doi:10.3772/j.issn.2095-915x.2018.06.004

基于关键词共现的学科领域研究空白 (Research Gaps) 发现

肖香龙[1,3] 李信[1,2,3] 高寒[1,3] 程齐凯[1,3]

1. 武汉大学信息管理学院 武汉 430072;
2. 华东师范大学学术评价与促进研究中心 上海 200241;
3. 武汉大学信息检索与知识挖掘研究所 武汉 430072

摘 要 文章首先对知识发现和研究空白（Research Gaps）发现的相关研究现状和问题梳理，设计和实现了一个基于关键词共现的学科领域 Research Gaps 发现算法；然后，以 Web of Science 中图书情报领域在 1900-2017 年的全部文献为研究对象，利用提出的算法进行图书情报领域 Research Gaps 识别，并对识别得到的 Research Gaps 填补文献与普通文献从时间分布、类别分布和著者分布等方面进行对比分析。研究结果表明：（1）本文提出的 Research Gaps 发现算法能够较好的发现图情领域中有实际意义的 Research Gaps，为推广至其他学科奠定基础；（2）Research Gaps 填补文献较普通文献具有较高的影响力。

关键词： 关键词共现；Research Gaps 发现；非相关文献；图书情报领域
中图分类号： TP182，G35
开放科学（资源服务）标识码（OSID）

Research on Scientific Gaps Recognition Based on Keywords Co-occurrence

XIAO Xianglong[1,3] LI Xin[1,2,3] GAO Han[1,3] CHENG Qikai[1,3]

1.School of Information Management, Wuhan University, Wuhan 430072, China;
2.Institute for Academic Evaluation and Development, East China University, Shanghai 200241, China;
3.Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan 430072, China

基金项目： 国家自然科学基金青年项目“基于深度语义挖掘的引文推荐多样化研究”（1704137）；华东师范大学学术评价与促进研究中心 2017 年开放项目“面向学术文本深度语义挖掘的学术评价方法创新研究”（IAED2017B04）。
作者简介： 肖香龙（1992-），硕士研究生，研究方向：信息检索与文本挖掘；李信（1991-），博士研究生，研究方向：文本挖掘与语义计量，E-mail: lucian@whu.edu.cn；高寒（1991-），硕士研究生，研究方向：信息检索与文本挖掘；程齐凯（1989-），讲师，博士，研究方向：信息检索，文本挖掘与自然语言处理。

doi:10.3772/j.issn.2095-915x.2018.06.004

Abstract The article first sorted out both related research status and problems in terms of knowledge discovery and research gaps discovery, then designed and implemented a research gap discovery algorithm based on keyword co-occurrence in subject areas. Moreover, this paper took the literature in the field of library and information science in Web of Science from 1990 to2017 as the object of study, The article uses the proposed algorithm to identify research gaps in the field of library and information science, and compared differences between literature filling research gaps and other literature concerning time distribution, category distribution and author distribution. The research results showed that:(1) The research gaps discovery algorithm proposed in this article can well find research gaps with practical significance in the field of library and information science and lays the foundation for the promotion to other disciplines; (2) Literature filling research gaps is more impactful than the other literature.

Keywords: Keywords co-occurrence; research gaps discovery; non-interactive literature; library and information science

引言

科研工作者通过阅读所在学科领域的相关科学文献，对所在领域的知识结构和研究前沿进行全面和准确的把握，是科学研究过程的必要环节。发现某一学科领域现存的研究空白(Research Gaps)并进行填补，是科学研究的重要目的之一。然而，在大数据时代，学术资源极速积累，依靠传统的人力方法探究已有文献中存在的 Research Gaps 已经表现出明显的高成本，低效率等问题[1]。如何高效准确的发现某一学科领域中的 Research Gaps，已经成为知识发现领域的研究热点和难点。由此研究者们开始对 Research Gaps 发现的计算方法进行研究，例如 20 世纪 80、90 年代 D. Swanson[2-4] 提出称为 ABC 模式的非相关文献知识发现思想、2009 年 C. Chen 等[5] 提出科学转化发现的解释和计算理论等。

Y. Peng 等[6] 在 2013 年提出利用 MeSH 词频进行计算，通过查找本该以较高频次共现、而实际未曾共现的 MeSH 词对的方法，进行生物医学领域 Research Gaps 发现，这一研究为在其他学科领域的 Research Gaps 发现提供了基础，但是存在以下不足：

（1）应用领域较窄，此研究仅进行了生物医学领域 Research Gaps 发现，在人文社科领域等其他领域是否可用尚未验证。

（2）对高质量受控词表依赖较为严重，此研究是基于 MeSH 词进行 Research Gaps 发现的算法设计，使用 MeSH 具有弊端，一方

面 MeSH 更新不及时，对最新的研究方向标注不足，另一方面大多数学科领域不具备类似 MeSH 的庞大受控词表，此方法通用性差。

（3）Research Gaps 的发现方法缺乏客观说服力和通用性，在 Y. Peng 等人的研究中，Research Gaps 发现算法设计中高频共现次数的阈值由研究者主观经验确定，缺乏客观理论支持。基于此，本研究在 Y. Peng 等人研究的基础之上，将 MeSH 的应用扩展为各学科领域文章中必不可少的关键词的应用，并对其算法进行改进，以弥补上述提到的不足之处。本研究以图书情报领域为例，利用关键词词频进行高频共现的关键词对计算，查找属于高频共现而实际并未共现的关键词对，此关键词对即为图情学科的 Research Gaps，并对算法设计中高频共现的关键词对频次阈值的确定进行了探索。

