基于引用共词网络的领域基础词汇发现研究*

程齐凯　王佳敏　陆　伟
(武汉大学信息管理学院　武汉 430072)
(武汉大学信息检索与知识挖掘研究所　武汉 430072)

摘要: 【**目的**】从学术文献中发现领域基础词汇，为把握学科知识结构和发展脉络提供支持。【**方法**】将引文网络引入到共词分析中，构造关键词之间的引用共词网络，采用 PageRank 算法对候选词汇重要性进行排名，基于约 11 万篇计算机领域文献集进行实证研究。【**结果**】从定性和定量的角度与词频法和共词分析法进行对比，结果表明本文方法效果较好，能更好地拟合专家人工筛选结果，盲选实验的平均准确度达 72.6%。【**局限**】仅以计算机领域为例进行实验。【**结论**】本研究提出一种融合引用共词网络和 PageRank 算法的领域基础词汇发现策略，能够提高领域基础词汇发现的效率和质量。

关键词: 基础词汇　引用共词网络　PageRank　词频法　共词分析
分类号: G350
DOI: 10.11925/infotech.2096-3467.2018.1159

1　引　言

领域基础词汇是刻画、表征领域知识的基本信息承载单元，是领域知识结构和发展脉络中的核心单元，也是信息检索和信息抽取的重要单元。词汇是科学知识的载体[1]，而关键词是文献核心内容的浓缩和提炼，能直接反映领域的知识点分布和知识结构[2-3]，因此领域基础词汇发现主要是利用领域相关文献中关键词之间的语义关系对文献集合进行分析，进而发现学科领域基础词汇，以把握学科知识结构和发展脉络。

以关键词作为基本知识单元的研究主要集中在知识结构和演化[4-5]、主题和热点发现[6-7]等研究中，常用的方法为词频法或共词分析法，一般根据主观经验或一定的规则筛选部分关键词进行分析。但词频法仅仅考虑词汇的出现频次，容易忽略词频不高但较为重要的领域词汇，而共词分析法只关注文献自身关键词之间的关系，忽略了不同文献之间的间接关联，在实际中两种方法得到的结果往往包含较多语义过于宽泛的词汇或者上位词，但是这些词汇并不具备领域特色，难以有效揭示领域的研究特征[8]，也就无法很好地表征领域研究基础。实际上，不同学术文献的学术价值存在差别，被引次数较多的文献往往学术价值较高，而学术价值较高的文献所包含的关键词比学术价值较低的文献关键词更能反映学科的研究内容[9]。

为此，本文将文献之间的引文关系引入到共词分析方法中，构造文章之间引用关系构成的关键词共现网络即“引用共词网络”，并通过 PageRank 算法对该网络中的领域基础词汇进行发现。在计算机领域 11 万余

通讯作者：王佳敏，ORCID: 0000-0003-3954-0381，E-mail: wangjm@whu.edu.cn。

*本文系国家自然科学基金面上项目“面向词汇功能的学术文本语义识别与知识图谱构建”(项目编号: 71473183)、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”(项目编号：71704137)和中国博士后科学基金资助项目“基于词汇功能的科研资源推荐”(项目编号: 2016M602371)的研究成果之一。

篇学术文献集上进行实验，并与传统的词频分析法和共词分析法进行对比分析。

2 相关研究

与本文最相关的研究主要集中在关键词筛选任务中。在基于关键词的领域知识分析研究时，需要从大量关键词中提取出最能表征数据特征的小部分作为分析对象[10]。词频是关键词筛选最直接的依据，例如，Wang 等[11]对所有术语词频进行统计并从高到低排序，根据个人经验选取前 N 个高频词作为分析的样本数据。Hu 等[3]在分析信息检索领域的主题结构和演化时，从原始关键词中选择词频不小于 10 次的关键词共 150 个作为分析对象。这类方法虽然简单可行，但凭借研究者的经验进行选择，主观性较强，往往会忽略掉一些词频不高但能够表征领域特色的基础词汇。为更客观地确定高频词的阈值，Donohue[12]根据齐普夫第二定律[13]提出高频低频词分界公式。Yang 等[14]根据 Donohue 高低频词分界公式获取医学信息学领域频次超过 36 次的 35 个高频 MeSH 词作为研究对象。Yan 等[15]根据 Donohue 公式得到高频词阈值为 120，但只有 7 个关键词超过该阈值。这种定量方法在一定程度上避免了主观经验，但当研究领域范围过大时，使用这类方法容易获得太过抽象、具体的词以及领域外不相关的词[16]。此外，还有学者将关键词集合转化为网络，采用网络指标(如网络节点度数、中介中心性、特征向量中心性等)[17]或相关方法(如 K-core 分解[18]、核心/边缘结构[19]、惩罚性矩阵分解[20])进行关键词筛选。这类方法通过网络结构发现重要的节点，取得了一定成效，但由于在关键词构建的网络中，上述指标与词频仍然线性相关，因而抽取到的关键词与高频词并无太大差异[16]。

