网络百科词条质量评价研究综述 *

黄令贺 李明泽 于金平
（河北大学管理学院，保定 071002）

摘 要：[**目的 / 意义**] 对以往网络百科词条质量评价研究进行梳理，了解词条质量评价研究如何展开，以及存在的问题。[**方法 / 过程**] 运用文献调查、定性分析和定量分析法对网络百科词条质量评价研究现状进行梳理，从人工评价、自动评价两个方面展开分析。[**结果 / 结论**] 与权威信息资源相比，网络百科词条质量整体偏低；词条质量人工评价结果不一致的情况具有一定普遍性；词条质量自动评价的准确率与分类数量基本成反比；两类研究的方法虽然不同，但对词条质量的认知以及评价思路基本一致。

关键词： 网络百科 人工评价 自动评价 评价标准 评价指标
分类号： G250
DOI： 10.31193/SSAP.J.ISSN.2096-6695.2020.01.11

0 引 言

网络百科独特的编纂模式是其成功的关键，也是引发词条质量被专业人士质疑的根源。美国图书馆协会前主席 Gorman 将维基百科比作是营养缺乏的快餐[1]。更具代表性的是，大量教师禁止学生在论文写作中引用网络百科词条[2]。那么，与精英参与编纂的传统百科等权威资源相比，网络百科词条质量究竟如何？经过前期尝试性的文献检索，笔者发现已有大量研究关注这一问题。2005 年 12 月发表在 *Nature* 上的“Internet encyclopedias go head to head”一文认为，维基百科中自然科学类词条质量已经可以媲美《大英百科全书》[3]。但部分研究证实，医学类词条质量显著低于 Medscape Drug Reference、LexiComp Online 等全球权威资源数据库[4][5]。这种研究结果的不一致是个例，还是普遍存在？其背后的原因是什么？这些问

* 本文系河北省教育厅人文社会科学研究重点项目“信息生命周期视角下 Wiki 信息质量评价体系研究”（项目编号：SD181068）的成果之一。
[**作者简介**] 黄令贺（ORCID：0000-0001-5810-2630），男，副教授，博士，研究方向为信息资源管理，Email：linghehuang@hbu.edu.cn；李明泽（ORCID：0000-0001-8717-6209，通讯作者），男，硕士研究生，研究方向为信息资源管理，Email：2294157171@qq.com；于金平（ORCID：0000-0002-3576-7469），女，硕士研究生，研究方向为信息资源管理，Email：2583313235@qq.com。

题值得探究。

另一方面，为提升用户体验，网络百科逐渐形成了以“用户评价、词条质量分类、推荐使用”为核心的工作机制[6]。但随着词条数量的急剧增加，低效的人工评价方式已不能满足词条分类的需求[7]。以维基百科为例，质量等级明确的词条尚不足总数的六分之一[8]，而有待评价的词条仍以指数级增长[9]。因此，词条质量自动评价成为一个亟待解决的问题，并引发了众多研究者的关注。前期探索性的文献检索发现，这类研究的评价准确性也有较大差异，然而并无相关研究予以全面总结。

以上两类研究都属于信息质量评价研究的范畴。为了解两类研究中评价结果差异的情况、具体原因、两类研究的整体现状以及研究者对网络百科词条质量的认知和评价思路，本研究从评价结果出发，在评价对象、评价方法等方面对这两类研究进行系统梳理，以期为进一步研究提供方向指导。

1　文献来源与分析方法

1.1　文献来源

本文研究的网络百科是指基于 Media Wiki 或类似技术而建立的支持用户协作编纂词条的一种内容生产网络平台，代表有维基百科和百度百科。在国外，网络百科通常用 Wiki、Wikipedia、Online Encyclopedia、Internet Encyclopedia、Web-based Encyclopedia、Open-source Encyclopedia、Community-based Encyclopedia 等词来表示。而在国内，对应的词主要有“维基”“维基百科”“在线百科”“网络百科”，以及更为具体的“百度百科”等。基于以上分析，考虑到英文词可能出现的复数情况，确定网络百科对应的外文检索词为“Wiki*”与“encyclopedia*”。考虑到中文文献可能直接使用英文关键词的情况，确定网络百科对应的中文检索词为“百科”、“维基”与“Wiki*”。

