引用语境中的线索词分析*
——以 *Journal of Informetrics* 为例

胡志刚[1,2] 孙太安[1,2] 王贤文[1,2]
[1] 大连理工大学科学学与科技管理研究所 大连 116024 [2] 大连理工大学 WISE 实验室 大连 116024

摘要:[**目的/意义**]引用语境是科学论文中包含引用的句子,是对引文的描述性或评价性文字。通过提取和分析引用语境中的线索词,可以了解引用语境的一般特征。[**方法/过程**]以 *Journal of Informetrics*(JOI)期刊为例,选取人称代词、行为动词和连接词三类常用的线索词,分别计算它们在引用语境中的频次、占比和排序。通过比较各类线索词在引用语境和非引用语境中的存在度,识别引用语境中的常用句型和论证模式。[**结果/结论**]在 JOI 期刊中,引用语境主要表现出如下特点:侧重于第一人称和第三人称论述,既展现他人的工作,也展现作者的研究;偏重于研究方法类引用,常用的行为动词为"use""base"和"study";强调通过转折和列举等逻辑方式进行论证,最常用的连接词位"also"和"but"。分析引用语境中的线索词,对于更好地理解科学论文中的引用的功能和动机具有重要的价值和意义。

关键词:引用语境 线索词 词频分析 引用
分类号:G250
DOI:10.13266/j.issn.0252-3116.2017.23.003

在传统的引文分析方法中,关注的主要是文章层面的引用关系。然而,在实际的引用行为中,引用是在句子层面发生的,引用的基本单位不是文献,而是文献中的句子,即引用语境。引用语境是施引文献对被引文献的描述性或评论性文字。基于引用语境,可以分析和识别施引作者在引用的情感、评价、功能和动机等。

引用语境本质上是一条句子。从语义上来看,一个句子是由内容词和线索词构成的。内容词(content phrase)主要由专有名词组成,负责告诉人们是"什么东西",而线索词(cue phrase)主要由内容词之外的动词、副词、连词、代词等组成,负责告诉人们是"什么事"。在对引用语境进行功能分析时,不同的研究领域所用的内容词可能完全不同,因此这类词不能作为跨学科领域的通用分类特征;而线索词由于与研究领域和研究主题无关,只与语境的语义功能和情感有关,更适合作为一般意义上的引用语境分类标准。

在本研究中,笔者将以 *Journal of Informetrics*(JOI)期刊论文中的引用语境为例进行实证分析。通过统计常用的三种线索词(即人称代词、行为动词和连接词)在引用语境和非引用语境中的频次和占比,了解引用语境中最常使用的线索词及其类型,分析引用语境有别于非引用语境的一般特征。这一实证研究,对于理解引用语境的内涵,洞悉引用功能的类别,进而构建引用语境的分类模型具有一定的借鉴价值。

1 文献综述

引用语境中包含了丰富的信息和内涵,因此通过人工判读或机器学习的方法,可以有效地对引用的具体动机和类型进行识别和划分。其中提取并分析线索词(cue word 或 cue phrase),是了解引用的规范和文法的一种重要手段。

M. Bertin 和 I. Atanassova 等人曾以 PLOS 期刊为例分析了引用语境中最常用的动词。例如,在引言(introduction)部分,最常用的动词为"show""use""include""suggest"等,而在方法(methods)部分,最常用的

* 本文系国家自然基金青年项目"开放获取背景下的全文引文分析方法与应用研究"(项目编号:71503031)和国家自然科学基金面上项目"地理与网络二维空间及其交互影响视角下的科学论文扩散研究"(项目编号:71673038)研究成果之一。
作者简介:胡志刚(ORCID:0000-0003-1835-4264),讲师,硕士生导师;孙太安(ORCID:0000-0001-6769-2652),硕士研究生;王贤文(ORCID:0000-0002-7236-9267),教授,博士生导师,通讯作者,E-mail:xianwenwang@dlut.edu.cn。
收稿日期:2017-06-16 修回日期:2017-08-28 本文起止页码:25-33 本文责任编辑:王善军

动词为“use”“perform”“follow”“obtain”等[1]；对于每一个动词，他们还列举了最常采用的 3 - g 短语，比如“was used to”“ we used”“we used the”等[2]。

R. Mercer 和 C. Di Marco 则对引用语境中的边界线索词(hedging cues)进行了词频分析[3]。边界线索词最早是由 K. Hyland[4] 提出，他指出边界线索词是对事实进行陈述的行为动词以及表达确定程度的副词，如“assume”“unlikely”等。研究结果表明，引用语境中的边界线索词含量高于非引用语境。

利用某些特殊的线索词，作为判断引用行为和动机的标志，在前人的研究中并不鲜见[5-7]。1999 年，日本学者 H. Nanba 等利用引用位置和线索词，构建了一种引用自动分类系统[8]，将引用分成三类：Type B (base on)、Type C(compare to)和 Type O(others)。他们发现，从线索词来看，Type B 和 Type C 分别对应于不同的线索词，Type B 对应的线索词包括“based on”“used in”“we can use”等，Type C 对应的线索词包括“but”“however”“although”等。文中，他们共列举了 Type B 所对应的线索词 84 个，Type C 所对应的线索词 76 个。