近年，部分学者提出将引文关联关系引入词语共现或实体共现分析中，提出结果更为有效、思路更为可靠的新方法[9,21]。例如，Ding 等[21]提出实体计量用来衡量不同层次知识单元的影响，以 Metformin 药物为例构建实体–实体引文网络(Entity-Entity Citation Network)，通过对比验证了该方法可以有效发现知识实体之间的关联。Song 等[22]提出施引文献和被引文献的知识实体之间存在相关关系，并构建了生物医学文献中基于基因实体的引用共词网络(Gene-Citation-Gene Network)，通过与传统的共词网络(Gene-Gene Network)对比，发现前者更能揭示知识实体之间的一些隐含关系。李树青[9]利用引文分析思想计算文献的学术价值，并以此计算文献和引用文献的词语共现对权重值，完成本体结构中层次概念联系的表达和设计。吴清强等[23]认为高影响因子期刊上或被引次数较高的文献中的词更具有代表性，根据文献的来源期刊、被引次数等属性赋予关键词不同的权值，从而构建基于论文属性的加权共词分析模型。葛菲等[24]提出引文分析能较好地反映文献集中存在引用关系的主题，内容词分析方法反映的是已有文献集中关心的主题，将二者结合起来在揭示科学结构方面能产生更好的效果。

综上，传统基于词频或共词分析的方法关注的都是文献自身关键词的频次或关键词对之间的共现关系，且没有对不同学术价值文献的关键词进行区分，得到的领域词汇往往外延过大，不能够很好地表征领域研究基础。不同学者从关键词加权、文献属性差异以及将引文关系考虑到共词分析方法中等角度进行了有益尝试并取得了一定的改进效果，为本文基于引用共词网络的领域基础词汇发现提供借鉴。引文分析通过文献之间的引用关系，以一种间接方式反映了不同文献知识单元之间的关联，而共词分析法是对当前文献的直接计量，反映已有文献集中知识单元之间的关系。将二者结合起来，能够同时发挥引文分析在挖掘文献之间间接关联关系和隐藏的重要知识节点上的优势与共词分析在挖掘语义关联、揭示知识结构上的优势，能够丰富领域知识单元之间的关联网络，从而更加完整、准确地发现领域基础词汇。

3 研究方法

3.1 基本思路

本文以计算机领域为例，对 ACM 数据集中包含的关键词进行抽取，分别通过关键词的共现对和引用共现对构建关键词共词网络和引用共词网络，然后通过 PageRank 算法对网络节点重要度进行计算，根据 PageRank 值高低排名抽取出领域基础词汇，整体研究流程如图 1 所示。

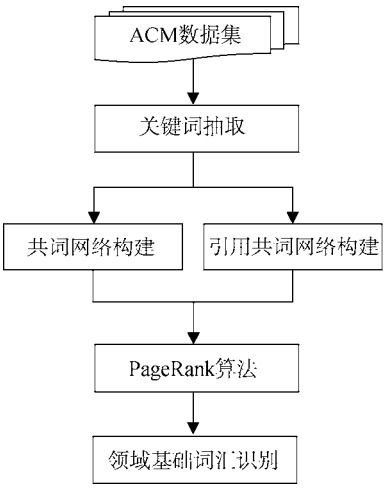

图 1　整体研究流程

3.2　共词网络和引用共词网络构建

在关键词抽取的基础上，构建两种类型的共词网络。一种是传统的共词网络，其原理是如果两个关键词在同一篇文献中出现，则这两个关键词形成共现关系。将所有具有共现关系的“关键词–关键词”(Keyword-Keyword，KK)对关联起来，就形成了关键词共词网络。另一种是基于引用关系的关键词共词网络，即认为当两篇文献存在引用关系，则施引文献的关键词和被引文献的关键词之间可以通过这种引用关系构建关键词对共同出现的情况，综合文献集中所有的“关键词–引用–关键词”(Keyword-Citation-Keyword，KCK)对，便可以生成引用共词网络。两种类型网络的关键词对构建过程如图 2 所示。