词条质量属于信息质量的范畴，而信息质量又是一个多维概念[10]。为了更广泛地收集相关文献，通过对信息质量及相关概念的分析[11]，最终确定关键检索词为质量、准确性、完整性、权威性、可信、可靠、信任、价值，对应的英文词为 quality（qualities）、accuracy（accurate）、completeness（complete）、authority(authoritative)、credibility（credible）、reliability（reliable）、trust（trustworthiness、trustworthy）、value（valuable）、worth（worthy）。另外考虑到部分文献（例如发表在 *Nature* 上针对维基百科词条质量评价研究的那篇文章）的题名与摘要中均未出现“评价”，但确实属于网络百科词条质量评价研究，因此在文献检索时，并没有添加“评价”这一关键词，其目的是使相关文献尽可能多地进入检索范围。

选取 WOS、Elsevier、Emerald、Wiley、ACM Digital Library、ProQuest Digital Dissertation、CNKI、万方、维普等数据库为主文献源。检索方式采用主题检索，英文检索式为：“TS=((Wiki* OR encyclopedia*) AND (quality* OR accura* OR complete* OR authorit* OR credib* OR reliab* OR trust* OR value* OR worth*))”，中文检索式为：“SU=（‘百科’+‘维基’+‘Wiki*’)*(‘质量’+‘准确性’+‘完整性’+‘权威性’+‘可信’+‘可靠’+‘信任’+‘价值’)”。此外，选取 Google Scholar 和百度

学术作为辅助文献源，检索方式采用题名检索，检索式与主文献源相同。文献发表时间截止到2018 年，具体检索时间为 2019 年 3 月 14~15 日。考虑到相关文献来源的广泛性，对文后参考文献进行了分析与回溯检索。在全面的文献检索后，根据表 1 展示的步骤与细则进行筛选，最终得到与网络百科词条质量评价高度相关的文献 109 篇，其中外文文献 94 篇、中文文献 15 篇。这里有两个问题需要说明：（1）文献筛选标准：由于本文希望从两类研究的评价结果出发，所以去掉了只有单纯理论探讨、缺乏具体评价框架或实证结果的研究；（2）在文献筛选的过程中，发现第一类研究基本以人工评价为主，只有小部分研究采用了一些软件辅助统计，基于此，将引言中提到的两类研究分别命名为“网络百科词条质量人工评价”和“网络百科词条质量自动评价”。

表 1　文献筛选步骤与细则

步骤	名称	细则
1	去重	如果一篇文献既有会议版本，又有期刊版本时，保留期刊版本
2	文献类型筛选	剔除掉书评、新闻报道、会议综述等类型的文献
3	研究内容筛选	剔除有关网络百科平台、系统运行、用户贡献等方面的质量评价研究以及探索用户对网络百科质量感知情况等相关性较低的研究
4	研究类型筛选	剔除只有单纯理论探讨、缺乏具体评价框架或实证结果的研究

1.2　分析方法

考虑到两类评价研究存在的差异，为了更好地提取文献信息，特对评价结果、评价对象、评价方法等概念进行了界定，具体如下。①评价结果：在词条质量人工评价研究中，评价结果指的是与其他信息资源相比，网络百科词条质量如何，常用“显著高”“相差不多”“显著低”等词语进行描述；在词条质量自动评价研究中，评价结果则为词条质量分类的准确率。②评价对象：被评价词条所属的网络百科平台。③评价方法：这里的“评价方法”并不是某种具体的方法，而是包含评价标准、评价指标、对比资源、评价人以及相关技术方法在内的一个整体概念。其中，评价标准为判定词条质量高低的准则与尺度，反映了研究者对词条质量的认知和评价思路，其特点具有一定抽象性[12]，因此，评价标准通常由具体评价指标来实现。评价指标是对词条质量判断的依据，是词条质量某方面特征的具体化。对比资源、评价人只出现在词条质量人工评价研究中。对比资源是指与网络百科进行质量比较的权威信息源，常见的有大英百科全书、各种专业数据库以及专业词典等；评价人是指进行比较评价的执行人。技术方法在词条质量自动评价研究中是指采用的机器学习技术及相关算法。此外，两类研究通常还涉及评价范围和样本数量，其中，评价范围是指被评价词条的主题范围，样本数量是指评价研究中实际使用词条的个数。

