研究方向为文献分析和网络借贷；林春培，男，1985 年生，副教授，主要研究方向为创新管理与情报分析。

mining methods for thematic document analysis and effectively improves the topic recognition effect in related fields such as journals and other disciplines.

Key words: theme analysis; keyword importance; affinity propagation clustering; core theme

1 引 言

主题分析是图书情报、管理与创新等学术研究领域中备受关注的热点问题[1-3]，其目的是从相关文件、报道和文献等数据资料中获取某个时期内人们所关心的主要话题和研究内容，以便更好地掌握时代特色和发展趋势对当前社会、经济、金融以及自然科学等方面的影响。同时，主题演化也是图书情报领域中的一个重要研究方向[4-5]，它以时间顺序为主线，分析不同时间点所关心的重要主题，以便人们更好地发现相关学科和管理科学等领域中的重要研究内容及研究线路。主题演化分析可以科学地归纳历史研究主题中存在缺失、薄弱和突出等环节，能够为相关领域研究问题提供更好的科学管理和决策支持。

对于文献主题分析的大部分研究是采用科学计量方法，如词频[6]、共引[7]和共词[8]等分析方法。词频分析方法是主要是用来统计文献资料中出现关键词的频率，依据频率大小来划分关键词的重要性，通常采用词云和统计直方图等可视化技术来实现关键主题的分析。共引分析则是通过文献之间引用和被引用关系的分析，并通过聚类和多维尺度分析等方法对具有相似主题进行归纳总结。共词分析是文献计量中最常用的方法，主要通过对关键词共同出现在同一篇文献中的频次度量不同关键词之间的相似关系，使用 Ochiai 系数、Jaccard 系数、余弦指数和 TF/IDF 等相似性度量方法来构建重要关键词之间的相似性矩阵，进而采用多维尺度分析和层次聚类等多元统计手段进行主题划分和提取[9]。与此同时，随着大数据时代下文本挖掘的重要性，基于语义和人工智能的主题分析[10-11]也成为该研究领域的重要方法，这些方法通常依赖于精确可靠的语料库和机器学习的性能。

相比较于语义分析和人工智能等方法来说，科学计量方法是一种原理简单直观且易于被研究者接受的方法。鉴于词频分析不能够深入分析关键词之间的相关性以及共引分析需要大量共引文献的复杂性，共词分析成为文献主题分析研究中最重要且最常用的科学计量方法。然而，目前共词分析方法大多数是建立在关键词频率的基础之上，缺乏关注关键词对不同文献存在不同重要性影响的研究。虽然部分学者[12-13]通常使用词频来度量这些词语的重要程度，但对于词语是否能成为反映文献主题的关键词是一个有待深入分析的问题。虽然文献作者提供的关键词能够较好地反映相关文献的研究主题，但传统共词分析将所有关键词的重要性视为同等重要，使得最终的分析结果不能很好地反映文献的真实主题。与此同时，共词分析得到的关键词相似性矩阵结合层次聚类、多维尺度分析和社会网络分析图谱等分析方法能够对研究问题进行主题提取，但聚类个数和尺度大小等人为设定因素势必影响了相关主题的分析质量。因此，如何在研究过程中增强关键词重要性分析和减少人为主观因素的影响成为主题分析领域中值得关注和深入研究的问题。

鉴于不同关键词对同一篇文献主题的不同反映程度和关键词反映核心主题的客观性，本文提出一种基于关键词重要性和近邻传播聚类的主题分析方法。通过对同一篇文献的关键词进行重要度计算，以此为基础建立主要关键词的共现相似性矩阵，结合自适应近邻传播聚类方法对相似性矩阵进行自动聚类，使用每个聚类簇中心代表关键词来描述该簇的核心主题，进而实现相关文献资料的主题提取和分析。另外，以图书情报类某一核心期刊 2012 年至 2016 年期间刊发的文献资料作为数据资源，使用新方法对该期刊的主题进行分析研究，发现该期刊在这段时间内的主题演化过程及规律，为相关期刊发展和学科主题的研究提供决策支持和参考建议。

2 研究方法

2.1 研究出发点

基于关键词的传统主题分析方法是将文献数据中提取的关键词通过共词分析来建立关键词之间的相似性矩阵，再结合经典的多维尺度分析或层次聚类等数据分析手段进行聚类，并且根据指定的类别数目或多维尺度进行主题提取和分析。本文提出基于关键词重要性和近邻传播聚类的主题分析方法，通过对关键词的重要性进行计算，使用关键词的权重和共现关系来计算它们之间的相似性，进而得到基于关键词重要程度的相似性矩阵。

鉴于传统聚类分析手段不能很好地处理簇类数

目的问题，使用近邻传播方法（Affinity Propagation，AP）[14]在相似性数据矩阵的基础上进行聚类分析，并且根据簇中心代表关键词具有较好的簇成员代表性，使得聚类结果中的簇中心代表对象可以直接作为相应关键词簇的核心主题。如图 1 所示，实线与虚线流程分别表示新方法与传统方法的主题分析思路。新方法不仅考虑了关键词的重要性，还结合近邻传播 AP 聚类方法解决核心主题抽取的问题。

