维基百科历史类文献的参考文献分析

Analysis on References in Wikipedia's History Articles

杨 阳

(中国科学技术信息研究所 北京 100038)

摘 要 分析了维基百科中历史类文献的参考文献,随机选择了50篇标有参考文献的文章进行分析。统计结果显示:无论是参考文献的数量还是使用次数,维基百科与历史学专业期刊《史学月刊》存在较大差距,维基百科参考文献的主要来源是图书,占全部参考文献的75.68%。与《史学月刊》几乎没有来源于网络的参考文献不同,维基百科的参考文献有22.11%的参考文献来源于网络,并且主要来自网络新闻媒体。研究结果表明:维基百科50篇文章的质量与《史学月刊》的文章存在较大差距。最后提出了一些参考意见以改善维基百科的信息质量。

关键词 维基百科 参考文献 历史 信息质量

中图分类号 G250 **文献标识码** A **文章编号** 1002-1965(2010)10-0051-03

维基百科在最开始的时候就确立了自己的三项核心内容方针,可供查证(verifiability)就是其中之一,其意思是:加入维基百科的内容须要发表在可靠来源中能被读者查加,而不能仅由维基百科的工作人员认为它真实正确[1]。这也就是说:在维基百科上发表的内容必须标注"参考文献"以证明其来源。这激发了我们的一个想法:即希望能通过情报学中的引文分析方法对维基百科的参考文献进行分析,看能否发现一些有趣的现象。

为了保证内容的真实性,维基百科提供了一系列的指导方法,但对于这种方法是否真的可行,目前很少有人进行研究。本论文希望能通过对于维基百科历史方面文章的参考文献分析,对这个问题进行回答。

研究发现:首先,并不是所有的内容都清晰地标注了参考文献;其次,这些参考文献有很多来源于网络信息资源,尤其是网络新闻媒体。本文认为:需要对包括网络新闻媒体在内的网络环境多加注意,因为这些网络环境为维基百科提供了引文的资源,这些内容进而被维基百科的用户们逐步扩散出去。认识到这个问题有助于帮助我们改进维基百科的内容质量,帮助用户获得更加客观真实的信息,从而为全民信息素养的提高产生积极影响。

1 相关研究综述

维基百科易于编辑的特性是它迅速流行和发展的原因之一,这种特性带给用户一种满足感,进而激励了用户进一步帮助维基百科的发展。但并不是所有的人都对维基百科持赞赏态度。有一些观点认为这种开放性的系统助长了错误信息的泛滥。例如:2005年,曾经担任过罗伯特·肯尼迪助手的记者老约翰·席根塔勒(John Seigenthaler)在《今日美国》上撰文,指出维基百科中有一篇文章错误地把他和罗伯特·肯尼迪及约翰·F·肯尼迪遇刺联系起来[2]。在2006年,人们发现有超过1 000篇文章被美国国会成员修改,目的是消除对一些政客的负面评论[3]。2007年,一个维基百科的资深用户被发现使用欺骗性的身份,目的是诱使人们接受他的某些观点[4]。美国佛蒙特州的Middlebury大学,历史系直接禁止学生在论文中使用维基百科作为参考文献[5]。

尽管针对维基百科内容真实性的评论有很多,但是很少有人进行系统性的研究。2005年在Nature上发表的一篇文章进行过一次有益的尝试,这篇文章挑选了维基百科和大英百科全书的50个科学领域的文章,将它们送到各领域的专家手中,得到了关于这些文章的42份报告,专家们指出了这些文章出现的错误(维基百科有162处,大英百科全书有123处)[6]。

关于维基百科内容的质量问题,有很多批评意见。虽然维基百科内容的质量问题可能由很多因素引起,但是本文只是从参考文献的角度进行分析,我们的一个假设是:如果维基百科的内容能尽可能引用核心作者的文章或者权威的期刊和出版物,或者采用经过同行评议的内容,那么在未来,维基百科的文章可能会作