值得注意的是，评价标准在文献中并不直接出现，通过分析具体的评价指标进行确定。例如，Giles 采用“错误数量”与“遗漏事实数量”两个指标对维基百科与大英百科进行对比[3]。

从字面上分析，“错误”即是与客观实际不符合[13]。如果“错误数量”不为零，则意味着词条内容并没有准确地表达出某些知识或信息的客观实际。“遗漏”是指应该列入或提到却因疏忽而没有列入或提到。如果“遗漏事实数量”不为零，则意味着词条内容并没有完整地表达出知识或信息的客观实际。基于以上分析，结合信息质量内涵[14]，推断“错误数量”与“遗漏事实数量”分别对应“准确性”与“完整性”两个评价标准。以上文献信息提取完毕后，接下来将从词条质量人工评价、词条质量自动评价两个方面进行梳理。

2　网络百科词条质量人工评价研究

2.1　评价对象与评价结果

这类研究的文献数量为 44 篇，占总体的 40.4%，所有研究的样本数量均值为 88，极差为 1172，标准差为 199。其中，样本数量超过 100 的研究仅占 11.4%。

在评价对象方面，均为维基百科，其中绝大多数是英文版本；大部分研究围绕单一语言版本的维基百科展开，涉及多语言版本的研究只有 3 篇[15][16][17]；多数研究只涉及单一大类词条，其中医学类是最受关注的，如表 2 所示。整体上看，评价范围相对集中，但具体评价的词条主题却非常分散。以医学类词条评价为例，涉及的主题就包括健康学[18]、临床医学[4]、药理学[16]等。

表 2　词条质量人工评价对象情况

	文献数量（篇）	百分比（%）
评价对象		
语言版本		
英文版维基百科	40	90.9
西班牙文版维基百科	3	6.8
其他版本	5	11.4
版本数量		
单语言版本	41	93.2
多语言版本	3	6.8
评价范围		
医学主题	33	75.0
历史主题	5	11.4
哲学等	6	13.6
样本数量		
<11	13	29.5
11~100	26	59.1
>100	5	11.4

在对词条质量的评价结果方面，56.8% 的研究显示，网络百科词条质量显著低于大英百科等权威信息资源，如表 3 所示。

表 3　词条质量评价结果分布

		质量			合计（篇）	百分比（%）
		显著低	相同	显著高		
学科名称	医学主题	18	13	2	33	75.0
	历史主题	4	1	0	5	11.4
	哲学等	3	3	0	6	13.6
合计		25	17	2	44	100
百分比（%）		56.8	38.6	4.6	100	

综合以上分析，可以认为与大英百科等权威信息资源相比，网络百科词条质量整体偏低。网络百科词条质量评价研究结果的不一致具有一定普遍性，其背后的原因可能为此类研究在评价范围、样本数量等方面差异较大。相对于网络百科语言版本多、主题范围广、词条数量大的特点，此类研究的评价对象较为单一，评价范围相对集中，样本规模较小。

2.2　评价方法分析

在此类研究中，评价方法具体包括评价人、对比资源、评价标准和评价指标。如表 4 所示，评价人有研究者、领域专家、大学生三种类型，数量通常在 6 人以下。尽管评价人的类型并不多样，但在质量评价研究中，评价人不一致的情况却非常普遍[5][18][19]。由于对比资源种类众多，并未在表 4 中予以体现，但就此类研究的汇总情况来看，所采用的对比资源差异较大。