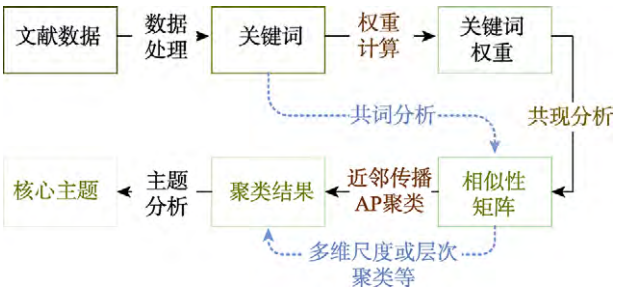

图 1　基于关键词和 AP 聚类的主题分析研究思路

为了更好地了解本文提出方法的思路和过程，本节先对关键词的重要性进行了定义和权重计算，再结合共词分析确定关键词之间的相似性，进而构造出主要关键词之间的相似性矩阵。另外，为了更好地理解新方法聚类结果与传统聚类结果的差异性，简单介绍了近邻传播聚类的基本思想对主题分析中的作用和意义。

2.2　关键词重要性

通常情况下，每篇文献数据都具有多个关键词来共同描述同一个文献主题的研究，但这几个关键词对文献主题描述的贡献程度或重要程度不相同。例如，查先进等[15]提出题目为“网络信息行为研究现状及发展动态述评”一文中，作者按先后顺序给出描述该文献主题研究的 3 个关键词，即“信息行为”、“网络环境”和“信息需求”。不难发现，针对“网络信息行为研究现状及发展动态述评”这一文献主题，关键词“信息行为”最能够描述该文献的研究内容和研究对象，是最重要的关键词。其次，“网络环境”是该文献研究内容和研究对象的定语，其重要性要低于研究对象“信息行为”的重要性。“信息需要”关键词则是该文献主题中经常涉及的词主，通常与信息行为相关，其重要性要比“信息行为”和“网络环境”低。实际上，文献作者在对关键词提炼时通常会考虑与文献研究主题最为关联的若干关键词，并且潜意识地根据文献研究主题的相关性和重要性来罗列关键词。因此，关键词的重要程度在某种意义上与人的行为决策相关，其重要性的强弱在一定程度上与作者给出关键词的顺序相对应。

根据上述分析，若某文献 p 中作者使用了 K_p 个关键词来描述该文献主题，且按顺序先后形成关键词集合 $\mathrm{Key}_{p\bullet}=\{\mathrm{Key}_{p1},\mathrm{Key}_{p2},\cdots,\mathrm{Key}_{pK_p}\}$，则在文献 p 中第 k 个关键词对文献主题的相关性描述的重要度可定义为：

$$w\mathrm{Key}_{pk}=\frac{K_p-k+1}{\sum_{k=1}^{K_p}k} \tag{1}$$

式(1)根据作者给出关键词的先后顺序来计算它们之间的权重。例如，在文献数据集中，若第 p 篇文献为“网络信息行为研究现状及发展动态述评”一文，则 $K_p=3$ 且第一个关键词 key_{p1} 为“信息信为”，第二个关键词 key_{p2} 为“信息环境”和第三个关键词 key_{p3} 为“信息需求”，它们的权重可计算得到且分别为 1/2、1/3 和 1/6。

2.3　关键词相似性

传统方法通常根据不同关键词在文献数据集中共同出现的频次来衡量它们之间的相关性或相似性，最常用的方法为 Ochiai 系数，其计算公式为

$$\mathrm{Ochiai}(i,j)=\frac{C_{ij}}{\sqrt{C_i}\sqrt{C_j}} \tag{2}$$

其中，C_{ij} 表示关键词 i 和关键词 j 在文献集合中共同现出的频次，C_i 和 C_j 分别表示关键词 i 和 j 在文献集合各自出现的频次。容易发现，Ochiai 系数仅考虑了关键词共同出现的频次关系，忽略考虑关键词在不同文献的重要程度。为了使得度量的相关性系数或相似性度量能够体现相同关键词对不同文献和不同关键词对同一文献的重要程度差异性，提出基于重要性的关键词相似性度量方法，即

$$\mathrm{Sw}(i,j)=\frac{\sum_{i,j\in\mathrm{Key}_{p\bullet}}^{N}w\mathrm{Key}_{pi'}\times w\mathrm{Key}_{pj'}}{\sqrt{\sum_{i\in\mathrm{Key}_{p\bullet}}^{N}w\mathrm{Key}_{pi'}^2}\sqrt{\sum_{j\in\mathrm{Key}_{p\bullet}}^{N}w\mathrm{Key}_{pj'}^2}} \tag{3}$$