收稿日期:2010-05-12 修回日期:2010-06-22

作者简介:杨 阳(1984-),男,硕士,研究实习员,研究竞争情报、科技政策。

为学术研究的参考文献使用。从这种角度来看,维基百科制定的“可供查证”的核心内容方针是非常重要的,这种制度在一定程度上解决了文章来源的真实性问题。而通过对维基百科关于历史领域文章的参考文献分析,可以对这些文章的内容质量做初步的评价。

本文采用历史类文章作为分析样本的原因有两点:第一,Spoerri 在2007年的研究表明:政治和历史方面的综合性文章比计算机科学等自然科学类文章拥有更高的访问量[7];第二,由于历史本身具有解释性的特点,因此历史类文章参考文献的选择,直接影响了其内容的真实性。

2 研究方法

维基百科的中文版对于各国历史有一个专门的分类页面[8],这个页面包含112个子分类,介绍了106个国家历史类的文章,本文选择50篇包含参考文献的文章,其中40篇关于中国历史,10篇关于世界历史(这样选的原因是方便我们与历史学的核心期刊《史学月刊》进行比较,该刊刊载的中国历史与世界历史方面的文章大致也是这个比例)。这50篇文章全部是2010年4月14日的版本,对于每篇文章,我们重点采集以下信息:50篇文章总的参考文献数量(不包括重复出现的参考文献);这些参考文献在50篇文章里一共被引用了多少次;这些文章每百字所使用的参考文献数量;这些参考文献有多少是基于网络内容的;这些基于网络内容的参考文献是否来源于付费的网络资源;非网络来源的参考文献的文献格式(书籍、期刊、等等);网络参考文献的链接现在是否依然可用。

本文随机挑选了《史学月刊》2009年刊登的50篇文章与维基百科的50篇文章进行比较,《史学月刊》主要发表中国古代史、中国近代史、世界史、史学理论、史学史、各种专业史等方面的研究成果,是中文历史期刊中比较权威的一种刊物[9]。我们的研究主要对两者的以下方面进行比较:参考文献的数量;参考文献在文章中的被引用次数;文章每百字所使用的参考文献数量。

当然,维基百科的文献与历史学专业期刊的文献在目的上就是截然不同的,并且维基百科发表的文章多为概念定义等类型的文章,而《史学月刊》上的文章多为学术研究性文章。这两种类型的文章,在引文(参考文献)的结构方面应该有不同的特点,将它们直接进行比较可能会存在一些问题,因为维基百科的内容质量可能根本无法与专业期刊进行比较,但是作为一种越来越被大众接受并广泛使用的工具,维基百科有理由做得更好。

通过本文的分析,可以展现维基百科和历史专业期刊文章在引文结构上的不同,并且能通过对比分析,为进一步提升维基百科的内容质量提供参考性意见。

3 数据分析

维基百科中随机选择的50篇文献包含了407篇参考文献,被引用了530次。《史学月刊》中随机选择的50篇文献包含了964篇参考文献,被引用了985次。考虑到文章的长度不同,对参考文献的数量可能会有一定的影响,采用文献中每百字引用的参考文献数量进行比较,《史学月刊》的文献每百字引用的参考文献为0.1740。维基百科的文献每百字引用的参考文献为0.1305。从统计的结果来看:维基百科的文献,其参考文献的数量要少于《史学月刊》的文献。

为了对参考文献的质量进行分析,本次研究将《史学月刊》和维基百科的参考文献分为两种类型:基于网络资源的参考文献和基于非网络资源的参考文献。在维基百科使用的407篇参考文献中,有90篇来自网络资源,占总数的22.11%,而在《史学月刊》的964篇参考文献中,只有1篇来自网络资源。分析的结果表明:维基百科中的文献相比较《史学月刊》的文献,其内容引自网络资源的比例相对偏大。对这90篇来自网络资源的文献进行分析,研究发现这些文献全部来自于免费的网络资源(除去不能打开的页面链接)。