表 4　评价人及评价标准数量情况

	文献数量（篇）	百分比（%）
评价人		
研究者	32	69.6
领域专家	10	21.7
大学生	4	8.7
评价人数量		
<6 人	33	75.0
6~10 人	6	13.6
>10 人	5	11.4
评价标准数量		
<4 个	39	88.6
4~6 个	4	9.1
>6 个	1	2.3

如表 4 所示，大部分研究采用的评价标准不超过 4 个，表明此类研究在评价标准数量方面差异不大。表 5 展示了此类研究中采用的评价标准及对应指标，括号中的数字代表文献数量。可以

看出，此类研究在评价标准方面的数量分布相对集中。综合以上两方面，本文认为此类研究在网络百科词条质量认知以及从哪些方面进行评价已达成一定共识，其评价思路，以及对各个评价标准的概念及对应指标具有一致性，具体归纳如下。

（1）准确性是对词条内容与事物客观实际情况符合程度的描述，侧重对词条内容的衡量，关注点在于知识点或事实本身[20]。从涉及文献数量来看，准确性应该是研究者最看重的评价标准。在具体评价过程中，绝大部分研究采用“错误数量”直接衡量准确性，其基本逻辑是错误越多，准确性越差。使用“编辑者数量”或“编辑次数”间接对准确性进行衡量的研究只有 2 篇[21][22]，其基本逻辑是编辑者越多或编辑次数越多，词条应该越准确。但部分研究也已证明，编辑者越多、编辑次数越多，编辑者之间更易产生冲突，反而不利于词条质量的提升[23]。因此，“编辑者数量”与“编辑次数”用于准确性的衡量可能存在偏差。

表 5　第一类研究的评价标准与评价指标

（单位：篇）

序号	评价标准	评价指标	序号	评价标准	评价指标
1	准确性（33）	错误数量（32） 编辑次数（2） 编辑者数量（1）	6	及时性（6）	最近更新时间（3） 更新频率（2） 是否与最新知识一致（1）
2	完整性（21）	遗漏事实数量（20） 文字数量（1）	7	一致性（3）	前后内容描述是否一致（3）
3	可读性（19）	图片数量（3） 表格数量（1） 语言风格（1） 词条结构情况（1） Flesch–Kincaid grade level（13） Flesch reading ease（6） Automated readability index（4） Fog scale level（3） Coleman–Liau index（3） Gobbledygook（2） SMOG Index（1）	8	客观性（2）	是否掺杂个人主观内容（1） 是否存在偏见内容（1）
4	可验证性（14）	参考文献数量（11） 是否有合适的参考文献（3）	9	相关性（1）	词条内容是否与主题相关（1）
5	权威性（10）	参考文献质量（9） 搜索引擎结果排名（1）	10	简洁性（1）	文字描述是否简洁（1）

（2）完整性是指词条与事物客观实际相比，其内容完备或没有残缺的程度，侧重对词条内容的衡量。从涉及文献数量来看，完整性的重要程度仅次于准确性。在具体评价过程中，绝大部分研究采用“遗漏事实数量”直接衡量完整性，其基本逻辑是遗漏事实数量越多，完整性越差。使

用“文字数量”间接地对完整性进行衡量的研究只有 1 篇[24]，其基本逻辑是文字数量越多，词条内容越完整。显而易见，文字数量与事实数量并非完全一致。因此，“文字数量”用于完整性的衡量也可能存在一定偏差。

（3）可读性是指词条内容清晰、易读与易理解的程度，侧重对词条描述质量的衡量，关注点在于描述语言或佐证材料（图、表等）是否有利于词条的阅读与理解。从涉及文献数量来看，可读性也是非常重要的一个评价标准。在具体评价过程中，大部分研究采用了如 Flesch - Kincaid grade level 等比较成熟的测量指标[25]。这类指标一般通过统计学方法计算句子中的字数、音节数等，从而达到判断词条阅读难度的目的。此外，还有少部分研究采用了图片数量、表格数量、词条结构情况等指标[26]，其基本逻辑是图片与表格的数量越多，词条结构性越强，越有利于用户阅读与理解。