其中，$w\mathrm{Key}_{pi'}$ 和 $w\mathrm{Key}_{pj'}$ 分别表示关键词 i 与关键词 j 共同出现在第 p 篇文献关键词集合 $\mathrm{Key}_{p\bullet}$ 中对应第 i' 个关键词和第 j' 个关键词的权重，$i,j\in\mathrm{Key}_{p\bullet}$ 表示两个关键词 i 和 j 共同出现在第 p 篇文献关键词集合 $\mathrm{Key}_{p\bullet}$ 中，N 为文献集合中的文献数量。式(3)中关键词之间的相似性度量取值范围为[0,1]，值越

大表示关键词之间的相似性越强。根据式(3)可以进一步计算文献数据集合中 m 个重要关键词之间的相似性矩阵，即

$$S=\begin{bmatrix} \mathrm{Sw}(1,1) & \mathrm{Sw}(1,2) & \cdots & \mathrm{Sw}(1,m) \\ \mathrm{Sw}(2,1) & \mathrm{Sw}(2,2) & \cdots & \mathrm{Sw}(2,m) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{Sw}(m,1) & \mathrm{Sw}(m,2) & \cdots & \mathrm{Sw}(m,m) \end{bmatrix} \quad (4)$$

2.4 近邻传播聚类

近邻传播 AP 是一种基于数据对象相似性矩阵分析的聚类方法，通过传递和更新近邻信息使得最终每个数据点所包含的信息达到平衡。该方法的基本思想于 2007 年由 Frey 等[14]提出并发表在 Science 权威刊物，后来也有不少学者[16-17]对其算法的性能和效率进行了改进，并且常被应用于各种领域的数据聚类分析。AP 聚类具有的特点包括初始将每个数据点作为初始代表点，不需要设置初始中心点；使用客观存在的数据点作为簇中心代表对象，而不是如 *K*-means 等算法形成新的簇中心；无需初始化用户指定最终聚类数目；算法不易受数据的初始值影响；对数据之间的相似性矩阵没有对称性要求；能够很好地解决非欧空间和大规模稀疏矩阵计算等问题。特别地，通过 AP 聚类算法获得的簇中心代表对象能够较好地反映对应簇中所有成员的公共信息，使用簇中心代表对象能够较好地描述簇的特征，进而可以用来反映同类数据的主题。

与传统划分聚类相比，AP 聚类能够根据数据之间的相似性，自动找出若干个簇中心代表对象，并且每个簇成员数据对象与该簇中心代表对象的相似性最大。在 AP 聚类运算过程中，通过对两种信息的迭代和更新来完成，即代表程度 $r(i,k)$ 和合适程度 $a(i,k)$。前者表示 x_k 的积累信息，说明了 x_k 作为 x_i 代表点的程度；后者代表了 x_i 的积累信息，用来表示 x_i 选择 x_k 作为代表点的合适程度。它们的计算公式和迭代过程可表示成：

$$\begin{cases} R^{(t+1)}=f(S,A^{(t)}) \\ A^{(t+1)}=g(R^{(t+1)}) \end{cases} \quad (5)$$

其中，S 表示数据对象之间的相似性矩阵，R 和 A 分别为数据对象在 AP 聚类迭代过程中的代表程度矩阵和合适程度矩阵，t 表示迭代次数。f 和 g 分别为信息更新函数，具体公式可参阅文献[14]。另外，在首次迭代 $t=0$ 时，合适程度矩阵 A 初始化为零矩阵。最后，通过计算数据点的代表程度和合适程度之和，选择具有最大信息程度的点 $x_{k'}$ 作为 x_i 的代表点，即

$$k'=\arg\max_k(r(i,k)+a(i,k)) \quad (6)$$

如图 2 所示，二维平面坐标系中的 20 个数据点通过计算它们之间的相似性，利用 AP 聚类方法对相似性矩阵进行聚类，在每次迭代过程中 AP 聚类能够使每个数据点找到邻近最具代表性的数据点作为同簇中心代表点。被指向的数据为代表点，箭头出发点为被代表点。通过这种方式，AP 聚类最终可以将数据点自动划分成若干个簇，每个簇中具有一个簇中心代表点，也就是簇中其他所有成员指向该代表点，同时也说明了该代表点在本簇中具有最强的代表性。