剑桥大学的 S. Teufel 很早就开始关注引用语境分析[9]，是活跃在该研究领域的重要专家之一。早在 1999 年，S. Teufel 在博士论文中就曾详细研究过引用语境的分类规则[10]。这一分类规则长达 25 页，其中包括大约 150 条规则，包含例子、决策树和特别解释说明等。作为一种基于线索词对引用语境进行分类的方法，她的分类规则包含 1 762 个线索词、185 个识别符(POS-based recognizer)以及 20 个行为动词。在人工识别的过程中，又有 75 类共 892 个线索词被添加进去。这应该是迄今为止所有规则中最为复杂的一个。

此外，加拿大滑铁卢大学 R. Radoulov [11] 和美国威斯康星大学密尔沃基分校计算机系的 H. Yu 等人[12-14]，也借助线索词等文本特征，利用机器学习的算法构建了引用的自动分类系统，并展现出了较好的分类效果。

总的来说，由于开放获取运动的兴起，全文数据正变得越来越易得，提取并分析正文中的引用语境正成为引文分析的新趋势[15-18]。在单词和短语层面上，对于引用语境中的线索词或线索短语进行提取和分析，是引用语境分析的基础[19]。如何提取并分析引用语境中的这一细粒度信息，对于设计并构建引用语境的分析框架，具有非常重要的基础性意义。

2 数据与方法

2.1 数据来源

笔者所使用的样本是 JOI 期刊自 2007 年创刊至 2013 年的 350 篇论文。JOI 是信息计量学领域的一种 SSCI 国际重要期刊，由世界著名的科技出版商爱思唯尔运营。在爱思唯尔旗下，运营着包括 JOI 在内的 2 500多种学术期刊，所有这些期刊可以通过 ScienceDirect 和 ConSyn 两个数据库进行检索或全文下载。其中，ScienceDirect 提供 PDF 格式的全文下载；Elsevier ConSyn 提供的 XML 格式的全文下载。在本研究中，分析采用的 XML 格式文本正是在 Elsevier ConSyn 中批量下载得到的。

采取 XML 格式的文本数据进行分析，是因为相对于 PDF 格式文本，引用语境的提取更为容易和准确。图 1 是 XML 格式文本的一个片段，其中包含了 4 个引用语境。

```
The time-evolution of the h-index, called h-index sequence
in the following, has received much attention in recent
years as a way to quantify the progress of scientists'
careers (<ce:cross-refs refid="bib0025 bib0035">Egghe,
2010; Hirsch, 2005</ce:cross-refs>). <ce:cross-ref refid="
bib0040">Liang (2006)</ce:cross-ref> proposed the concept
of the h-index sequence.
<ce:cross-ref refid="bib0005">Burrell (2007a)</ce:cross-ref>
 used a stochastic model to investigate the influence of
parameters like publication and citation rates as well as
the researcher's career length on the h-index's time
evolution. Moreover, he studied some properties of the
h-index sequences of 19 leading physicists included in the
dataset used in the original Hirsch-index paper (
<ce:cross-ref refid="bib0035">Hirsch, 2005</ce:cross-ref>).
```

图 1 XML 格式论文文本中的引用语境样式

引用语境的提取的规则如下：

(1)以句号(.)/问号(?)作为句子的分隔符，切分出句子，并写入 sentence 数据表中。由于句号除作为句子结束符外，还可能出现在人名(如“Iijima S.”)、数字(如“0.123”)或其他缩写中(如“etc.”“e. g.”“Fig. 1”)中。对于这类情况，主要采取词表替换(主要针对缩写中的句号)和正则表达式替换(主要针对人名和数字中的句号)相结合的方法，将干扰句号首先替换掉其他特殊符号，切分之后再进行恢复。

(2)识别每个句子中是否包含了引用代码(如 <ce:cross-ref refid = “bib0040” > Liang (2006) </ce:cross-ref >)。如果句子中包含引用代码，则标记为引用语境。最终，笔者在 JOI 期刊的 350 篇论文中共提取了 67 612 条句子，其中含有引用的句子即引用语境 10 408条，占全部句子的 15.4%，其余非引用语境 57 204条，占全部句子的 84.6%。

2.2 引用语境中的线索词

在本研究中，笔者选取了以下三种类型的线索词

进行分析：人称代词、行为动词和连接词。对这些线索词的提取，主要采用正则表达式查询的方法。其中，对于行为动词还利用通配符的方法进行了模糊匹配，以识别其各种时态和语态的不同的形式。

线索词是和内容词相对应的，它们一个表示的是“意义”，一个表示的是“对象”。图 2 中形象了给出了内容词和线索词的区别：在隐去了线索词之后，可以知道在说“什么东西”（h-index、h-values、productivity、citation practices 等），而不知道在说“什么事”，像是传统的词频分析，只保留了信息内核和语义信息；而在隐去了内容词之后，可以知道在说“什么事”（however、several limitations of the something 等），但不知道在说“什么东西”，只能识别引用时的状态情感和语用信息。