对维基百科中来自非网络资源的317篇参考文献的种类进行分析,结果如表1所示。

表1 来自非网络资源的参考文献的种类

种类	数量	占非网络资源的百分比(%)	占全部参考文献的百分比(%)
图书	308	97.16	75.68
学术期刊	7	2.21	1.72
大众期刊	2	0.63	0.49

结果表明:在来自非网络资源的参考文献中,图书是使用比例最高的参考文献,占到了总数的97.16%,如果算上非网络资源的参考文献,这个比例有所下降,达到75.68%。学术期刊的使用偏少,只有7篇,占非网络资源参考文献总数的2.21%。

对于来自网络资源的参考文献,将它们按照网站的创建者进行分类,因为根据域名,可以方便对网站类型进行处理,例如:.gov 域名的网站来自政府,.org 域名的网站来自非盈利组织,.ac/ .edu 域名的网站来自研究院所/学校。由于网页的种类繁多,对于一些很难鉴别的网页,我们把它们标注在其它类型中。分析得到的结果如表2所示。

在网络新闻媒体中,新华网的出现频率最高,有8条参考文献来自于新华网;在历史类主题网站中,国学网的出现频率最高,有4条参考文献来自于国学网,由

于其他几种类型的网络资源所在网站比较分散,这里不做进一步的分析。

表 2　来自网络资源的参考文献的种类

种类	数量	占网络资源的百分比(%)	占全部参考文献的百分比(%)
网络新闻媒体	31	34.44	7.62
非盈利组织	13	14.44	3.19
政府组织	10	11.11	2.46
学校/研究院所	9	10.00	2.21
历史类主题网站	9	10.00	2.21
博客	2	2.22	0.49
其它	16	17.78	3.93

在维基百科标注的来自网络的参考文献中,并不是所有的网页链接都能正常打开,在实验中,有 22 篇网页链接无法打开,占所有网络资源参考文献的 24.44%。对于不能正常打开的网页链接,在本次研究中通过打开这些网页的主页来推断这些网页的种类,对于一些连主页也无法显示的网页,我们将它们放在其它类别中。

4　对结果的讨论

尽管在随机挑选的 50 篇维基百科的文章中,没有挑选那些不含参考文献的文章,但是得到的结果依旧表明维基百科文章的参考文献数量(407 篇)远远比不上《史学月刊》文章的参考文献数量(964 篇),即使采用每百字引用的参考文献作为指标,这一结论也没有发生改变,《史学月刊》每百字引用的参考文献为 0.1740 。维基百科每百字引用的参考文献为 0.1305。研究还发现:维基百科文章的编辑者更多的使用来自网络的信息资源,而《史学月刊》的论文作者几乎从来不使用来自网络的信息资源作为参考文献。

如果假设维基百科所有的来自网络的参考文献都是免费获取的,那么就需要关注一下信息商品化问题,现代信息技术已经使得信息作为一种商品在市场上销售,信息的商业价值在某种程度上远远大于其政治和社会价值,文本数字化技术的重要目的之一就是使信息商品化。网络的发展方便了我们及时获取所需的信息,但是大部分高质量的信息都存在于付费的数据库中,虽然有一些尝试是关于信息的开源化,例如 Google 的数字图书馆项目[10],但这些尝试仅仅还处于初级阶段。因此,如果维基百科的文章编辑者无法获得商业数据库中的信息资源,那么他们文章的可信度和质量便会降低。

当进一步分析参考文献的类型时,我们发现:在来源于网络的参考文献中,有 34.44% 的文献来自网络新闻媒体,但网络新闻媒体的信息发布是具有很大随意性的,它们没有经过严格的审查,特别是没有经过同行评议。事实上,即便是对于网络环境下信息质量的评价体系,目前也是众说纷纭[11-13]。很显然,将网络新闻媒体的文章作为重要的参考文献来源,对于百科全书来说是,文献来源的可靠性会有所降低。

5　建议

第一,虽然研究表明:从所引用的参考文献来看,维基百科的文献质量不高,但是我们不能否认维基百科拥有广大的用户群体,并且在知识的共享方面有着非常杰出的贡献,因此历史学家们有必要为文献质量的提高做出一些努力,相比较于专业期刊,维基百科可以作为一个很好的普及历史知识的平台。