1 相关研究

本研究主要研究内容为遍历数据集中的所有关键词，找到从未在同一篇文章中共现，但是其理论上的共现频次属于高频的关键词对，本研究认为此关键词对即为当前学科领域 Research Gaps，并对上述算法中涉及的关键词对共现次数的高低频分界阈值进行了探索。本研究属于非相关文献的知识发现探索，因此，本研究从关键词共现、高频词界定和非相关文献知识发现三个方面进行了文献调研。

1.1 关键词共现相关研究

科研论文的关键词是研究者观点的凝结表达，也是文本计量和挖掘的重要指标。两个及以上数量关键词同时出现在同一篇文章中的现象称为关键词共现。通过关键词之间的共现关系和频次计量描述来表达一个学科领域集合内部的相互联系和结构，可以进行热点主题的揭示和发展动态的预测。

关键词共现是在探究学科发展结构和前沿中常用的研究方法。20 世纪 80 年代，J. Whittaker 等 [7] 认为在同一篇文章中的共现的关键词被作者认为具有联系，能被其他研究者接受的话可以影响未来相同关键词标引的研究。1996 年，I. Dagan 等 [8] 通多关键词共现方法进行了非结构化文本数据集中的知识发现研究。2011 年，Y. Ding 等 [9] 将共词分析和传统叙词表结合，提高信息检索多样性。近年来国内学者对关键词共现方法的使用与改进进行了较多研究。2011 年，冷伏海等 [10] 基于文献关键词进行了知识发现领域的三元共词分析。2013 年唐晓波等 [11] 融合了取自标题的增补关键词与论文作者的自标引关键词进行共词分析方法的改进。2014 年，邱均平等 [12] 运用作者关键词耦合、关键词分析等方法，对知识管理领域中的作者外在合作关系和潜在合作关系进行了挖掘。2016 年，郑彦宁等 [13] 以研究主题年龄和研究主题关注作者数量为指标，构建了基于关键词共现的研究前沿识别和跟踪方法。

综上，关键词能够反映学术论文的研究主题，研究者利用关键词共现的方法对学科发展结构的探索和前沿热点的追踪进行了大量研究，而用关键词共现的方法或者思路进行所在学科领域 Research Gaps 发现的研究很少。

doi:10.3772/j.issn.2095-915x.2018.06.004

1.2 高频词界定相关研究

当前高频词的选取主要有两种方法，自定义选取法和公式选取法。

自定义选取法即学者根据研究需要自行界定高频词的排序位次或者频次阈值。2002 年，邱均平等 [14] 通过排序前 20 位关键词对 2002 年的情报学领域研究热点进行了分析总结。2006 年，马费成等 [15] 利用词频不小于 10 次的关键词构建关键词表作为研究工具，对国内外知识管理领域的研究热点进行了分析。2014 年，W. Zhang 等 [16] 在其研究中选择词频不低于 26 的 178 个关键词进行分析。2016 年，M. Sedighi[17] 在其研究中将词频排序为前 20 的术语作为高频词，用得到的高频词作为分析的样本数据。

公式选取法即通过公式进行高频词的词频阈值计算。1967 年，A. Booth[18] 在研究中通过齐夫第二定律解释了低频词的分布规律，并且进行了实证数据的探索。1973 年，J. Donohue[19] 探索了高频词与低频词的临界值，对高频词的频次阈值进行了探究。1978 年，M. Pao[20] 结合 A. Booth 提出的低频词分布规律和齐夫定律，对 J. Donohue 的公式进行了改进，提出了新的计算高频词与低频词的分界公式。1992 年，孙清兰 [21] 在总结以上研究的基础之上提出了计算高频词、低频词分界值的公式，孙清兰在实证研究中验证此公式和 J. Donohue 公式效果一致。叶飞等 [22] 在 2013 年基于齐普夫定律提出了确定高频词的新方法。

综上，公式方法虽然没有形成统一的标准，但是相较于自定义方法，公式法避免了高频关键词选取中主观性过大的缺点。J. Donohue 和 M. Pao 等人较早的对高频词界定公式展开了研究，此后关于高频词的界定所用的公式法多利用以上公式或者由以上公式演变。在比较各方法效果之后，本研究采用 M. Pao 的高频词界定公式进行高频词的选定。