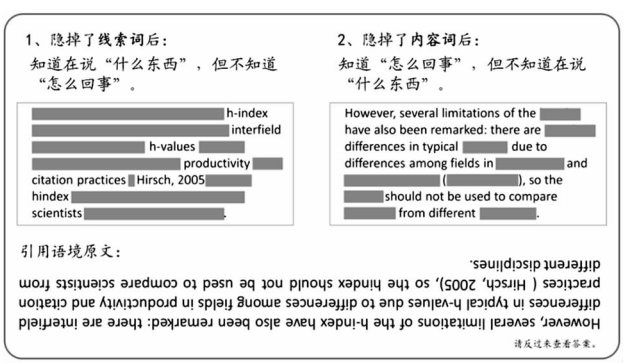


图 2　引用语境中的内容词和线索词样例

人称代词表示的是引用中的主宾关系，即主体和客体。人称代词按照人称可以分为第一人称代词（如 I、We）、第二人称代词（如 You）和第三人称代词（如 He、She 和 They）。按照人称格可以分为主格代词（如 I）、宾格代词（如 me）和物主代词，物主代词又分为形容词性物主代词（如 my）和名词性物主代词（如 mine）。科学论文中的代词的例子如下：

- ***We*** *focused mainly on what information the engineers needed and used in* ***their*** *work and how* ***they*** *obtained and managed this information.*
- ***Our*** *goal is to analyze not the publication and citation counts but the group's connectedness to the research mainstream, both statically and over time.*

行为动词表示的是引用语境中的谓语结构。对于行为动词的提取，笔者采用了托伊菲尔在博士论文[10]中所采用的行为动词列表，包括 18 类 364 个行为动词，这 18 类的行为动词参见表 1。

根据这一词表，在引用语境中提取得到的行为动词的例子如下：

- *Details of our proposal are* ***given*** *in Section 2, where we also* ***show*** *how this can be* ***determined*** *under the stochastic model* ***used*** *by Burrell (2007b).*

表 1　在引用语境中所提取的行为动词的类别

行为动词	例子
Affect 影响	afford | believe | decide | feel | hope | imagine | regard | trust | think…
Argumentation 论证	agree | accept | advocate | argue | claim | conclude | comment…
Aware 清楚	be unaware | be familiar with | be aware | be not aware…
BetterSolution 更佳方案	boost | enhance | defeat | improve | go beyond | perform better…
Change 改变	adapt | adjust | augment | combine | change | decrease | elaborate…
Comparison 比较	compare | compete | evaluate | test…
Continue 继续	adopt | agree with | base | be based on | be derived from | be originated in
Contrast 对比	be different from | be distinct from | confict | contrast | clash…
FutureInterest 未来兴趣	plan on | plan to | expect to | intend to…
Interest 兴趣	aim | ask | address | attempt | be concerned | be interested | be motivated
Need 需要	be dependent on | be reliant on | depend on | lack | need | necessitate
Presentation 展示	describe | discuss | give | introduce | note | notice | point out | present
Problem 问题	abound | aggravate | arise | be cursed | be incapable of | be forced to
Research 研究	apply | analyze | analyse | build | calculate | categorize | categorise…
Similar 类似	bear comparison | be analogous to | be alike | be related to…
Solution 方案	accomplish | account for | achieve | apply to | answer | alleviate | allow for
Text Structure 文章结构	begin by | illustrate | conclude by | organize | organise | outline | return to
Use 使用	apply | employ | use | make use | utilize…

- The Journal Citation Report promulgated by ISIin2008 ***used*** *a new journal evaluationindex, Eigenfactor, the calculation of which* ***is based on*** *the PageRank algorithm, but eliminates selfcitationsinjournals.*

连接词表示的科学论文中的逻辑关系。对于连接词的提取，笔者采用了 A. Knott 在其硕士论文[20]中所采用的线索词列表，包括 10 类 123 个连接词，如表 2 所示：

表 2 在引用语境中所提取的连接词的类别

连接词	例子
AdditionalInformation 补充	even|on the contrary|in fact|in actual fact|actually|as a matter of fact
Cause 因果	because|,for|insofar as|to the extent that|considering that|given that
Digression 离题	incidentally|by the way
Hypothetical 假设	if ever|if only| in that case|as long as|on condition that|supposing that
Negative Polarity 转折	but| whereas|instead|however|though|nevertheless|yet|otherwise
Restatement 重述	to summarise|to sum up|summing up|to recap|or rather|at least
Result 结果	therefore|consequently|to this end|as a result|for example|thus|hence
Sequence 顺序	besides|in addition|first|firstly| to begin with|first of all|furthermore
Similarity 类同	just as|the way|also|too|as well
Temporal 时间	while|meanwhile|beforehand|previously|ever since|since

根据这一词表,提取得到的表示转折关系的引用语境例子如下:

- ***However***, *this algorithm is very slow and has been outperformed by more recent methods* (*Newman*, *2006*).
- *As noted above* (*in footnote 4*), *algorithms for weighted betweenness centrality are available in the literature* (*Brandes*, *2001*), ***but*** *have not* ***yet*** *been implemented for matrices of this size* (*Brandes*, *personal communication*, *25 February 2010*).