第二,本文只是关于历史方面的一个研究,如果扩充到所有的学科领域,需要更多的领域专家共同努力以提高维基百科的文献质量。作为传播知识的重要成员,图书馆员们也应当为维基百科信息质量的提高做出自己的贡献。虽然信息的商品化对维基百科的文章编辑者造成了很大的影响,但是开源运动的不断发展必定会将这种影响不断降低。

第三,信息素养作为一个缩小数字鸿沟,走向知识社会的重要手段,越来越受到各个国家的关注[14]。维基百科在信息构建和信息共享方面具有自己独特的优势,与专业研究成果可以形成良好的互补,如果能更加关注维基百科在提高全民信息素养方面起到的积极作用,那么不论是对维基百科内容贡献者热情的激励还是对贡献内容质量的提高都非常有好处。

6　结论

本文主要分析了维基百科历史类文献的参考文献,研究表明:许多历史类的文献并没有标注参考文献,即使我们选择 50 篇标有参考文献的文章进行分析,结果也不是十分理想,许多参考文献来源于网络新闻媒体,这些参考文献的质量得不到保证。但是,维基百科作为一个免费的信息获取平台,其影响群体远大于那些可以获取付费数据库资源的研究人员,文献质量的好坏直接影响了公众对事物的认知,因此对于维基百科文章质量的提高是一项非常重要的任务。文章的第 5 部分提供了一些关于改进维基百科文献质量的一些参考性意见。

参 考 文 献

[1]　Wikipedia:可供查证[EB/OL]. Wikipedia http://zh.wikipedia.org/wiki/Wikipedia:%E5%8F%AF%E4%BE%9B%E6%9F%A5%E8%AF%81,2010-4-15

[2]　Seigenthaler J. A false Wikipedia'biography'[EB/OL]. USA Today:http://www.usatoday.com/news/opinion/editorials/2005-11-29-wikipedia-edit_x.htm,2010-4-15

(下转第 93 页)

(上接第53页)

[3] Mehegan D. Bias, sabotage haunt Wikipedia's free world[EB/OL]. The Boston Globe:http://www.boston.com/news/nation/articles/2006/02/12/bias_sabotage_haunt_wikipedias_free_world/, 2010-4-15

[4] Lih A. TheWikipedian Revolution: How a Bunch of Nobodies Created the world's Greatest Encyclopedia. NewYork: Hyperion Books,2009

[5] Read B. Middlebury College History Department limits students' use of Wikipedia[EB/OL]. The Chronicle of Higher Education: http://chronicle.com/article/Middlebury-College-History/23736,[2010-4-15]

[6] Giles J. Lnternet Encyclopedias go Head to Head[J]. Nature, 2005,438(7070): 900-901

[7] Spoerri A. What is popular on Wikipedia and why? [EB/OL]. First Monday: [2010-4-15]. http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1765/1645

[8] 分类:各国历史[EB/OL]. Wikipedia: http://zh.wikipedia.org/zh-cn/Category:%E5%90%84%E5%9C%8B%E6%AD%B7%E5%8F%B2,[2010-4-15]

[9] 李 佳. 我国历史学期刊的h指数分析——基于CSSCI(2000-2007年)的数据[J]. 西南民族大学学报,2009(8):92-95

[10] 周军兰. Google数字图书馆项目的多方博弈分析[J]. 大学图书馆学报,2006,24(5):20-27

[11] 宋立荣,褚军亮. 网络信息环境下信息质量管理的初步认识[J]. 现代情报,2009,29(10):53-56

[12] 戴维民. "网络为王"时代的媒体公信力认定——网络媒体评价指标与方法[J]. 图书情报工作,2004,48(1):33-38

[13] 刘雁书,方平. 网络信息质量评价指标体系及可获取性研究[J]. 情报杂志,2002(6):10-12

[14] 杨 阳,张新民. 信息素养的生命周期[J]. 图书情报工作,2009,53(3):30-33,24

(责编:白燕琼)