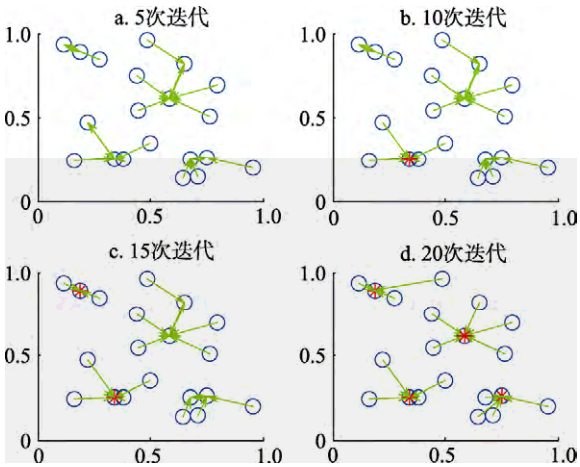

图 2 近邻传播 AP 聚类迭代过程

鉴于 AP 聚类结果中每个簇的中心代表点是该簇中最具有代表性的数据对象，故将其应用于文献数据核心主题的提取研究。通过建立关键词之间的相似性矩阵 S，使用 AP 聚类便可获得若干个簇中的中心代表关键词，这些关键词被称为代表相应簇其他关键词所反映主要研究内容的核心主题。与此同时，簇中成员之间具有较强的相似性，说明同簇中的关键词很大程度在描述同一主题，而且该主题由中心代表关键词来描述是最为合适的。因此，通过 AP 聚类并结合关键词重要性的相似性矩阵，便可实现基于关键词重要性分析的主题研究，即 idx = $AP(S)$，其中 idx 为 AP 聚类的结果向量，记录了每个关键词的中心代表关键词。例如，idx=[2,2,5,2,5] 表示关键词聚类结果中第 1、2 和 4 个关键词为同一类和第 3 个和第 5 个关键词为一类，这两类的核心主题分别由第 2 个和第 5 个关键词来描述。

3 主题分析

基于关键词重要性和近邻传播聚类的主题分析研究不仅可以用来对某个学科或某个领域的热点主

题进行分析，还可以针对某个期刊近年来刊发的论文实现主题研究，进而发现和探寻目标学科和目标期刊近年来主要聚焦问题和主题研究的发展趋势等，有利于作者、读者和编辑部对研究热点方向的关注和把握，进而促进学术研究的发展。

3.1　数据来源

选取 2012—2016 年图书情报类某核心刊物发表的论文作为文献数据分析资料，从中国知网中下载该期间的文献数据并提取关键词和发表年份等信息，数据收集时间为 2017 年 10 月 11 日。数据文献统计信息如表 1 所示，共收集到 2818 篇文献，作者提供了 11700 个关键词，不计重复的关键词有 6314 个。按年份先后顺序列出了重复出现 x 和不重复的关键词 y，结合每年文献数量 c，容易发现单篇论文提供的关键词数量 x/c 约为 4.15，并且每年波动不大，说明大多数作者都会提供 4~5 个关键词来描述相应的研究内容。另外，从单篇不重复的关键词数量 y/c 来看，随着时间变化作者提供了越来越多的不重复关键词，说明该期刊的研究主题近年来呈现出越来越丰富的状态。

表 1　数据文献及关键词统计信息

年份	2012	2013	2014	2015	2016	总数
重复 x	3157	2853	2392	2121	1177	11700
不重复 y	2170	1978	1725	1540	867	6314
文献数量 c	761	683	581	511	282	2818
x/c	4.15	4.18	4.12	4.15	4.17	—
y/c	2.85	2.89	2.96	3.01	3.07	—

3.2　关键词重要性分析

为了验证“关键词的重要性强弱是否与作者给出关键词的顺序相对应”的假设，通过实验数据来说明关于关键词重要性计算前提的可行性。通常情况下，文献的标题和摘要能够较全面地对文献主题进行描述和说明，因此可以使用作者给出的关键词在文献标题和摘要中出现的概率（即呈现概率）来分析关键词出现次序对重要性计算的影响。本次实验中，使用了 2818 篇文献中用户给出的关键词集合、标题和摘要等数据信息。通过对每篇文章作者给出的前 8 个关键词分别在标题、摘要和标题或摘要等三种情况中呈现概率进行统计，用来说明不同顺序的关键词对相关文献主题的不同描述力。需要说明的是，若作者给出的关键词少于 8 个，则只需对作者给出关键词进行顺序的重要性分析。实验结果可以发现，如图 3 所示，第 1 顺序的关键词在这三种情况下呈现概率最高，说明第 1 顺序的关键词对文献的重要性最强。随着关键词顺序的推移，相应顺序的关键词在这三种情况下出现的概率逐步下降，进而验证了关键词的重要性强弱与作者给出的关键词顺序相关。即关键词的顺序越靠前，该关键词对文献主题的描述力就越强。