3 结果

在 JOI 期刊论文的引用语境中,分别提取其中的人称代词、行为动词和连接词,统计它们的词频及名次。同时,列出它们在非引用语境中的词频及名次进行对比,以展现引用语境有别于非引用语境的一般特征。

3.1 人称代词

表 3 列出了在引用语境和非引用语境中出现的人称代词列表。在引用语境中出现频次最高的两个人称代词分别是“we”和“their”,分别出现在 841 个和 548 个引用语境中,占全部引用语境的 8.08% 和 5.27%。这反映了引用语境的二相性:属性上是“他人的”工作,功能上则是为“我们”所用;不仅强调是“他人”的工作,还强调为“我们”所用。此外,这两个高频词还反映了科学研究中的继承关系,即“我们”的工作一般是基于“他人的”的成果而开展的。

表 3 引用语境和非引用语境中的人称代词列表

排序	人称代词	频次(引用语境)	词频/排序(非引用语境)	排序	人称代词	频次(引用语境)	词频/排序(非引用语境)
1	we	841	7 656/1	**12**	**himself**	**12**	**15/16**
2	**their**	**548**	**2 672/3**	12	my	12	88/12
3	I	410	3 405/2	14	she	9	68/13
4	they	359	1 733/4	**15**	**ours**	**4**	**8/20**
5	our	213	1 689/5	**15**	**ourselves**	**4**	**13/17**
6	**his**	**117**	**413/8**	17	him	3	17/15
7	**he**	**111**	**252/9**	17	me	3	25/14
8	us	110	664/7	19	herself	1	9/18
9	them	96	669/6	19	mine	1	9/18
10	her	48	196/10	**19**	**myself**	**1**	**5/21**
11	themselves	22	105/11				

频次从高到低排列,其中,在引用语境中的位次高于其在非引用语境中的位次的人称代词用粗体表示。可以发现,“their”“his”“he”等第三人称代词在引用语境中的排序都高于其在非引用语境中的排序;而与此相反,“I”“us”等第一人称代词在引用语境中的排序低于其在非引用语境中的排序。这说明在引用语境中,第三人称的相对存在度更高,第一人称的相对存在度较低。

此外,结果还表明,在引用语境中,主格代词和形容词性物主代词的频次远超宾格代词,例如 their 出现了 548 次,they 出现了 359 次,而 them 仅出现了 96 次;

复数形式的人称代词超过单数形式的代词，如 we 出现了 841 次，而 I 出现了 410 次；男性人称代词超过女性人称代词，如 he 出现了 111 次，而 she 仅出现 9 次。

3.2 行为动词

表 4 列出了在 JOI 期刊论文中行为动词在引用语境中和非引用语境中分别出现的频次和位次。在引用语境中，出现次数最高的行为动词为“use”，出现了 1 029次，占全部引用语境的 9.88%，远超过排在第二的行为动词。在非引用语境中，“use”一词也以3 575 次的出现次数位列首位。

在引用语境中，排在“use”之后的两个行为动词分别是“base”和“study”，分别出现了 558 次和 504 次，高于其在非引用语境中的位次（第 4 位和第 6 位），这表明“base”和“study”更容易出现在引用语境中。在表 4 中，这种在引用语境中的排序高于其在非引用语境中的排序的行为动词，用粗体进行了标记。除了“base”和“study”之外，还包括“model”“propose”“review”“define”“suggest”等。

表 4 引用语境和非引用语境中的高频行为动词列表

排序	行为动词	频次（引用语境）	词频/排序（非引用语境）	排序	行为动词	词次（引用语境）	词频/排序（非引用语境）
1	use	1029	3 575/1	**21**	**describe**	**152**	**602/27**
2	**base**	**558**	**1 891/4**	22	obtain	151	1 087/14
3	**study**	**504**	**1 807/6**	**23**	**discuss**	**148**	**467/33**
4	measure	503	1 992/3	24	note	147	849/18
5	**model**	**411**	**1 689/7**	**24**	**test**	**147**	**719/25**
6	show	397	2 964/2	26	calculate	138	900/17
7	**propose**	**393**	**742/23**	**27**	**analyze**	**126**	**490/32**
8	give	296	1 875/5	28	expect	114	822/22
9	consider	263	1 516/8	29	change	111	835/20
9	**review**	**263**	**971/16**	30	observe	106	720/24
11	**define**	**262**	**1 165/12**	31	determine	104	570/30
12	provide	255	1 220/10	**31**	**examine**	**104**	**340/37**
13	count	251	1 109/13	31	need	104	677/26
14	**suggest**	**219**	**579/29**	**34**	**follow**	**102**	**461/34**
15	present	218	1 362/9	**35**	**based on**	**98**	**337/38**
16	compare	211	1 003/15	36	indicate	94	830/21
17	**structure**	**192**	**838/19**	37	analyze	86	525/31
18	**introduce**	**184**	**333/39**	**37**	**investigate**	**86**	**312/40**
19	increase	172	1 167/11	37	state	86	405/36
20	**report**	**169**	**595/28**	40	allow	81	411/35