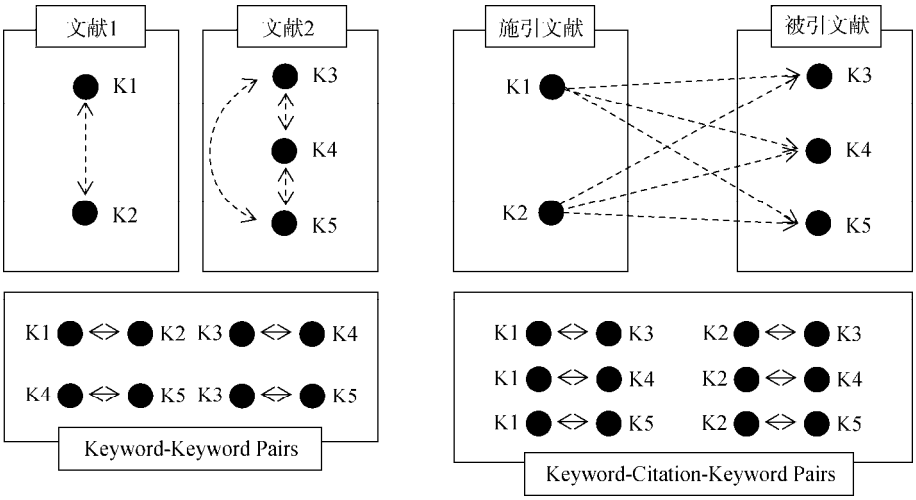

图 2　KK 对及 KCK 对构建过程

3.3　基于 PageRank 算法的基础词汇发现

PageRank 算法是 1998 年由 Brin 等[25]提出的一种基于链接分析的网页排序算法，通过分析网络的链接结构获得网络中网页的重要性排名。基本思想是将所有网页及网页之间的链接视为一个有向图，节点是网页，节点重要性由链接该节点的其他节点的重要性和数量决定。由于关键词共现网络与网页链接网络本质相同，均为有向图，在关键词共现网络有向图中，一个节点代表一个关键词，节点之间的连线代表关键词的共现关系或引用共现关系，将 PageRank 算法应用在共词网络中，可以同时兼顾词汇的质量和数量。因此，本文将 PageRank 算法引入到共词网络中用于领域基础词汇的发现，得到词汇 PageRank 值的计算如公式(1)所示。

$$S(v_i)=(1-d)+d\times\sum_{j\in In(v_i)}\frac{S(v_j)}{|Out(v_j)|} \qquad (1)$$

其中，$S(v_i)$、$S(v_j)$ 分别表示关键词 v_i 和 v_j 的 PageRank 值，$In(v_i)$ 表示指向关键词 v_i 的关键词集合，$|Out(v_j)|$ 表示关键词 v_j 指向的关键词的集合，$|Out(v_j)|$ 为集合中元素的个数，d 为阻尼系数，一般设为 0.85。

在关键词构成的引用共词网络中，词汇节点之间的关系强度不是均匀的，因为同一种词语共现对会在不同引文关系中多次出现，而被引次数越多的关键词其重要性越高，因此，将共现词语对之间的权重考虑

进来[26]，构建基于加权的 PageRank 计算公式如公式(2)所示。

$$S'(v_i)=(1-d)+d\times\sum_{j\in In(v_i)}S'(v_j)\times\omega(v_i,v_j) \quad (2)$$

其中，$S'(v_i)$、$S'(v_j)$ 分别表示关键词 v_i 和 v_j 的加权 PageRank 值，$In(v_i)$ 表示指向关键词 v_i 的关键词集合，$\omega(v_i,v_j)$ 代表关键词 v_i 引用 v_j 时 v_j 的 PageRank 值传递给 v_i 的比重，其计算公式如公式(3)所示。

$$\omega(v_i,v_j)=\frac{W(v_i,v_j)}{\sum_k W(v_i,v_k)} \quad (3)$$

其中，$W(v_i,v_j)$ 表示关键词 v_i 和 v_j 之间的共现次数，$\sum_k W(v_i,v_k)$ 表示与关键词 v_i 共现的关键词对的次数总和。

4 实验分析

4.1 数据来源

本研究所使用数据来源于美国计算机学会(Association for Computing Machinery, ACM)的20余万篇英文会议论文，时间跨度为1951年–2012年。经过筛选，包含关键词的文献数量约为11万篇，关键词数量约48万个，论文之间涉及的引用关系约161万条，详细数据统计情况如表1所示。含关键词的文献年度分布情况如图3所示，文献数量基本呈现随年份逐渐增长的趋势，其中80%的文献集中分布在2004年–2011年之间。年度篇均关键词分布情况如图4所示，数量基本稳定在4-6个之间。