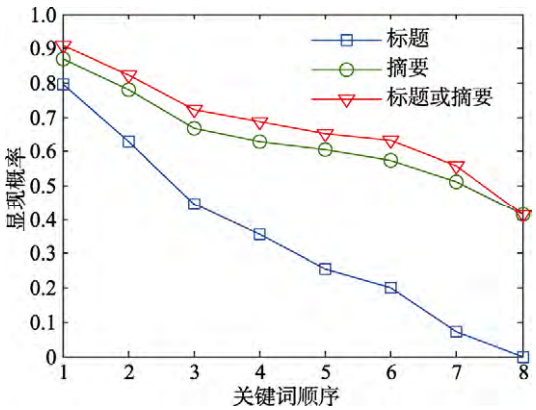

图 3　关键词顺序的重要性分析

根据式(1)，不重复关键词可以用来在计算这些关键词在每篇文献中的重要程度（权重），并将其在所有文献数据中出现的权重平方和视为这些关键词的重要性，即

$$w(i)=\sum_{i\in \mathrm{Key}_{p\bullet}}^{N} w\mathrm{Key}_{pi'}^{2} \tag{7}$$

其中，N 表示文献数量，$w\mathrm{Key}_{pi'}$ 表示第 i 个关键词在第 p 篇文献中对应相同关键词 i' 的权重，若不出现在该文献中，则对应的权重为 0。根据重要性的强弱来选择特定时期该刊物的重要关键词，如表 2 显示了 2012—2016 年期间以及每年前 10 个最重要的关键词，其重要性代表这些关键词在对应时间段里被该期刊作者和编辑所关注的重要程度。由表 2 的重要性数值可以看出，数值之间的差异性较大，说明这些关键词能对不同主题的进行区分描述，并且不同主题具有较大差别的被关注度。同时容易发现，对于此类具有较强特色的刊物，包括“高校图书馆”、“公共图书馆”、“数字图书馆”和“移动图书馆”等图书馆相关研究一直是该刊物研究的重要主题。除了关

表 2 2012—2016 年期刊文献前 10 个重要关键词

序号	2012 年		2013 年		2014 年	
	关键词	重要性	关键词	重要性	关键词	重要性
1	图书馆	7.71	图书馆	7.59	高校图书馆	6.76
2	高校图书馆	6.73	高校图书馆	6.93	图书馆	4.85
3	公共图书馆	4.04	公共图书馆	2.36	公共图书馆	3.19
4	数字图书馆	3.66	开放获取	2.08	数字图书馆	1.78
5	学科馆员	2.46	移动图书馆	1.68	移动图书馆	1.43
6	微博	2.38	关联数据	1.51	大数据	1.41
7	知识服务	1.65	学科服务	1.47	学科馆员	1.01
8	文本挖掘	1.48	数字图书馆	1.39	图书馆联盟	0.95
9	学科服务	1.47	图书馆学	0.92	突发事件	0.92
10	竞争情报	1.36	移动服务	0.92	移动阅读	0.84
序号	2015 年		2016 年		2012—2016 年	
	关键词	重要性	关键词	重要性	关键词	重要性
1	高校图书馆	6.27	高校图书馆	3.95	高校图书馆	30.65
2	图书馆	3.82	图书馆	2.03	图书馆	26.00
3	公共图书馆	2.30	阅读推广	0.84	公共图书馆	11.93
4	大数据	1.52	政府数据	0.70	数字图书馆	7.24
5	全民阅读	1.21	图书馆员	0.65	移动图书馆	5.27
6	MOOC	1.19	移动图书馆	0.64	学科服务	5.02
7	阅读推广	1.13	学科服务	0.63	学科馆员	4.62
8	网络舆情	1.09	大学图书馆	0.61	大数据	3.83
9	移动图书馆	1.08	大数据	0.58	开放获取	3.60
10	信息素养	1.07	信息素养	0.58	微博	3.33

于图书馆的相关研究外，一些其他主题的研究会随着时间的变化而改变，如“学科服务”、“移动服务”和“大数据”等。然而，表 2 给出的重要关键词信息不易发现重要主题的变化趋势。因此，对于关键词演变而导致主题变化的研究有待进一步深入分析。

3.3 关键词演化分析

期刊主题演化通常表现为相应关键词随时间变化的演化过程，因此，通过对关键词演化分析可以发现期刊主题随时间变化的规律和趋势。通过不重复关键词分别在 2012—2016 年每年该刊物刊发文献中出现的权重，统计每个关键词的重要性，并按照重要性的强弱进行排序，分别选择前 10 个主要关键词来研究本年度该期刊受关注的主题。同时，对重要性实现均篇重要性分析，即用表 2 中的重要性除以表 1 中对应当年的文献数量 $c(\text{year})$，即可得到均篇重要性 $\bar{w}(i)=w(i)/c(\text{year})$。如图 4 所示，纵坐标是 2012—2016 年每年的前 10 个主要关键词形成的 27 个不重复的关键词，说明了 5 年期间该期刊研究主题所对应的主要关键词具有一定的重叠，即重叠率为(50–27)/50=46%，也反映了某些关键词的反映的主题被受到持续关注。实际上，若图 4 中每行出现 2 次及上以圆圈，则说明纵坐标对应的关键词在不同时期内受到关注。另外，圆圈大小代表对应关键词出现在当年刊发文献中的均篇重要性，圆圈越大越能代表它被重视的程度越大。从图 4 中易知，“图书馆”、“高校图书馆”和“公共图书馆”等关键词被关注度最大，并且“高校图书馆”被关注度随着时间变化越来越大，而对于“图书馆”的综合研究相对越来越小。