1.3 非相关文献知识发现相关研究

Research Gaps 的发现是非相关文献知识发现的重要研究方向，因此本研究对非相关文献知识发现相关研究进行了调研。1986 年 D. Swanson[23] 论述了 undiscovered public knowledge 概念，此后提出了非关联文献知识发现方法 [2-3]，之后多位学者对 D. Swanson 所提出的方法进行了更多场景的应用和更丰富的研究。D. Swanson 和 N. Smalheiser 建立了一个科学发现系统，此系统将不相关的文献集组合起来进行探索，可以得到之前单一数据集不能揭示的发现 [4,24]。2007 年，V. Torvi 等 [25] 提出了一个定量模型，用来评价 arrowsmith 两点法产生的 B-term data，使此方法的使用变得简单有效，此模型也可以应用到其他基于文献的知识发现系统中去。在国内，学者对于 D. Swanson 的情报学思想和方法进行了应用。黄水清等 [26] 验证了在中文文献中 D. Swanson 的非关联文献知识发现法的可行性。张晗等 [27] 利用 Arrowsmith 系统所挖掘的科研合作与交流的内容，找到两个机构研究内容的相似点（可以合作之处）和不同点（可以相互交流、学习之处）。

除了两点法，研究者还提出了不同的非相关文献知识发现方法。安新颖等 [28] 对非相关文献的知识发现方法进行了总结和介绍，包括基

于单词的词频统计方法、基于短语的词频统计方法、基于概念的知识发现方法和基于概念的词频统计方法。Y. Peng 等 [6] 通过探索在一个数据集中 MeSH 词对实际中没有达到预期中共现次数而发现一个学科领域的 Research Gaps（GAPs），并且对 GAPs 进行了分类和特征探究。李信等 [29] 基于文献词汇的不同语义功能，构建起基于学术文本词汇功能的知识发现体系和方法。

综上，在非相关文献知识发现领域的研究，很多研究存在以下不足：

（1）重复应用 Swanson 在 1986 年提出的方法到具体领域。

（2）研究方法依赖受控词表，集中在医学领域，应用范围较窄。而关键词在学术文章中作用类似于生物医学领域文章中 MeSH 词的作用，可以揭示文章研究主题，关键词共现的方法也被用来探究当前学科领域的框架结构与研究热点。因此，本研究尝试利用关键词为工具进行非相关文献的知识发现，拟基于图书情报学科领域文献关键词词频统计，利用高频词界定公式选取期望中共现频次较高而实际中并未共现的关键词对，进行图情学科中 Research Gaps 的发现。本研究创新点如下：

（1）将关键词共现方法应用到学科领域 Research Gaps 的发现研究中，减少对受控词表的依赖。

（2）基于 Pao 的高频词界定公式确定高频的共现关键词对的共现频次，对关键词对共现频次的高频阈值进行了探索。本文的下一部分将对利用关键词共现方法发现学科领域 Research Gaps 的具体过程进行详细阐述。

2 研究方法描述

2.1 数据来源

本研究数据来源于 Web of Science 核心合集数据库，学科领域选择为 Information science and Library science，时间选择为 1900-2017，文献类型选择为 article、proceedings paper 和 review。获取的文章数据元素包括标题，作者，作者关键词（keywords），增补关键词（keywords plus），摘要，期刊，发表时间，参考文献数量，被引次数和使用计数（Usage Count）。如表 1 所示，对部分文章元素做了解释。数据获取的截止时间为 2017 年 12 月 14 日，共获取文章数据 163529 条。

表 1 部分文章数据元素的解释

文章元素	作用 / 含义
作者关键词	文章作者自己标注的关键词
增补关键词	由标题或摘要中提取的关键词，对文章主题进行补充说明
Usage Count	用户认为此文章满足了其信息需求的次数统计（如点击链接获取全文或者导出文章题录信息操作）

本研究统计了全部文章、只包含作者关键词文章和包含作者关键词与增补关键词文章数量的年份分布及变化趋势，见图 1。

从图 1 中可以看出，图书情报学科领域的所发表的文章，在 20 世纪上半叶数量少且增长幅度小，进入下半叶之后，文章数量明显上升，且增长稳定。在最近十年，文章数量出现 2008 年和 2016 年两个峰值，其余年份文章数量变化比较稳定。全部文章、只包含作者关键词的文