数据集中不重复的关键词约为16万个，平均每篇论文涉及的不重复关键词为1.49个，这说明计算机领域的论文共同和重复使用大量词汇作为关键词以对论文进行标识，研究的学科主题相对比较集中，以关键词作为基本知识单元发现领域基础词汇具有较高可行性。

表1 数据集描述性统计

统计量	统计量的值
发文年限(年)	1951-2012
文献数量(篇)	238 309
含关键词的文献年份(年)	1971-2012
含关键词的文献数量(篇)	110 360
关键词数量(个)	479 743
不重复关键词数量(个)	164 146
引用关系数量(条)	1 615 030

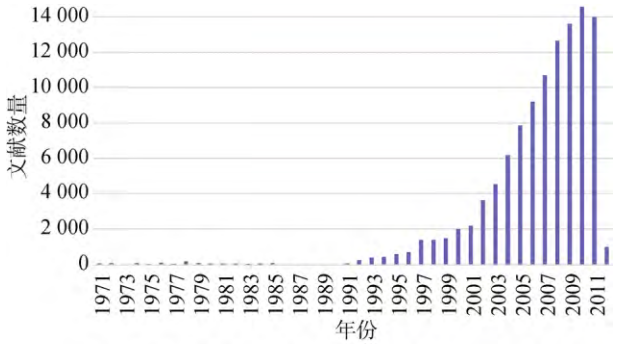

图3 含关键词的文献年度分布情况

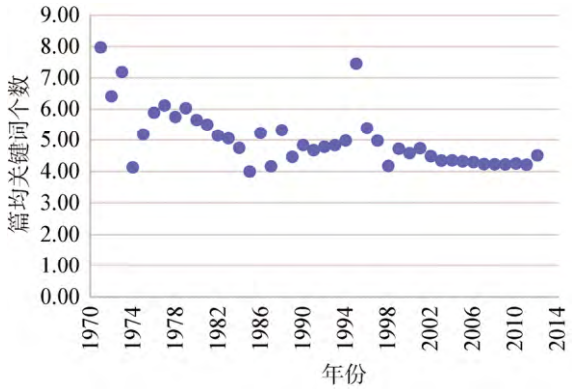

图4 年度篇均关键词分布情况

4.2 实验结果及分析

采用引用共词网络方法对计算机领域的基础词汇进行识别，通过对计算机领域约11万篇会议论文进行关键词抽取和引文关系识别，构建关键词共词网络和引用共词网络，并采用加权和未加权的PageRank算法对领域关键词进行重要性排名，得到领域基础词汇候选集合，将基于词频的方法和基于关键词共词分析的方法作为基准实验进行对比分析。

(1) 基于词频的方法

基于词频的方法通过统计关键词的出现频次，并按照频次高低对关键词进行排名，取排名前50的关键词作为候选领域基础词汇，如表2所示。可以发现，基于词频法得到的候选词汇中，TOP10的词汇中有6个词汇："evaluation"(评价)、"design"(设计)、"simulation"(模拟)、"collaboration"(合作)、"usability"(可用性)和"optimization"(优化)均是语义过于宽泛的词汇，在其他学科中也属于高频词，并不能很好地代表计算机领域的研究基础，只有"security"、"visualization"、"privacy"和"information retrieval"分别表征了计算机领域中的"计算机安全"、"可视化"、"隐

私保护”和“信息检索”4 个基础研究领域，可以称之为基础词汇。将范围进一步扩大到 TOP30 的候选词中，也仅有“clustering”(聚类)、“wireless sensor networks”(无线传感器网络)、“data mining”(数据挖掘)、“sensor networks”(传感器网络)、“interaction design”(交互设计)、“machine learning”(机器学习)、“ubiquitous computing”(普适计算)、“XML”、“virtual reality”(虚拟现实)、“augmented reality”(增强现实)、“social networks”(社交网络)、“Java”等词汇可以表征计算机领域的基础知识。同样地，扩展到 TOP50 样本中，基础词汇和非基础词汇也是交替出现。因此，整体来看基于词频的方法虽然能够发现领域中出现频次较高、研究热度较高的词汇，但这些词汇往往是跨领域的上位词或领域外的不相关词，对特定领域的研究基础表征能力不足，单纯依靠词频的方法在领域基础词汇识别研究中并不理想，尤其是当需要筛选小规模的基础词汇作为研究对象时，通过词频排名提取基础词汇并不能满足实际需求。