除了对于关于图书馆的相关研究外，其他一些主题备受重视。例如，“大数据”自 2014 年开始就受到期刊作者和编辑的关注，这也符合时代发展的研究情况。实际上，自 2012 年大数据概念的正式提出，大数据研究的受到各个领域的关注。但论文创作、编辑审稿和和刊发过程需要一定的时间。研究发现，该期刊把“大数据”作为主要关键词却从 2014 年才开始，可以从侧面说明该刊物的作者和编辑对于论文创作效率和刊发时效有待进一步提高。同样，“移动图书馆”、“阅读推广”、“网络舆情”和“政府数据”都分别持续和逐渐受到有关学者关注，并且也符合“互

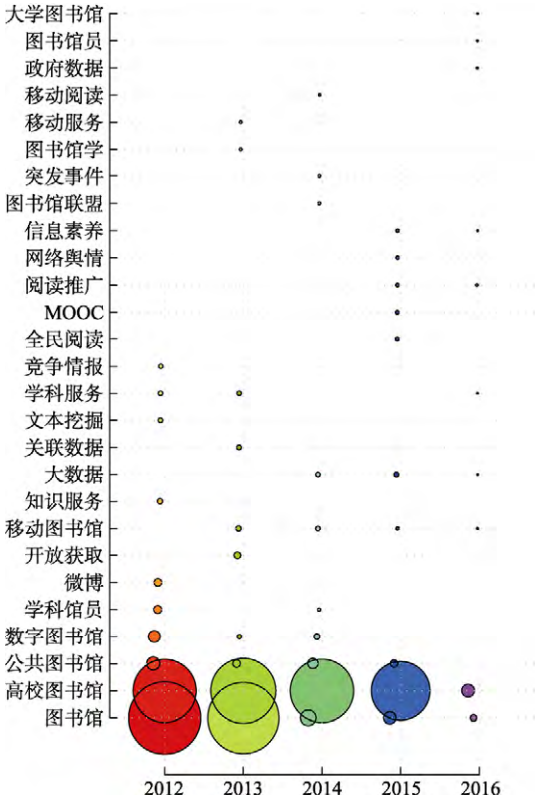

图 4　主要关键词的演化趋势

联网+”和移动网络等环境下的时代特征研究。然而，比较明显的是关于“微博”、“学科馆员”和“竞争情报”等相关主题的研究被逐步淡化，这也说明这些领域很可能近期一段时间内不被关注的可能性较大。因此，通过基于权重影响的关键词演化分析，可以用来归纳总结期刊主题演变趋势，也可用来预测期刊或学科领域未来研究的主体方向，以便更好地支持期刊学术发展和学科主题的良性研究。

3.4　核心主题演化分析

为了进一步确定哪些重要关键词具有共同的研究主题，使用基于重要性影响的关键词分析构建主要关键词之间的相似性矩阵。通过近邻传播 AP 对反映重要关键词相关性的相似性矩阵进行聚类分析，最终得到若干个簇，并且每个簇的中心代表关键词用来表示该簇其他成员共同描述的核心主题。

根据关键词的重要性大小排列，选取该刊物中每年刊发文献中的前 50 个主要关键词进行相似性度量，构建相似性矩阵后使用 AP 聚类的结果如图 5 所示。图 5a~图 5e 分别描述了 2012—2016 年每年主要关键词聚类结果呈现出来的核心主题，图 5f 是这五年内前 50 年重要关键词聚类结果呈现的核心主题。图 5 中每个簇都具有中心代表对象，即核心主题。例如，2012 年重要主题聚类结果中“学科馆员”核心主题包括了“学科服务”、“学科查新”、和“学科化服务”等相关研究，2013 年“移动图书馆”核心主题包括了“移动服务”和“移动阅读”等相关研究，2014 年“图书馆联盟”核心主题涵盖了“美国”和“战略规划”等相关研究，2015 年“大数据”核心主题包含了“科研数据”、“网络舆情”、“数据挖掘”、“健康信息”、“专利引文”和“智库”等相关研究，2016 年的核心主题“用户行为”却与“协同信息检索”、“社会化标签”和“网络百科”等推荐系统的研究有关。然而，综合从这五年内来看，高校及其图书馆是该刊物最为重要的核心主题，MOOC、美国、数据挖掘、本体和学科馆员也是该期刊近年来较为重要的核心主题。