doi:10.3772/j.issn.2095-915x.2018.06.004

章和同时包含作者关键词与增补关键词的文章数量的年份变化趋势基本相同。基于此，本文确定了 2.2 的关键词选择策略和 2.3 数据集划分策略。

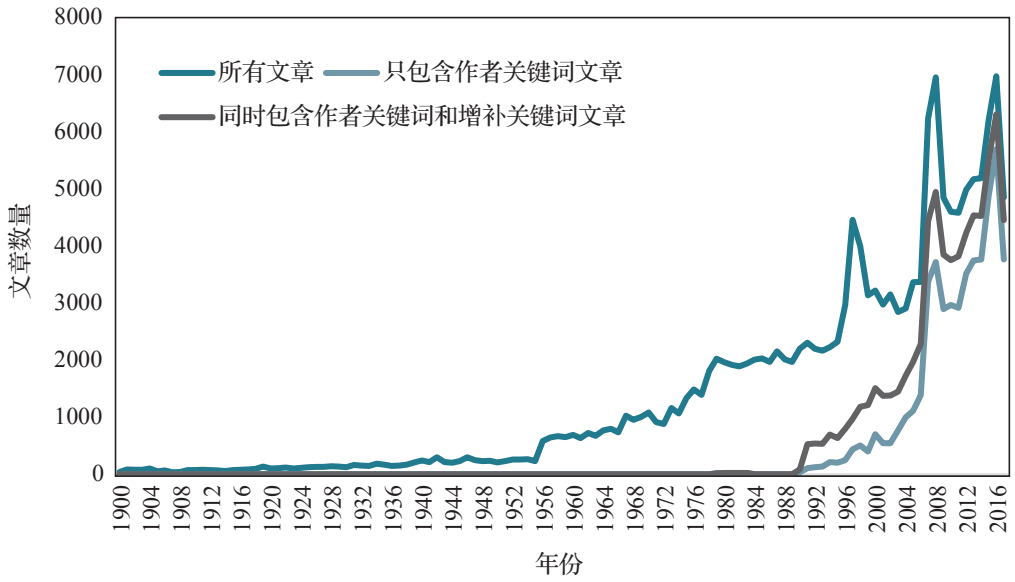

图 1 文章数量年份分布

2.2 关键词的选择

从图 1 来看，在 1990 年之前，几乎没有文章包含作者关键词或增补关键词，整个数据集中包含关键词的文章在全部文章中占比不高。数据统计结果表明，只包含作者关键词的文章数量较少，同时包含作者关键词和增补关键词的文章数量较之增加了 38.8%。表 2 展现了不同类型文章数量对比。

表 2 不同关键词类型文章数量

文章类型	文章数量
全部文章	163529
只包含作者关键词文章	49926
同时包含作者关键词和增补关键词文章	69290

基于此，本研究采用 Web of Science 核心合集数据库中图情学科文章的作者关键词与增补关键词作为发现 Research Gaps 工具。由于数据集中作者关键词数量较少，且作者关键词易受到作者主观意愿的限制，在反映文章内容全面性上效果受限，因此本研究采取了作者关键词与增补关键词组合作为发现 Research Gaps 的工具。作者关键词 + 增补关键词的模式对于学科领域特点的揭示效果更有效率，更能够表现其发展趋势和变化特点 [30-31]。

本研究对文章中的作者关键词和增补关键词字段进行抽取，为了减少在标引过程中，由于词的时态或者词性带来的不规范的影响，利用 Stanford NLP 对关键词进行词干还原，并且进行了单个关键词出现频率的统计。

2.3 数据集在时间维度的划分

本研究根据图 1 中对文章数量逐年变化趋

势进行了观察与总结，对文章数量与关键词数量的逐年统计，将获取的数据以时间维度划分为两个数据集，2012 年及之前文章为时间段 1 数据集 ,2013 年及之后文章为时间段 2 数据集，两个数据集均去除既没有作者关键词也没有增补关键词的文章。本研究拟在时间段 1 数据集中进行 Research Gaps 的发现，在时间段 2 数据集中验证所得 Research Gaps 是否被填补，并对填补 Research Gaps 的文章进行特征分析。文章数量和关键词数量在两个数据集中的分布见表 3。

表 3 文章数量与关键词数量在两个数据集中的分布

时间段	文章数量	作者关键词 + 增补关键词数量（不重复）
1900-2012	43968	62850
2013-2017	25322	57051
1900-2017	69290	102008

2.4 学科领域Research Gaps识别方法

本研究抽取文章题录数据中的所有作者关键词和增补关键词，找出预期中共现次数在某一阈值 N 以上，但是在时间段 1 数据集中并没有共现的关键词对，此关键词对即被认为是在时间段 1 数据集中的 Research Gaps。此方法具体描述如下：

Algorithm 1 Research Gaps 识别

Input：dataset of period 1 ——DS1 // 时间段 1 数据集

set of keywords ranked by frequency descendly from dataset of period 1——List<Keyword> keywordList; // 从时间段 1 数据集中得到按频次倒序排列的关键词集合

Output：set of pair of keywords——List <Keywords> KeywordsList // 被认为 Research Gaps 的关键词对集合

1. $n = \frac{1+\sqrt{1+8I_1}}{2}$ //n 为高频关键词与低频关键词区分词频值，$_1$ 为词频为 1 的关键词数量
2. N = //N 为高频关键词对与低频关键词对区分频次临界值，向上取整
3. for keywordList // 对关键词集合进行循环
4. $M = \frac{K(i).freq*K(j).freq}{DS1.size}$ // 两两关键词组队 , M 为预期中关键词对共现频次
5. if // 因为 N 为临界值，所以判断关键词对为高频采用比较值为 N+1
6. if p = 0 // p 为在时间段 1 数据集中 K(i) 和 K(j) 关键词对实际共现次数
7. KeywordsList.add(K(i)K(j)) // 符合上述条件关键词对为 Research Gaps
8. end if
9. end if
10. end for