表 2　基于词频法的候选领域基础词汇

基础词汇	频次	基础词汇	频次
security	1 138	social networks	563
visualization	1 051	multimedia	556
evaluation	965	ontology	539
design	948	Java	538
privacy	907	routing	501
simulation	850	classification	500
collaboration	841	web services	500
information retrieval	837	mobile computing	476
usability	792	software engineering	471
optimization	768	information visualization	470
clustering	764	children	469
Wireless sensor networks	730	interaction	457
data mining	727	learning	444
education	686	user experience	424
performance	657	communication	422
sensor networks	655	access control	420
interaction design	649	peer-to-peer	416
machine learning	646	FPGA	416
ubiquitous computing	640	accessibility	407
XML	634	modeling	406
scheduling	624	user interface	399
virtual reality	612	mobile devices	397
augmented reality	601	middleware	396
genetic algorithms	572	mobility	395
semantic web	564	recommender systems	392

(2) 基于共词分析的方法

采用 PageRank 算法对关键词共词网络节点重要度进行排名，按照排名高低得到领域基础词汇候选集，截取 PageRank 值排名前 50 的候选词汇如表 3 所示。

表 3　基于共词分析的候选领域基础词汇

PR 排名	基础词汇	PR 排名	基础词汇
1	security	26	simulation
2	design	27	accessibility
3	privacy	28	user experience
4	collaboration	29	virtual reality
5	usability	30	genetic algorithms
6	information retrieval	31	semantic web
7	evaluation	32	clustering
8	education	33	optimization
9	XML	34	interaction
10	visualization	35	augmented reality
11	interaction design	36	mobile
12	ubiquitous computing	37	wireless sensor networks
13	machine learning	38	pedagogy
14	children	39	training
15	Java	40	embedded systems
16	data mining	41	social media
17	multimedia	42	verification
18	scheduling	43	classification
19	routing	44	FPGA
20	software engineering	45	low power
21	sensor networks	46	mobile computing
22	performance	47	UML
23	peer-to-peer	48	access control
24	web services	49	animation
25	social networks	50	wireless

可以发现，基于共词分析得到的候选词汇 TOP10 中识别出了“security”、“privacy”、“information retrieval”、“XML”和“visualization”这 5 个领域基础词汇，较词频法多一个。表 3 的 TOP30 候选词汇中，“software engineering”(软件工程)、“peer-to-peer”(对等网络)等基础词汇在表 2 中排在 TOP30 之后，而在表 2 的 TOP30 中出现的“wireless sensor networks”、“clustering”、“augmented reality”等基础词汇在共词分析结果中排在 TOP30 之后。在整体 TOP50 样本中，两种方法所得到的候选词汇重合比例为 78%，即有 39 个候选词汇同时归属两种方法，区别在于部分词汇的位

列排序发生变化。因此，整体来看基于共词网络的PageRank指标排名结果较基于词频的方法略好，能够通过PageRank值将一些频次不高但比较重要的节点排在靠前的位置。但同时可以发现两种方法所得到的候选词汇出现大量重复，说明共词网络指标与词频依然线性相关，所得到的候选词汇对领域研究基础的表征能力仍然有限。

(3) 基于引用共词网络的方法

采用加权和未加权的PageRank算法对关键词引用共词网络节点重要度进行排名，按照排名高低得到领域基础词汇候选集，加权和未加权的PageRank值排名前50的候选词汇如表4所示。

表4 基于引用共词网络的候选领域基础词汇

未加权PR排名	基础词汇	未加权PR排名	基础词汇	加权PR排名	基础词汇	加权PR排名	基础词汇
1	non-photorealistic rendering	26	visualization	1	ubiquitous computing	26	recommender systems
2	ubiquitous computing	27	aspect-oriented programming	2	sensor networks	27	image-based rendering
3	sensor networks	28	texture synthesis	3	non-photorealistic rendering	28	visualization
4	augmented reality	29	CS1	4	augmented reality	29	texture synthesis
5	CSCW	30	security	5	CSCW	30	virtual reality
6	social networks	31	virtual reality	6	children	31	security
7	privacy	32	recommender systems	7	social networks	32	interaction design
8	information retrieval	33	volume rendering	8	privacy	33	volume rendering
9	children	34	text entry	9	information retrieval	34	concurrency
10	awareness	35	interaction design	10	awareness	35	web characterization
11	information visualization	36	eye tracking	11	collaborative filtering	36	global illumination
12	transactional memory	37	speech and pen input	12	Java	37	eye tracking
13	Java	38	global illumination	13	transactional memory	38	interaction techniques
14	collaborative filtering	39	clustering	14	web search	39	text entry
15	web search	40	interaction techniques	15	information visualization	40	participatory design
16	routing	41	usability	16	wireless sensor networks	41	input devices
17	wireless sensor networks	42	concurrency	17	routing	42	peer-to-peer
18	design	43	participatory design	18	ethnography	43	usability
19	ethnography	44	geometric modeling	19	design	44	geometric modeling
20	texture mapping	45	input devices	20	CS1	45	garbage collection
21	web characterization	46	XML	21	animation	46	clustering
22	image-based rendering	47	garbage collection	22	texture mapping	47	motion capture
23	animation	48	Fitts' law	23	evaluation	48	Fitts' law
24	evaluation	49	motion capture	24	collaboration	49	speech and pen input
25	collaboration	50	human-robot interaction	25	aspect-oriented programming	50	XML