（4）可验证性指是否有恰当或充分的证据证明词条内容的准确程度，侧重对词条描述质量的衡量。可验证性涉及的文献数量占总数的三分之一，表明了这一标准的重要性。在具体评价过程中，大部分研究使用“参考文献数量”间接衡量这一标准，其基本逻辑是参考文献越多，证明词条内容准确程度的证据越充分。此外，3 篇研究采用了“是否有合适的参考文献”这一指标对可验证性进行衡量[19][27][28]。其基本逻辑是，词条内容与对应参考文献的匹配程度越高，词条可验证性越强。

（5）权威性是指词条具有使人信服的力量和威望。根据网络百科的特点，一般认为能够为词条权威性提供证据的主要是参考文献，具体来说是参考文献的质量。因此，这一标准侧重的也是词条描述质量的衡量。在涉及这一标准的 10 篇研究中，9 篇采用参考文献质量来衡量词条内容的权威性，1 篇采用搜索引擎结果排名来代表词条整体的权威性[29]。

（6）及时性指的是当人们对事物的认识或事物自身客观情况发生变化时，词条内容能否及时修改以符合实际情况。如果词条未能及时修改，则意味着其内容在准确性或完整性方面有所欠缺。因此，可以认为及时性是准确性或完整性在时间维度上的映射。在具体评价过程中，常采用“最近更新时间”“更新频率”等指标对其进行衡量[15][30]。

（7）一致性是指词条的格式或语义前后一致的程度，侧重对词条描述质量的衡量，通常采用“前后内容描述是否一致”来衡量。简洁性指的是词条内容描述简明扼要的程度，也侧重对词条描述质量的衡量。在具体评价过程中，两个标准的衡量都需要评价人发挥重要作用。

（8）客观性是指词条内容不包含个人主观偏见、中立的程度。相关性指的是词条内容与表达主题相关联的程度[31]。这两个标准都是对词条内容的衡量。在具体评价过程中，客观性和相关性一般需要评价人基于自身专业知识和经验进行判断。

基于以上分析，可以看出关于网络百科词条质量的评价研究是从词条内容自身和词条描述两个方面展开评价。词条内容质量可以概括为词条内容与其要反映的客观事实相对应的程度，评价标准主要包括：准确性、完整性、及时性、客观性与相关性。词条描述质量是对语言组织、图表配置以及参考文献提供等情况的衡量，能够反映出用户对词条质量的感知情况，评价标准主要包括可读性、可验证性、权威性、一致性与简洁性。结合以往信息质量评价研究的成果[32]，可以

认为此类研究的思路包括“数据”和“用户”两种视角。此外，评价指标包括直接指标和间接指标两类。直接指标对相应标准的衡量更加准确，但严重依赖评价人的专业水平，普适性差。间接指标对评价人的依赖程度低，普适性强，但准确性差。因此，未来研究在选择指标时，需要进行综合考虑。

3　网络百科词条质量自动评价研究

3.1　评价结果与评价对象

这类研究的文献数量为 65 篇，占总体的 59.6%。评价对象以维基百科为主，并且绝大多数为英文版本。研究的样本数量均值为 13720，极差为 168183，标准差为 31436.1。其中，样本数量超过 1000 的文献数量占总数的 56.9%。此外，部分研究因未进行实证或未对评价结果、评价对象、样本数量予以明确，因此对这些未明确的内容在表中用“无”表示，如表 6 所示。

表 7 展示的是词条质量评价结果分布情况。在分类数量方面，二分类研究数量最多，占比 46.2%；五分类研究数量最少，占比 1.5%。值得注意的是，分类数量相同的研究在类别划分上存在一定差异。以二分类为例，类别划分主要包括特色词条与非特色词条[33]、高质量词条与低质量词条[34]、高质量词条与一般词条[35]。准确率是衡量分类评价效果的主要指标，以往进行实证的研究都会计算其数值，以展现研究水平。本研究对准确率数值进行汇总，得到不同分类数量对应准确率的均值和最高值。可以看出，二分类研究的准确率最高，达到 97%；五分类研究的准确率最低，为 78.6%。综合来看，准确率与分类数量基本成反比，即随着分类数量增多，准确率降低。值得注意的是，准确率与文献数量基本呈正比，在一定程度上反映出研究越多，评价准确率提升越大。