本研究采用的高频关键词频次阈值计算方法为 Pao 提出的公式：

$$n = \frac{1+\sqrt{1+8I_1}}{2}$$

代表频次为 1 的关键词数量 I_1 为 44934，n 向上取整数，n 为 301。基于得到的 n 可以得出预期中共现的关键词对的高频分界值（设为 N）利用下式得到：N=301 × 301/43896，N 向上取整为 3。由于 3 为高低频分界的临界值，所以本文以 N+1 作为预期中高频的关键词对共现次数阈值，即在时间段 1 数据集中预期共现次数在 4 次及以上，而实际中并未共现的关键词对为

doi:10.3772/j.issn.2095-915x.2018.06.004

Research Gaps。

2.5 Research Gaps及填补文章的特征分析

在时间段 2 数据集中验证由时间段 1 数据集得到的 Research Gaps 是否被填补，即代表 Research Gaps 的关键词对在时间段 2 数据集文章中有共现现象。对由时间段 1 数据集得到的 Research Gaps 和由时间段 2 数据集得到的填补 Research Gaps 的文章进行特征分析。

3 实验结果分析

3.1 Research Gaps识别结果

通过上述方法，在时间段 1 数据集中，共获取不重复作者关键词与增补关键词 102008 个，可以得到关键词对 907797732 个，符合上述阈值条件的关键词对为 7035 个，约占全部关键词对的 0.0000078；在 7035 个符合阈值条件的关键词对中，实际在时间段 1 数据集中并未共现的有 519 个，即 Research Gaps 为 519 个，约占符合阈值条件关键词对的 0.074。由上述数据可知，Research Gaps 在已有文献的可能研究方向中所占比例极小，以传统人力阅览或者向专家咨询的方式是极其难以高效和准确发现的。

本研究在时间段 1 数据集中发现 519 个 Research Gaps，在时间段 2 数据集中，共有 499 篇文章填补了 250 个 Research Gaps，填补 Research Gaps 的概率为 48.2%。根据所获取到的文章数据，本研究将从时间分布、Research Gaps 类别分布、文章影响力、文章著者等方面对 499 篇填补 Research Gaps 文章进行分析。

3.2 Research Gaps特征分析

本研究对得到的 Research Gaps 进行了观察，发现其并不是完全相似的，在两个关键词组合成为关键词对的场景上，可以分为三类：

（1）专业型：此类指构成 Research Gaps 的关键词对，单独关键词在图书情报领域都代表比较热门的研究方向。在两个单独的研究方向上，前人已经在做出了较多的研究，但是缺乏对他们的组合探索。如 internet / knowledge transfer，knowledge management / digital divide 等。

（2）交流型：此类指构成 Research Gaps 的关键词对由来自于图书情报学科领域内和领域外的关键词交流组合。当前的知识积累极其丰富，知识单元的划分也越来越细，由于时间、精力和技术的限制，在两个领域之中的研究人员可能没有意识到另外一个关键词所代表的领域的研究方向或者未能进行跨领域的交流沟通。如 library / physician 等。

（3）标记型：此类 Research Gaps 在关键词的标记中作为标记的副产物出现。两个单独在图情领域出现频次较高的关键词，由于属于上下位关系或者组合在起没有实际意义，却满足本文中所认为的 Research Gaps 的条件，例如 information retrieval / retrieval。

Research Gaps 在上述类别中的数量与占比分布如表 4 所示：

从表 4 中可以看出，专业型 Research Gaps 占比最多，约为 2/3；其次为交流型 Research Gaps，约为 1/4；专业型 Research Gaps 和交流型 Research Gaps 共同占比 89.4%；由关键词标引所造成的副产物标记型 Research Gaps 占比

约为 1/10。由于缺乏高质量的受控词表，图情领域文献的关键词标引缺乏指导标准，形成了缺乏实际意义的标记型 Research Gaps，但其总体占比较小。专业型 Research Gaps 和交流型 Research Gaps 占据全部的绝大部分，可见本研究得到的 Research Gaps 大部分都是有实际含义的，此方法能够较为准确高效的发现图书情报学科领域中的 Research Gaps。接下来本文将进行在时间段 2 数据集中所得到的填补 Research Gaps 的文章的特征分析。

表 4 Research gaps 类别分布

类别	举例	数量	占比
专业型	internet / knowledge transfer	333	64.2%
交流型	library / physician	131	25.2%
标记型	information retrieval / retrieval	55	10.6%
全部		519	100%

3.3 填补空白文章特征分析

3.3.1 填补 Research Gaps 文章的时间分布

本研究统计了填补所得 Research Gaps 的 499 篇文章在时间段 2 中各年份的分布并且将其与时间段 2 数据集文章整体时间分布情况作对比，结果如表 5 所示：

表 5 填补 research gaps 文章年份分布

年份	填补空白文章数量 / 占比	时间段 2 文章数量 / 占比
2013	98 / 19.6%	4536 / 17.9%
2014	75 / 15.0%	4524 / 17.9%
2015	111 / 22.2%	5502 / 21.7%
2016	123 / 24.6%	6306 / 24.9%
2017	92 / 18.4%	4454 / 17.6%
总和	149 / 100%	25322 / 100%