从图 5 中的聚类结果可以分析得知，近邻传播 AP 聚类获得的中心代表关键词能够较好地发现核心主题，使其能很好地代表簇成员的其他关键词的主体思想。例如，2015 年重要关键词聚类结果中，“阅读推广”作为核心主题时，其成员包括了“农家书屋”、“读书会”、“信息素质教育”、“虚拟社区”、“高校图书馆”和“全民阅读”等，这些关键词成员在一定程度上都与阅读推广研究具有较强的相关性。与此同时，随着时间的变化，部分核心关键词也会演变成簇成员，而有些簇成员也会变成核心关键词。如 2014 年“大数据”关键词是核心关键词“数字图书馆”的簇成员，但随着大数据的发展，“大数据”关键词在 2015 年和 2016 年成为了核心关键词。同样，随着主题研究的不断深入，主题之间也不断存在交叉研究，例如，核心主题“高校图书馆”在每年都呈现出不同相关内容的研究。

另外，对每年聚类结果中获取的核心主题关键词进行演化分析，可以发现哪些核心主题持续受关注，哪些主题却不再受关注和关注程度如何等信息和知识。图 5 中分别对 2012—2016 年的 50 个主要关键词使用 AP 聚类，每个簇的中心代表对象为一个核心主题，进而可以分别提取了 14、13、16、8 和 10 个核心主题，共计 61 个核心主题。例如，图 5d 显示了 8 个核心主题，分别为“大数据”、“阅读推广”、“机构知识库”、“影响因素”、“资源融合”、“MOOC”、“图书馆”和“科学数据”。将每个核心主题所包含的所有成员关键词的权重加和作为该核心主题的重要性，再对它们进行演化统计分析，其核心主题演化结果如图 6 所示。通过演化趋势分析，可以得到 42

a. 2012年主要关键词聚类结果

b. 2013年主要关键词聚类结果

c. 2014年主要关键词聚类结果

d. 2015年主要关键词聚类结果

e. 2016年主要关键词聚类结果

f. 2012-2016年主要关键词聚类结果

图 5 基于主要关键词聚类的核心主题变化情况

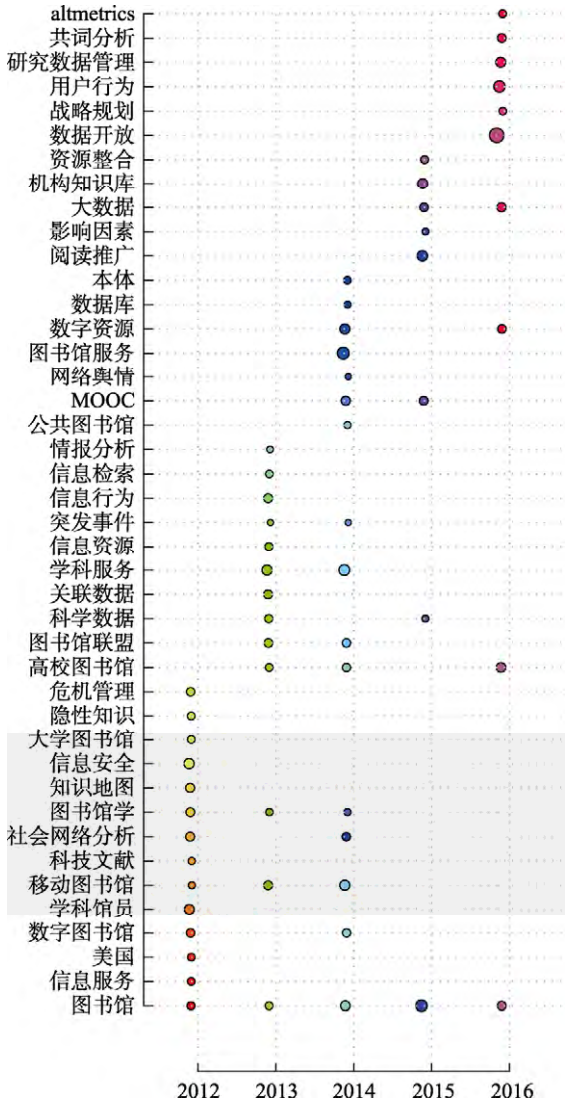

图 6　核心主题演化趋势

个不重复的核心主题，其圆圈大小代表该主题受关注的重要程度，圆圈半径越大代表该核心主题受到关注的程度越高。

容易发现，“图书馆”核心主题一直受到该期刊的青睐，在一定程度上也体现了该期刊以图书情报为研究主象的办刊宗旨。关于大数据及数据管理方向的主题研究也逐渐受到该期刊的重视，并且呈现出较大的研究热度。部分主题的研究显示出阶段性的研究热度，如“信息检索”、“信息资源”、“突发事件”和“学科服务”等核心主题出现一时的研究热度。然而，“移动图书馆”和“数字图书馆”等核心主题研究在近两年没有受到关注。因此，通过核心主题演化分析可以进一步掌握期刊或学科等相关领域中存在的主要研究内容和趋势，为期刊和科学的学术研究发展提供主题研究重要线索。