可以发现，基于引用共词网络的加权和未加权的PageRank算法得到的候选词汇TOP10完全相同，区别仅在于部分词汇的排序不同，其中“non-photorealistic rendering”(非真实感绘制技术)、“ubiquitous computing”、“sensor networks”、“augmented reality”、“CSCW”(计算机支持协同工作)、“social networks”、“privacy”和“information retrieval”均表征计算机领域的基础研究方向或基础技术，可以界定为领域基础词汇，TOP10中只有“awareness”(意识)和“children”(儿童)两个词汇不属于领域基础词汇。将范围进一步扩大到TOP30的候选词中，也只有“design”、“ethnography”(民族志)、“evaluation”、“collaboration”等少数词汇不能作为领域基础词汇。在TOP50样本中，可以看到候选词汇中基础词汇的比例高于非基础词汇，而重要的基础词汇排名均比较靠前。因此，整体来看基于引用共词网络分析的方法要比词频法和共词分析法效果好，能

够发现频次不高但在网络中处于核心节点的一些较为重要的知识单元，并且排名靠前的词汇大部分均为基础词汇，说明本文所提引用共词网络分析方法有效可行，在需要提取小范围基础词汇为研究对象的任务中能够发挥出较大优势。

对比表 4 中加权和未加权的排名结果，TOP50 候选词汇中重复词汇达 49 个，只有“human-robot interaction”和“peer-to-peer”分别单独属于未加权和加权的 TOP50 候选词汇，且这两个词汇均是领域基础词汇，由此说明加权和未加权的识别结果整体效果相差不大，区别仅在于部分基础词汇的排名顺序不一样。

4.3　基于盲选实验的量化评估

上述分析从定性角度对实验结果进行了探讨，为进一步对上述方法的实验结果进行量化评估，本文参考文献[27]设计了一种基于盲选实验的量化评估方法。由于加权和未加权的引用共词网络效果相差不大，在盲选实验中仅以词频法、共词分析法和未加权的引用共词分析法三种实验结果为对象进行评估。具体评估过程为：将三种实验得到的领域基础词汇候选集进行混合，并打乱次序，得到不重复的 87 个候选词，邀请实验者从这些候选词中选出能够表征计算机领域的基础词汇。受邀者为从事计算机领域相关研究且具备多年研究经验的科研人员，共计三人。

统计每位实验者选择的词汇中，分别归属三种方法所包含的候选基础词汇的数量和比例。由于候选词集中各种方法提供的候选基础词汇数量相等，因此可以认为实验者选出的词来自哪个方法更多，则该方法效果更好。盲选实验结果如表 5 所示，方法 1 至方法 3 分别对应基于词频的方法、基于共词分析的方法和基于引用共词分析的方法。

表 5　盲选实验结果

实验者编号	选中词数	与方法 1 重合情况		与方法 2 重合情况		与方法 3 重合情况	
		数量	比例	数量	比例	数量	比例
1	70	40	57.14%	39	55.71%	40	57.14%
2	48	25	52.08%	24	50.00%	36	75.00%
3	50	28	56.00%	30	60.00%	34	68.00%
平均值	56	31	55.07%	30	55.24%	31.75	66.71%

可以看出，通过盲选实验得到的基础词汇中，与传统词频方法和共词分析方法重合的比例相差不大，而与引用共词网络方法重合的比例远高于前两者，其平均准确率达 66.71%，在一定程度上说明引用共词网络方法能更好地拟合专家人工筛选的结果。在实际应用中往往需要筛选的仅是一小部分基础词汇，因此进一步采用 P@N(N=10, 20, 30, 40, 50)指标来观察三种方法在第 N 个位置上的正确率，结果如表 6 所示。

表 6　盲选实验正确率

方法	P@10	P@20	P@30	P@40	P@50
方法 1	0.37	0.60	0.68	0.63	0.62
方法 2	0.43	0.62	0.60	0.62	0.62
方法 3	0.73	0.75	0.69	0.73	0.73