从表 5 数据来看，填补 Research Gaps 的文章和时间段 2 数据集全部文章在年份分布上整体变化趋势是相一致的，且每年数据占总量的比例相近。在全部文章分布较多的 2015 年、2016 年，填补 Research Gaps 的文章数量也为 5 年中最多，且二者占总数比例相近。此数据说明 Research Gaps 的填补并不是偶然爆发的，而是整个学科发展趋势而变化的一部分，在时间分布上面没有明显的集中或者分散趋势。

3.3.2 填补 Research Gaps 文章的 Research Gaps 类别分布

本小节对时间段 2 数据集中填补 Research Gaps 的文章进行了在 Research Gaps 类别上的分布统计，如表 6 所示：

表 6 填补 research gaps 文章在 research gaps 类别上的分布

类别	文章数量 / 所填补 Research Gaps 数量	Research Gaps 数量	填补率
专业型	341 / 171	333	51.4%
交流型	123 / 62	131	47.3%
标记型	35 /17	51	33.3%
全部	499 / 250	519	48.2%

由表 6 可以看出，在填补 Research Gaps 文章的 Research Gaps 分类分布上面，填补专业型 Research Gaps 和交流型 Research Gaps 的文章占全部填补 Research Gaps 文章的 93%，标记型 Research Gaps 填补文章占比很小。35 篇文章填补了 17 个标记型 Research Gaps，通过对填补标记型 Research Gaps 的文章进行观察分析，被填补的标记型 Research Gaps 由 system、model、information 等在关键词中频繁出现，但是其组

doi:10.3772/j.issn.2095-915x.2018.06.004

合成的关键词对不能表示为明显的研究方向，填补标记型探究空白的 35 篇文章，在其增补关键词中会出现代表 Research Gaps 的关键词对或者其中一个关键词，增补关键词是自动标引的产物，所以填补标记型 Research Gaps 的文章是关键词标记的产物，而非科研人员进行有明确方向的研究所产生的。这类整体占比较小，对整体结果影响较小。

从单个 Research Gaps 的填补文章数量来看，平均一个 Research Gaps 被两篇文章填补；从填补概率看，专业型 Research Gaps 和交流型 Research Gaps 被填补概率高于整体被填补概率，标记型 Research Gaps 被填补概率明显低于整体被填补概率，由于关键词标记的原因产生了填补标记型 Research Gaps 的文章，所以其填补概率明显低于整体的填补概率。专业型 Research Gaps 的填补概率与交流型 Research Gaps 的填补概率相近。以上数据说明，填补 Research Gaps 的文章在 Research Gaps 类型的分布上面是均匀的，没有明显的集中或者分散的趋势；Research Gaps 被填补的可能性在 Research Gaps 的类型分布上也是均匀的。

3.3.3 填补 Research Gaps 文章的文章影响力分析

基于所获取到的文章数据，本节将从被引量和 Usage count 两个角度分析填补 Research Gaps 的文章的影响力，数据结果如表 7 和表 8 所示：

表 7 填补 research gaps 文章被引量统计及与整体对比

类别	篇数（有引用信息的）	总引用数	平均引用数
填补空白的文章	499	2049	4.11
时间段 2 数据集文章	21844	63174	2.89
全时间段数据集文章	69290	898125	12.96

从表 7 中可以看出，填补 Research Gaps 的文章相比时间段 2 数据集的全部文章的平均引用数量提高了 42%，幅度明显，但是相比全时间段文章的平均引用量来说明显较低，仅相当于后者数据的 1/3。引文行为的表现中存在马太效应 [32-33]，即被引用次数越高的文章受到引用的可能性越高，受到引用持续的时间越长。经过统计，在引用量由高至低排列的前 1000 篇文章中，仅有 9 篇文章来自 2013-2017 年，其余皆来自于 1900-2012 年，前 1000 篇文章平均引用量为 284.39。在时间段 1 数据集的文章中，较早发表的文章中存在许多经典文章，其受到的引用量高出平均水平很多，而随着时间的积累和马太效应的影响，拉高了总体的平均引用量。综上，在与同时段数据集全部文章的比较中，填补 Research Gaps 的文章获得了明显更高的平均引用数。

表 8 填补 research gaps 文章 usage count 统计及与整体对比

类别	篇数（部分文章无此数据）	总数（近 180 天 /2013 年以来）	篇均（近 180 天 /2013 年以来）
填补空白的文章	496	2557 / 14474	5.16 / 29.2
时间段 2 数据集文章	21680	64053 / 383867	2.95 / 17.71
全时间段数据集文章	68871	110402 / 919336	1.60 / 13.35

Usage count 是利用 Web of Science 数据库查找信息的人认为此文章对其有用的次数，表现为导出文章题录信息或者通过连接查找全文等操作。从表 8 数据来看，填补 Research Gaps 的文章在满足信息查询者信息需求的表现上较时间段 2 数据集全部文章和全时间段数据集文章的平均值均有明显的大幅提高。此结果表示填补 Research Gaps 文章填补了尚未探索的研究方向或者领域，较好的满足了在此研究方向或领域寻找参考信息的用户的信息需求。

综上，由于填补 Research Gaps 的文章所研究主题为之前尚未进行探索的研究领域和方向，其在被引量和 Usage count 数据表现上较之于全部文章的平均水平有明显提升，影响力较全部文章的平均水平有明显提升。