4　结　论

本文提出了一种基于关键词重要性和近邻传播聚类的主题分析方法，该方法从文献关键词具有重要程度差异性的角度出发，通过建立不同关键词在同一文献或不同文献的权重信息，筛选出重要关键词，并构建重要关键词之间的相似性矩阵。与此同时，通过利用能够发现簇中心代表点的近邻传播聚类方法对该相似性矩阵进行聚类分析，可以发现若干个不同研究核心主题的簇，每个簇的中心代表关键词能较好地反映该簇的核心主题，进而实现期刊或学科的主题提取研究。本研究创新性内容主要体现在：①提出了关键词的重要性度量方法，使得关键词的相似性考虑了关键词对不同文献的重要性影响，提高了主要关键词作为主题成员的描述力。②关键词相似性矩阵结合近邻传播聚类方法能够根据关键词之间的关系信息传播来自适应地找到主要关键词作为核心主题，避免传统分析手段因受人为主观因素影响而带来主题分析不准确的情况，提高核心关键词对主题思想的代表性。另外，本次从中国知网中收集了图书情报类某个核心期刊 2012—2016 年刊发的文献数据，通过数据处理及统计分析、关键词演化分析和核心主题演化分析来实现对基于关键词重要性和近邻传播聚类的主题分析方法在该刊物的主题研究工作。分析结果表明，新方法在考虑关键词重要程度的同时，还能有效地提出符合其他重要关键词的核心主题，有利于创作者和编辑部对期刊主题研究方向的把握，进而有助于提高学术研究和办刊质量。

该主题研究方法在对关键词的权重定义是依赖于作者给出关键词的顺序，意味着关键词的重要性计算某种程度会受到作者意识行为的影响。虽然大部分学者在提供关键词时会按重要顺序给出关键词，这方面有利于提高关键词权重的质量，但仍可能存在部分学者会不按与文献主题相关的重要程度来给出关键词，使得该方法提供的关键词权重计算偏离真实性。因此，较为客观的文献关键词重要性度量是未来有待进一步研究的工作。与此同时，本次仅将该方法应用于期刊主题的分析研究，对于将新方法应用于某个学科、某个专业和某些学术方向等主题分析也是未来值得研究和探讨的内容。

参 考 文 献

[1] 郑晓月, 牟冬梅, 琚沅红, 等. 学科知识结构主题演化模式研究——以图书情报学领域“计量学”主题为例[J]. 图书情报工作, 2017, 61(12): 32-41.

[2] 李明鑫, 王松. 近十年国内知识图谱研究脉络及主题分析[J]. 图书情报知识, 2016(4): 93-101

[3] 张春博, 王续琨. 主题裂变: 科学技术管理学的新走势[J]. 科学学与科学技术管理, 2012, 33(7): 5-11.

[4] 唐果媛, 张薇. 基于共词分析法的学科主题演化研究进展与分析[J]. 图书情报工作, 2015, 59(5): 128-136.

[5] 刘自强, 王效岳, 白如江. 多维度视角下学科主题演化可视化分析方法研究——以我国图书情报领域大数据研究为例[J]. 中国图书馆学报, 2016, 42(6): 67-84.

[6] 方龙, 李信, 黄永, 等. 学术文本的结构功能识别——在关键词自动抽取中的应用[J]. 情报学报, 2017, 36(6): 599-605.

[7] 李思志, 李佳骏, 李艳红. 管理科学与工程领域的创新轨迹研究——基于 TOP 期刊的文献计量和文本挖掘视角[J]. 中国管理科学, 2014, 22(S1): 56-62.

[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] 王沙沙, 丰景春, 薛松, 等. 基于知识图谱的 PPP 研究热点主题分析[J]. 科技管理研究, 2017, 37(17): 167-173.

[10] 秦春秀, 祝婷, 赵捧未, 等. 自然语言语义分析研究进展[J]. 图书情报工作, 2014, 58(22): 130-137.

[11] 张敏, 罗梅芬, 张艳. 国际文本挖掘研究主题群识别与演化趋势分析[J]. 图书馆学研究, 2017(2): 15-21.

[12] 赵京胜, 朱巧明, 周国栋, 等. 自动关键词抽取研究综述[J]. 软件学报, 2017, 28(9): 2431-2449.

[13] 李纲, 李轶. 一种基于关键词加权的共词分析方法[J]. 情报科学, 2011, 29(3): 321-324.

[14] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976.

[15] 查先进, 张晋朝, 严亚兰, 等. 网络信息行为研究现状及发展动态述评[J]. 中国图书馆学报, 2014, 40(4): 100-115.

[16] Guan R C, Shi X H, Marchese M, et al. Text clustering with seeds affinity propagation[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23 (4): 627-637.

[17] Sun L L, Guo C H, Liu C R, et al. Fast affinity propagation clustering based on incomplete similarity matrix[J]. Knowledge and Information Systems, 2017, 51(3): 941-963.

（责任编辑 车 尧）