高频行为动词反映了引用语境的主要功能和动机，大致可以分为如下几类：①阐述方法，如“use”“base”“measure”“model”等；②揭示问题，如“study”“propose”“introduce”“discuss”“examine”等；③描述结论，如“show”“suggest”“report”“describe”等。显然，在 JOI 期刊论文的引用语境中，“阐述方法”的动词出现次数较多，表明了引用语境中注重方法的特点。

根据表 1 中的行为动词的类别划分，笔者统计了不同功能类型的动词在引用语境中的数量，如表 5 所示。排在前五位的行为动词分别是“研究”（research）、“方案”（solution）、“兴趣”（interest）、“展示”（presentation）和“使用”（use），分别占全部引用语境数量的 9.85% －3.65%。这表明了在科学论文中，引用语境主要是围绕与“研究”有关的内容进行讨论，如研究问题、研究方案、研究方向和研究方法等。

表 5 基于行为动词的句式类型数量及占比

序号	句子功能类型	引用语境	占比	序号	非引用语境	占比
1	Research 研究	1 025	9.85%	1	5 266	9.21%
2	Solution 方案	878	8.44%	2	4 279	7.48%
3	**Interest** 兴趣	**775**	**7.45%**	**4**	**3 665**	**6.41%**
4	Presentation 展示	594	5.71%	3	3 692	6.45%
5	**Use** 使用	**378**	**3.63%**	**7**	**1 437**	**2.51%**
6	Text Structure 文章结构	368	3.54%	6	1 529	2.67%
7	Change 改变	238	2.29%	5	1 590	2.78%
8	Comparison 比较	215	2.07%	8	1 030	1.80%
9	Argumentation 论证	143	1.37%	9	730	1.28%
10	Need 需要	117	1.12%	10	725	1.27%
11	**Problem** 问题	**74**	**0.71%**	**13**	**286**	**0.50%**
12	Affect 影响	73	0.70%	11	419	0.73%
13	Contrast 对比	66	0.63%	12	297	0.52%
14	BetterSolution 更佳方案	55	0.53%	14	199	0.35%
15	Continue 继续	30	0.29%	15	160	0.28%
16	FutureInterest 未来兴趣	6	0.06%	16	44	0.08%
17	Similar 类似	3	0.03%	17	28	0.05%
18	Aware 清楚	2	0.02%	18	18	0.03%

在非引用语境中,各类功能的行为动词的占比和排序与引用语境中比较一致。相比而言,表示“兴趣”(interest)、“使用”(use)、“问题”(problem)的三类行为动词在引用语境中的存在度相对较高,而表示“文章结构”(text structure)、“改变”(change)的动词在非引用语境中的存在度相对较高。

3.3 连接词

表6 中列出了在引用语境中出现次数最多的前40个连接词。排在前三位的连接词分别是“also”“but”和“however”,分别出现了 725 次、515 次和 290 次,分别占全部引用语境的 6.97%、4.96% 和 2.79%。从含义来看,“also”表示类同或列举,而“but”和“however”表示转折。这表明并列和转折是引用语境中常用的一种逻辑关系。

表 6 引用语境中的高频连接词列表

排序	连接词	词次(引用语境)	词频/排序(非引用语境)	排序	连接词	词次(引用语境)	词频/排序(非引用语境)
1	also	725	2 714/1	**21**	**for instance**	**91**	**282/29**
2	but	516	2 668/2	**22**	**in addition**	**81**	**321/25**
3	**however**	**290**	**1 464/6**	**23**	**instead**	**70**	**254/32**
4	first	281	1 869/4	**24**	**previously**	**57**	**133/42**
5	if	265	2 655/3	**25**	**yet**	**56**	**226/35**
6	**for example**	**227**	**776/14**	26	furthermore	55	301/26
7	**since**	**218**	**990/11**	26	hence	55	685/16
8	because	192	1 059/8	26	still	55	371/23
8	while	192	1 125/7	**29**	**moreover**	**51**	**267/31**
10	**, as**	**178**	**806/13**	29	whereas	51	301/26
11	**although**	**171**	**589/19**	**31**	**indeed**	**47**	**253/33**
12	**, for**	**170**	**774/15**	**32**	**in fact**	**46**	**182/39**
13	**even**	**155**	**650/18**	33	either	40	292/28
14	then	138	1520/5	**33**	**though**	**40**	**199/37**
15	**at least**	**136**	**676/17**	35	next	35	396/22
16	**as well**	**134**	**557/20**	**36**	**actually**	**34**	**191/38**
17	so	131	1 019/10	**37**	**nevertheless**	**31**	**151/40**
18	thus	124	1 051/9	38	clearly	30	345/24
19	**rather**	**101**	**514/21**	38	in that	30	200/36
20	therefore	100	830/12	38	so that	30	246/34