3.3.4 填补 Research Gaps 文章的著者分析

本小节对填补 Research Gaps 的文章的著者进行了统计分析，将其和构成此 Research Gaps 的关键词在时间段 1 数据集中所在文章的著者进行重合比对，结果如表 9 所示：

表 9 填补 research gaps 文章著者与时间段 1 数据集相同关键词文章著者重合分析

类别	Research Gaps 数量	重合比例
和两个关键词所在文章著者重合	121	48.4%
和其中一个关键词所在文章著者重合	230	92%

从表 9 中可以看出，92% 的填补 Research Gaps 文章和构成此 Research Gaps 的其中一个关键词所在的文章有着作者重合，与两个关键词所在文章全部由著者重合的占比为 48.4%。基于此可以得出，发表填补 Research Gaps 文章的作者在之前就已经在相关方面进行了研究。知识在积累和传播过程中具有时空的延续性 [34]，个人知识的积累也是如此。学者的研究领域和方向在时间维度上面具有延续性，填补一个领域的 Research Gaps 的学者通常在其填补 Research Gaps 的文章发表之前，就已经做出了相关研究。这一发现为科研工作者进行 Research Gaps 的填补研究提供了思路，即在进行以代表 Research Gaps 的关键词对为主题的相关研究时，可以通过查找构成 Research Gaps 的关键词对在之前研究中的作者的学术成果进行相关研究调研。

4 讨论

本研究利用关键词为研究对象设计学科领域 Research Gaps 发现方法，并且在图书情报领域进行了验证。在 1900-2012 年时间段的图情领域文章中，利用所提出的方法共发现 519 个 Research Gaps，在 2013-2017 年时间段的图情领域文章中，共有 499 篇文章填补了 250 个 Research Gaps，填补 Research Gaps 的概率为 48.2%。将 Research Gaps 分为专业型，交流型和标记型三类，并且分析了 Research Gaps 和填补 Research Gaps 文章在类别上的分布。此外，通过对填补 Research Gaps 文章的被引数量、Usage count 等特征进行分析，发现填补 Research Gaps 文章相对于普通文章，具有较高的影响力。此方法可以有效的发现图情领域的 Research Gaps，据此进行 Research Gaps 和填补 Research Gaps 文章的特征分析。此方法研究

doi:10.3772/j.issn.2095-915x.2018.06.004

对象为学术论文关键词，具备在其他学科领域应用的基础。

Y. Peng 等 [6] 在生物医学领域进行 Research Gaps 发现，与本研究所使用方法存在不同：一方面本研究将前者研究中研究对象由 MeSH 扩展为学术论文关键词，设计算法发现学科领域中的 Research Gaps，使 Research Gaps 发现方法的应用不再限制于该领域，不再依赖于高质量受控词表，可有效解决 Research Gaps 发现方法的学科适用问题；另一方面，本研究提出应用高频词界定公式来界定高频共现的关键词对频次阈值，前者利用主观经验进行高频共现的 MeSH 词对频次阈值的确定，本研究所使用方法更加具有客观性。

5 结语

本研究通过对当前对非相关知识发现领域和关键词共现领域相关研究的调研，认为当前研究方法存在应用学科范围较窄，对高质量受控词表依赖大等缺点。然后以图情领域为例，将数据集分为两个时段，利用计算关键词共现的期望频次与实际频次的方法发现时段 1 数据集 Research Gaps，并且对得到的 Research Gaps 和时段 2 数据集中填补 Research Gaps 的文章进行了特征分析，结果证明科研工作者应该关注在当前研究中彼此独立却可能存在潜在联系的研究方向或领域，填补 Research Gaps 的文章获得了更高的影响力和关注度。同时，关键词作为学术论文的重要组成部分，表明此方法具备在其他学科领域进行应用的基础。

但本研究仍存在一些问题，需要未来进一步探索：此 Research Gaps 发现方法在其他学科应用、跨学科应用等尚待探究；由于之前相关研究很少，对于 Research Gaps 发现结果的评估缺乏相应标准，因此可对 Research Gaps 发现结果评估标准和评估方法进行深入探索；在对 Research Gaps 和填补 Research Gaps 文章的特征分析上，在未来可以引入机器学习的方法，对其进行多种特征进行识别和学习，进行更深入的探索。

参考文献

[1] 李信, 李旭晖, 陆伟,等. 大数据驱动下的图书情报学科热点领域挖掘——面向WOS题录数据的实证视角[J]. 图书馆论坛, 2017, 37(4): 49-57.

[2] Swanson D R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge[J]. Perspectives in biology and medicine, 1986, 30(1): 7-18.

[3] Swanson D R. Migraine and magnesium: eleven neglected connections[J]. Perspectives in biology and medicine, 1988, 31(4): 526-557.

[4] Swanson D R, Smalheiser N R. An interactive system for finding complementary literatures: a stimulus to scientific discovery[J]. Artificial intelligence, 1997, 91(2): 183-203.

[5] Chen C, Chen Y, Horowitz M, et al. Towards an explanatory and computational theory of scientific discovery[J]. Journal of Informetrics, 2009, 3(3): 191-209.