可以看出，基于引用共词网络的方法在各个位置上的正确率均明显高于词频法和共词分析法在相应位置上的准确度，平均准确率达 72.6%，其中 P@10 和 P@20 指标上分别达到 73%和 75%，即前 10 个候选词中有 7 个词属于基础词汇，前 20 个候选词中有 15 个词属于基础词汇，达到较好的识别结果。共词分析法在 P@10 和 P@20 指标上稍高于词频法，而在 P@30、P@40 和 P@50 指标上二者相差不大，说明共词分析法在提取小部分基础词汇的任务中表现优于词频法，而当返回结果样本数量较大时，两种方法的差距不是很明显。整体来看，本文所提基于引用共词网络的方法在识别领域基础词汇时，能够通过 PageRank 排名更好地发现重要性高的基础词汇，避免了依靠词频或共词法所得结果中大量语义过于宽泛的词汇排名靠前的情况，在发现领域基础词汇任务中具有较好的表现和较高的应用价值。

5　结　语

科研领域基础词汇对把握学科结构和知识脉络具有重要意义，本文将引文网络引入到共词分析中，通过关键词之间的引用关系构建引用共词网络，采用 PageRank 算法对候选词汇重要度进行排名。从定性和

定量的角度对结果进行评价，融合引文网络的共词分析方法，较传统的词频法和共词分析法，能更好地拟合专家人工筛选结果，盲选实验在各个位置上的准确率均高于后两者。综合说明本文所提基于引用共词网络方法能更有效地综合关键词的频次和重要性，既能发现频次较低但重要性高的基础词汇，也能过滤掉频次较高但语义过于宽泛的非基础词汇。本研究仅以计算机领域为例进行实证研究，今后将选择更多的学科领域对本文方法的可行性和效果进行验证，以考察其在不同学科的适应性。

参考文献:

[1] Courtial J P. Comments on Leydesdorff's Article[J]. Journal of the American Society for Information Science, 1998, 49(1): 98.

[2] Su H N, Lee P C. Mapping Knowledge Structure by Keyword Co-occurrence: A First Look at Journal Papers in Technology Foresight[J]. Scientometrics, 2010, 85(1): 65-79.

[3] Hu J M, Zhang Y. Research Patterns and Trends of Recommendation System in China Using Co-Word Analysis[J]. Information Processing and Management, 2015, 51(4): 329-339.

[4] Sun Y W, Zhai Y. Mapping the Knowledge Domain and the Theme Evolution of Appropriability Research Between 1986 and 2016: A Scientometric Review[J]. Scientometrics, 2018, 116(1): 203-230.

[5] Khasseh A A, Soheili F, Moghaddam H S, et al. Intellectual Structure of Knowledge in iMetrics: A Co-Word Analysis[J]. Information Processing & Management, 2017, 53(3): 705-720.

[6] Ravikumar S, Agrahari A, Singh S N. Mapping the Intellectual Structure of Scientometrics: A Co-Word Analysis of the Journal Scientometrics (2005-2010) [J]. Scientometrics, 2015, 102(1): 929-955.

[7] Soriano A S, Álvarez C L, Valdés R M T. Bibliometric Analysis to Identify an Emerging Research Area: Public Relations Intelligence — A Challenge to Strengthen Technological Observatories in the Network Society[J]. Scientometrics, 2018, 115(3): 1591-1641.

[8] 胡昌平，陈果. 科技论文关键词特征及其对共词分析的影响[J]. 情报学报，2014, 33(1): 23-32. (Hu Changping, Chen Guo. Characteristics of Keywords in Scientific Papers and Their Impact on Co-word Analysis[J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(1): 23-32.)

[9] 李树青. 基于引文关键词加权共现技术的图情学科领域本体自动构建方法研究[J]. 情报学报，2012, 31(4): 371-380. (Li Shuqing. Research on Automatic Construction of Domain Ontology in Library and Information Science Based on Weighted Co-occurrence of Citation Keywords[J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(4): 371-380.)

[10] Yan B N, Lee T S, Lee T P. Mapping the Intellectual Structure of the Internet of Things (IoT) Field (2000-2014): A Co-Word Analysis[J]. Scientometrics, 2015,105(2): 1285-1300.

[11] Wang Z S, Zhao H, Wang Y. Social Networks in Marketing Research 2001-2014: A Co-Word Analysis[J]. Scientometrics, 2015, 105(1): 65-82.

[12] Donohue J C. Understanding Scientific Literature: A Bibliographic Approach[M]. Cambridge: The MIT Press, 1973: 101.

[13] Booth A D. A "Law" of Occurrences for Words of Low Frequency[J]. Information and Control, 1967, 10(4): 386-393.