对比引用语境和非引用语境中的连接词排序,可以发现,“however”“for example”“since”“as”“although”等表示转折或列举关系的连接词在引用语境中的词频相对更高,而“if”“then”“so”“thus”“therefore”“hence”等表示因果关系的连接词在引用语境中的出现频次相对较低。这说明,引用语境主要通过引用他人的“一系列”(表示列举的连接词)成果,并通过对它们的评价,“转而”(表示转折的连接词)说明作者自己的研究的优势和意义。从推理的角度来讲,引用语境中主要以用反衬的方法,而不是直接推理的方法(因果关系),作为文章的论述的逻辑方法。

按照表 2 中的连接词分类,在引用语境中出现最多的句式是“转折”句(negative polarity),共出现了 1 489次,占引用语境的 14.31%。转折句同时也是非引用语境中最常出现的句式,占非引用语境的 13.71%。排在第二的是表示“类比”(similarity)的句子,在引用语境中占 8.61%,但它在非引用语境的频次仅排在第四位,说明这种句式更常出现在引用语境中(见表 7)。

对比引用语境和非引用语境中的句式排名,在引用语境中表示“类同”(similarity)、“因果”(cause)和“时间”的句式相对较多,而表示“结果”(result)、“假设”(hypothetical)的句式相对较少。

3.4 以 Hirsch 2005 一文的引用语境为例

接下来,以 JOI 期刊中被引最多的论文-h-index 一文[21](下称 Hirsch 2005)的引用语境为例,展现是否可以基于线索词对引用时的动机和引文的功能进行判别。在本文所用的数据集中,有 121 篇论文引用了 Hirsch 2005 一文,其中 59 篇论文中提及该文不只一次,最终提取得到 243 条引用该文的引用语境。

统计这些引用语境中的高频线索词,并撷取其中代表性的引用语境例句。在表 8 中,Hirsch 2005 的引

表 7　基于连接词的句式类型数量及占比

排序	功能类型	引用语境	占比	排序	非引用语境	占比
1	Negative Polarity 转折	1 489	14.31%	1	7 780	13.60%
2	**Similarity 类同**	**896**	**8.61%**	**4**	**3 509**	**6.13%**
3	Result 结果	854	8.21%	2	5 781	10.11%
4	**Cause 因果**	**577**	**5.54%**	**6**	**2 957**	**5.17%**
5	Sequence 顺序	536	5.15%	5	3 351	5.86%
6	**Temporal 时间**	**472**	**4.53%**	**7**	**2 256**	**3.94%**
7	Hypothetical 假设	402	3.86%	3	3 764	6.58%
8	Additional Information 补充	287	2.76%	8	1 288	2.25%
9	Restatement 重述	137	1.32%	9	696	1.22%
10	Digression 离题	1	0.01%	10	13	0.02%

用语境中出现最多的人称代词为 his(24 次)、we(22 次)、her(18 次)和 their(18 次)等，“we”和“their”的高频出现与前面的情况一致，而“his”和“her”出现次数较多的原因是 h-index 的原定义中有“his or her N papers”的说法。出现最多的行为动词分别是 measure(33 次)、propose(25 次)、define(17 次)和 introduce(17 次)等，这些动词表明 Hirsch 2005 一文的主要贡献是提出并定义了一个新的测量指标。出现最多的连接词是 at least(23 次)、if(21 次)、also(19 次)和 first(13 次)等，前两个高频连接词的出现是因为 h-index 的原定义中包含有“A scientist has h-indexh **if** h of his n papers have **at least** h citations each...”这样的描述。

表 8　Hirsch 2005 一文的引用语境中的高频线索词

线索词		词频	e.g. 引用语境
人称代词	his	24	Hirsch, who other than being author of numerous articles in physics, is also the inventor of the renowned index known by his name (Hirsch, 2005).
	we	22	In particular we intend to show in this article that the h-index or Hirsch (2005) index can be considered hype in our field.
	her	18	A scientist has index h if h of his or her N papers have at least h citations each and the other papers have less than h citations each.
	their	18	Notably, both the g index (Egghe, 2006) and the h index (Hirsch, 2005), as well as the total citation count per researcher, show equivalent statistical properties in terms of their distributions.
	he	6	As a physicist, Hirsch himself published only three articles (Hirsch, 2005, 2007, 2010) related to the index he created.
行为动词	measure	33	Hence, the h index measures both productivity and impact of his/her work (Hirsch, 2005; Hirsch, 2007).
	propose	25	One of the most recent and successful indicators was proposed by Hirsch (2005) and it is called h-index.
	define	17	It is defined as the number h such that, for a general group of papers, h papers received at least h citations while the other papers received no more than h citations (Hirsch, 2005).
	introduce	17	The following definition provides a formal mathematical description of the h index introduced by Hirsch (2005) and the w index introduced by Woeginger (2008).
	use	15	Today, the h index (Hirsch, 2005) is a widely used measure of scientific performance.
连接词	at least	23	The h index for a publication set is the highest number n such that there are n papers in the set each of them received at least n citations (Hirsch, 2005).
	if	21	A scientist has h index h if h of his n papers have at least h citations each and the other n-h papers have fewer than h + 1 citations each (Hirsch, 2005).
	also	19	Just like for any author, the well-known Hirsch's index (Hirsch, 2005) may also be calculated for journals (cf. Braun, Glanzel, & Schubert, 2006).
	first	13	We refer to measures such as the h index, first suggested by Hirsch (2005) for the evaluation of individual scientists.
	since	11	For this reason, Hirsch (2005) already proposed dividing the h-index by number of years since a scientist's first publication and called it the m-quotient.