[6] Peng Y, Bonifield G, Smalheiser N R. Gaps within the Biomedical Literature: Initial Characterization and Assessment of Strategies for Discovery[J]. Frontiers in Research Metrics & Analytics, 2017, 2.

[7] Whittaker J. Creativity and Conformity in

Science: Titles, Keywords and Co-word Analysis[J]. Social Studies of Science An International Review of Research in the Social Dimensions of Science & Technology, 1989, 19(3): 473-496.

[8] Dagan I, Feldman R, Hirsh H. Keyword-based browsing and analysis of large document sets[C]. Proceedings of the symposium on document analysis and information retrieval (SDAIR-96), Las Vegas, Nevada. 1996.

[9] Ding Y, Chowdhury G G, Foo S. Incorporating the results of co-word analyses to increase search variety for information retrieval[J]. Journal of Information Science, 2011, 26(6): 429-451.

[10] 冷伏海, 王林, 李勇. 基于文献关键词的三元共词分析方法——以知识发现领域为例[J]. 情报学报, 2011, 10(10): 1072-1077.

[11] 唐晓波,肖璐. 融合关键词增补与领域本体的共词分析方法研究[J]. 现代图书情报技术, 2013, 29(11): 60-67.

[12] 邱均平, 刘国徽. 基于社会网络和关键词分析的作者合作研究——以国内知识管理领域为例[J]. 情报科学, 2014(6): 3-7.

[13] 郑彦宁, 许晓阳, 刘志辉. 基于关键词共现的研究前沿识别方法研究[J]. 图书情报工作, 2016, 60(4): 85-92.

[14] 邱均平, 赵蓉英, 侯经川. 2002 年国内外情报学发展动向分析[J]. 情报学报, 2003, 22(5): 515-519.

[15] 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.

[16] Zhang W, Zhang Q, Yu B, et al. Knowledge map of creativity research based on keywords network and co-word analysis, 1992–2011[J]. Quality & Quantity, 2015, 49(3): 1023-1038.

[17] Sedighi M. Application of word co-occurrence analysis method in mapping of the scientific fields (case study: the field of Informetrics)[J]. Library Review, 2016, 65(1/2): 52-64.

[18] Booth A D. A “Law” of occurrences for words of low frequency[J]. Information and Control, 1967, 10(4): 386-393.

[19] Donohue J C. Understanding Scientific Literature: A Bibliographic Approach[M]. The MIT press: Cambridge, 1973

[20] Pao M L. Automatic text analysis based on transition phenomena of word occurrences[J]. Journal of the Association for Information Science and Technology, 1978, 29(3): 121-124.

[21] 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, 18(2): 78-81.

[22] 叶飞, 宋志强. 一种基于齐普夫定律的确定语料中高低词频分界点的新方法——以科学计量研究为例[J]. 情报学报, 2013, 32(11): 1196-1203.

[23] Swanson D R. Undiscovered public knowledge[J]. The Library Quarterly, 1986, 56(2): 103-118.

[24] Smalheiser N R, Torvik V I, Zhou W. Arrowsmith two-node search interface: a tutorial on finding meaningful links between two disparate sets of articles in MEDLINE.[J]. Computer Methods & Programs in Biomedicine, 2009, 94(2): 190-197.

[25] Torvik V I, Smalheiser N R. A quantitative model for linking two disparate sets of articles in MEDLINE[J]. Bioinformatics, 2007, 23(13): 1658-1665.

[26] 黄水清, 程冲, 李志燕. 开放式非相关文献知识发现方法在中文文献中的验证[J]. 情报理论与实践, 2008, 31(2): 246-250.

[27] 张晗, 崔雷, 姜洋. 运用非相关文献知识发现方法挖掘科研机构潜在的合作方向[J]. 现代图书情报技术, 2006, 1(4): 45-48.

[28] 安新颖, 冷伏海. 基于非相关文献的知识发现原理研究[J]. 情报学报, 2006, 25(1): 87-93.

[29] 李信, 程齐凯, 刘兴帮. 基于词汇功能识别的科

doi:10.3772/j.issn.2095-915x.2018.06.004

研文献分析系统设计与实现[J]. 图书情报工作, 2017(1): 109-116.

[30] Ding Y, Chowdhury G G, Foo S. Bibliometric Cartography of Information Retrieval Research by Using Co-Word Analysis[J]. Information Processing & Management, 2001, 37(6): 817-842.

[31] 张超星, 谭宗颖, 朱相丽,等. Web of Science中关键词的利用方式对情报分析结果的影响及选择建议——基于超临界二氧化碳技术领域的实证分析[J]. 情报科学, 2017(6): 73-79.

[32] 于鸣镝. 三论引文选刊的局限性[J]. 图书情报工作, 1990, 34(6): 22-23.

[33] 李国红. 我国计算机领域学术论文引用中的马太效应——以《计算机学报》和《计算机研究与发展》为例[D]. 新乡: 河南师范大学, 2010.

[34] 陈则谦. 基于内容开放平台的公共知识传播动力机制研究[D]. 北京: 北京大学, 2012.