[14] Yang Y, Wu M, Cui L. Integration of Three Visualization Methods Based on Co-Word Analysis[J]. Scientometrics, 2011, 90(2): 659-673.

[15] Yan B N, Lee T S, Lee T P. Analysis of Research Papers on E-Commerce (2000-2013): Based on a Text Mining Approach[J]. Scientometrics, 2015, 105(1): 403-417.

[16] 李纲，巴志超. 共词分析过程中的若干问题研究[J]. 中国图书馆学报，2017, 43(4): 93-113. (Li Gang, Ba Zhichao. Co-word Analysis: Limitations and Solutions[J]. Journal of Library Science in China, 2017, 43(4): 93-113.)

[17] Choi J, Yi S, Lee K C. Analysis of Keyword Networks in MIS Research and Implications for Predicting Knowledge Evolution[J]. Information & Management, 2011, 48(8): 371-381.

[18] Zhu W, Guan J. A Bibliometric Study of Service Innovation Research: Based on Complex Network Analysis[J]. Scientometrics, 2013, 94(3): 1195-1216.

[19] Ocholla D N, Onyancha O B, Britz J. Can Information Ethics Be Conceptualized by Using the Core/Periphery Model? [J]. Journal of Informetrics, 2010, 4(4): 492-502.

[20] Liu J X, Zheng C H, Xu Y. Extracting Plants Core Genes Responding to Abiotic Stresses by Penalized Matrix Decomposition[J]. Computers in Biology & Medicine, 2012, 42(5): 582-589.

[21] Ding Y, Song M, Han J, et al. Entitymetrics: Measuring the Impact of Entities[J]. PLoS One, 2013, 8(8): e71416.

[22] Song M, Han N G, Kim Y H, et al. Discovering Implicit Entity Relation with the Gene-Citation-Gene Network[J]. PLoS One, 2013, 8(12): e84639.

[23] 吴清强，赵亚娟．基于论文属性的加权共词模型探讨[J]. 情报学报, 2008, 27(2): 89-92. (Wu Qingqiang, Zhao Yajuan. Research in the Weighted Co-word Analysis Based on the Attributes of Articles[J]. Journal of the China Society for Scientific and Technical Information, 2008, 27(2): 89-92.)
[24] 葛菲，谭宗颖．基于文献计量学的科学结构及其演化的研究方法述评[J]. 情报杂志, 2012, 31(12): 34-39. (Ge Fei, Tan Zongying. Review of Science Structure and Evolution of Bibliometric Methods[J]. Journal of Intelligence, 2012, 31(12): 34-39.)
[25] Brin S, Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine[C]// Proceedings of the 7th International Conference on World Wide Web. 1998: 107-117.
[26] Zhao W Y, Mao J, Lu K. Ranking Themes on Co-Word Networks: Exploring the Relationships Among Different Metrics[J]. Information Processing & Management, 2018, 54(2): 203-218.
[27] 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. (Chen Guo, Xiao Lu, Zhao Xueqin. A Keyword Selection Method Based on the Combination of Popularity and Domain Relevancy of Keywords: A Holistic Perspective[J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(9): 959-968.)

作者贡献声明:

程齐凯：提出研究思路，进行实验，论文撰写与修改；
王佳敏：参与实验，数据处理与分析，论文撰写与修改；
陆伟：设计研究方案，论文撰写与修改。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: wangjm@whu.edu.cn。
[1] 王佳敏. acm_article.sql. 文献数据.
[2] 王佳敏. acm_article_reference.sql. 引文数据.
[3] 王佳敏. refwords_pr_score.xlsx. 实验结果数据.

收稿日期: 2018-10-19
收修改稿日期: 2018-11-16

Discovering Domain Vocabularies Based on Citation Co-word Network

Cheng Qikai　Wang Jiamin　Lu Wei
(School of Information Management, Wuhan University, Wuhan 430072, China)
(Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan 430072, China)

Abstract: [Objective] This paper identifies basic vocabularies of a specific domain from academic papers, aiming to grasp the knowledge structure and development context. **[Methods]** We combined the citation network and the co-word analysis to construct a citation co-word network. Then, we used the PageRank algorithm to evaluate the importance of the candidate words. We examined the proposed method with 110,360 articles in computer science. **[Results]** Our new method was compared with the word frequency method and co-word analysis qualitatively and quantitatively. We found that the proposed method performed well, and the average precision of a blind selection experiment reached 72.6%. **[Limitations]** The proposed method was only examined with computer science articles. **[Conclusions]** The new strategies could improve the performance of basic vocabulary discovery in one specific domain.

Keywords: Basic Vocabulary　Citation Co-word Network　PageRank　Word Frequency　Co-word Analysis