可以看出，基于线索词可以在一定程度上展现引用语境的语义，进而推测引用的动机和功能，尤其是借助行为动词这一表意具体、功能明确的线索词类型，如“measure”“propose”等。而基于人称代词、连接词等含义宽泛、用法灵活的线索词，则较难直接推断引用语境的语义和功能，这也是在对引用进行分类时主要采用行为动词作为分类特征的原因。

4 结论

近年来,随着开放获取运动的兴起,学术论文的全文本数据正变得越来越易得,这为引用语境的提取和分析提供了更为方便的数据基础。在本文中,笔者以信息计量学领域的重要期刊 JOI 为例,对引用语境中人称代词、行为动词、连接词等三类主要的线索词进行了统计和分析。结果表明,最为常见的人称代词包括“we”和“their”,表明引用语境的基本模式是“我们”参照了“他们的”研究方法和成果。最常用的行为动词是“use”“base”和“study”,表明引用语境中重点关注的是研究方法和研究基础。最常用的连接词是“also”“but”和“however”,表明引用语境的最常用的逻辑是列举关系和转折关系。通过了解引用语境的一般特征,为将来比较不同学科或不同引文的引用语境的特征提供了基础。

实证研究还表明,借助线索词,尤其是行为动词,可以在一定程度上识别出引用语境的主要句式,推测引用的主要动机和引文的主要功能。但是,引用的类型是复杂多样的,如何综合线索词及其组合,构建引用语境的自动分类模型,还有待进一步的探索和研究。

参考文献:

[1] BERTIN M, ATANASSOVA I, LARIVIERE V, et al. The distribution of references in scientific papers: an analysis of the IMRaD structure[EB/OL]. [2016 - 11 - 20]. https://www.researchgate.net/profile/Iana_Atanassova/publication/249652069_The_Distribution_of_References_in_Scientific_Papers_an_Analysis_of_the_IMRaD_Structure/links/0c96051e6649364120000000/The-Distribution-of-References-in-Scientific-Papers-an-Analysis-of-the-IMRaD-Structure.pdf.

[2] BERTIN M, ATANASSOVA I, SUGIMOTO C R, etal. The linguistic patterns and rhetorical structure of citation context: an approach using n-grams[J]. Scientometrics, 2016, 109(3): 1417 - 1434.

[3] MERCER R E, DI MARCO C, KROON F W. The frequency of hedging cues in citation contexts in scientic writing[C]//17th conference of the Canadian Society for Computational Studies of Intelligence. Berlin: Springer, 2004: 75 - 88.

[4] HYLAND K. Hedging in scientific research articles[M]. Netherlands: John Benjamins Publishing, 1998.

[5] FINNEY B. The reference characteristics of scientific texts[D]. London: City University, 1979.

[6] GARZONE M, MERCER R E. Towards an automated citation classifier[C]//AI 2000. Berlin: Springer, 2000: 337 - 346.

[7] PHAM S B, HOFFMANN A. A new approach for scientific citation classification using cue phrases[C]//AI 2003. Berlin: Springer, 2003: 759 - 771.

[8] NANBA H, OKUMURA M. Towards multi-paper summarization using reference information[J]. Journal of natural language processing, 1999, 6(5): 43 - 62.

[9] TEUFEL S, SIDDHARTHAN A, TIDHAR D. Automatic classification of citation function[C]// Proceedings of the 2006 conference on empirical methods in natural language processing. New York: ACM, 2006: 103 - 110.

[10] TEUFEL S. Argumentative zoning: information extraction from scientific text[D]. Edinburgh: University of Edinburgh, 1999.

[11] RADOULOV R. Exploring automatic citation classification[D]. Ontario: University of Wateloo, 2008.

[12] YU H, AGARWAL S, FRID N. Investigating and annotating the role of citation in biomedical full-text articles[C]// 2009 IEEE International conference on bioinformatics and biomedicine. Piscataway: IEEE, 2009: 308 - 313.

[13] AGARWAL S, YU H. Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion[J]. Bioinformatics, 2009, 25(23): 3174 - 80.

[14] AGARWAL S, CHOUBEY L, YU H. Automatically classifying the role of citations in biomedical articles[EB/OL]. [2016 - 11 - 20]. https://www.ncbi.nlm.nih.gov/pubmed/21346931.

[15] DING Y, ZHANG G, CHAMBERS T, et al. Content-based citation analysis: the next generation of citation analysis[J]. Journal of the Association for Information Science and Technology, 2014, 65(9): 1820 - 1833.

[16] ZHANG G, DING Y, MILOJEVIC S. Citation content analysis (CCA): a framework for syntactic and semantic analysis of citation content[J]. Journal of the American Society for Information Science and Technology, 2013, 64(7): 1490 - 1503.

[17] JAMALI H R, NABAVI M. Open access and sources of full-text articles in Google Scholar in different subject fields[J]. Scientometrics, 2015, 105(3): 1635 - 1651.

[18] LIU X, ZHANG J, GUO C. Full-text citation analysis: a new method to enhance scholarly networks[J]. Journal of the American Society for Information Science and Technology, 2013, 64(9): 1852 - 1863.

[19] Nakov P I, Schwartz A S, Hearst M A. Citation sentences for semantic analysis of bioscience text[C]//27th annual international ACMSIGIR conference. New York: ACM, 2004: 13 - 17.

[20] KNOTT A. A data-driven methodology for motivating a set of coherence relations[D]. Edinburgh: University of Edinburgh, 1996.

[21] HIRSCH J E. An index to quantify an individual's scientific research output[J]. Proceedings of the National Academy of Sciences, 2005, 102(46): 16569 - 16572.

作者贡献说明:

胡志刚:负责研究设计、数据处理和论文撰写;

孙太安:负责数据处理和论文修改;

王贤文:负责研究设计、论文指导和论文修改。

Citation Context Analysis Based on Cue Words: A Case Study of *Journal of Informetrics*

Hu Zhigang　Sun Tai'an　Wang Xianwen

[1] Institute of Science of Science and Science and Technology Management, Dalian University of Technology, Dalian 116024

[2] WISE Laboratory, Dalian University of Technology, Dalian 116024

Abstract: [**Purpose/significance**] A citation context is a descriptive or evaluative sentence in scientific papers that contains references. By extracting and analyzing the cue words in the citation context, we are able to identify the citation behavior and motivations. [**Method/process**] Taking citation contexts in *Journal of Informetrics* as an example, we selected three kinds of commonly used cue words, such as personal pronouns, behavioral verbs and conjunctions, and respectively calculated the frequency, proportion and ranking in the citation context. By comparing the existence of cue words in the citation context and non-citation context, the common sentence pattern of citation contexts was characterized. [**Result/conclusion**] In *Journal of Informetrics*, the citation context mainly shows the following characteristics: a) focusing on the first-person and third-person perspectives, the citation context not only shows the works of others, but also shows the authors' own research; b) citing papers prefer to citing references involved research methods, and the most commonly used behavior verbs are "use" "base" and "study"; c) adversative and enumerative sentences are the main demonstration modes. The most used conjunctions are "also" and "but". In conclusion, the citation context analysis based on cue words has important value and significance for us to better understand the functions and motivations of citations in scientific papers.

Keywords: citation context　cue words　word frequency analysis　citation

《知识管理论坛》征稿启事

《知识管理论坛》(ISSN 2095－5472，CN11－6036/C)获批国家新闻出版广电总局网络出版物正式资质，2016年全新改版。本刊关注知识的生产、创造、组织、整合、挖掘、分享、分析、利用、创新等方面的研究成果。任何有关政府、企业、大学、图书馆以及其他各类实体组织和虚拟组织的知识管理问题，包括理论、方法、工具、技术、应用、政策、方案、最佳实践等，都在本刊的报道范畴之内。本刊实行按篇出版，稿件一经录用即进入快速出版流程，并实现立即完全的开放获取。

2017 年各期内容侧重于：互联网＋知识管理、大数据与知识组织、实践社区与知识运营、内容管理与知识共享、知识创造与开放创新、数据挖掘与知识发现。现面向国内外学界业界征稿：

1. 稿件的主题应与知识相关，探讨有关知识管理、知识服务、知识创新等相关问题。文章可侧重于理论，也可侧重于应用、技术、方法、模型、最佳实践等。

2. 文章须言之有物，理论联系实际，研究目的明确，研究方法得当，有自己的学术见解，对理论或实践具有参考、借鉴或指导作用。

3. 所有来稿均须经过论文的相似度检测，提交同行专家评议，并经过编辑部的初审、复审和终审。

4. 文章篇幅不限，但一般以 4 000－20 000 字为宜。

5. 来稿将在 1 个月内告知录用与否。

6. 稿件主要通过网络发表，如我刊的网站(www. kmf. ac. cn)和我刊授权的数据库。同时，实行开放获取、按篇出版和按需印刷。

请登录 www. lis. ac. cn 投稿，注明"知识管理论坛投稿"。

联系电话：010－82626611－6638　　联系人：刘远颖