表 6　评价对象与样本数量

	文献数量（篇）	百分比（%）
评价对象		
英文版维基百科	46	68.7
泰语、法文等版维基百科	15	22.4
百度百科	1	1.5
无	5	7.5
样本数量		
<100	6	9.2
100~1000	11	16.9
1001~10000	23	35.4
>10000	14	21.5
无	11	17.0

表 7　词条质量评价结果分布

分类数量	文献数量（篇）	文献数量占比（%）	分类准确率（%）	
			均值	最高值
2	30	46.2	89.0	97.0
3	6	9.2	81.0	92.0
4	2	3.1	84.0	85.0
5	1	1.5	78.6	78.6
6	10	15.4	75.6	88.1
无	16	24.6	–	–

3.2　评价方法分析

在网络百科词条质量自动评价研究中，评价方法包括技术方法、评价标准和评价指标。在技术方法方面，61.5% 的研究采用了机器学习方法。在具体的机器学习过程中，算法的选择多样，主要包括 C4.5 决策树[36]、SMO 算法[37]、随机森林算法[38]、最大熵模型[7]等。在评价标准方面，评价标准数量在 3 以内的研究占比 67.7%，6 以内的研究占比 95.4%。表明此类研究在评价标准数量方面差异不大。表 8 展示了此类研究中采用的评价标准及对应指标，括号中的数字代表文献数量。可以看出，此类研究在评价标准方面的数量分布相对集中。

综合以上两方面，此类研究在网络百科词条质量认知以及从哪些方面进行评价也已达成一定共识。对比表 5 和表 8，结合评价标准类型和涉及文献数量来看，两类研究采用的评价标准基本相同，可以认为两类研究在网络百科词条质量认知以及评价思路方面基本一致。但与第一类研究相比，第二类研究没有评价人介入，那么其评价标准如何衡量，评价指标又有何特点呢？

首先，与第一类研究相比，评价标准数量增加了 4 个，具体为“稳定性”、“可用性”、“可信性”与“增殖性”。其中，稳定性表示的是词条内容在一段时间内稳定不变的情况。采用这一标准的研究认为，词条内容在短时间内不应有剧烈变动。如果有，则很可能存在破坏行为。在实际测量方面，一般采用“版本回滚次数”来衡量[39]。可用性是指词条内容可以被用户利用的情况，侧重于词条内容价值的表达。可能由于测量难度太大，仅有 1 篇研究使用“浏览量”来度量[40]。但浏览量的影响因素非常多，不仅涉及可用性，还与词条所属主题等因素有关，因此通过浏览量来反映可用性存在较大偏差。可信性与增殖性出自同一篇研究[41]，其中可信性表达的是词条内容可以使用户信服的程度，与权威性类似；增殖性指的是词条在传播使用过程中产生新价值及附加价值的潜力。与可用性相同，增殖性侧重的也是词条内容价值的表达，但现有研究缺乏具体的评价指标。

表 8　第二类研究的评价标准与评价指标

（单位：篇）

序号	评价标准	评价指标	序号	评价标准	评价指标
1	完整性（44）	文字数量（35） 段落数量（12） 编辑次数（11） 编辑者数量（10） 句子数量（5） 名词所占比例（4） 事实数量密度（2） 词条年龄（2） 积极编辑次数（1）	8	权威性（41）	注册编辑者比例（11） 编辑者编辑次数差异（11） 外链数量（8） 编辑者协作网络中心度（6） 编辑者所编辑版本留存时间（6） 参考文献质量（4） 外链密度（1）
2	可读性（31）	图片数量（18） 内链数量（13） 是否存在信息框（9） 各类词所占比例（4） 表格数量（3）内链密度（2） 图片密度（1） 拼写错误数量（2） 语法错误数量（1） 注释数量（1） 段落长度的标准差（1） Flesch–Kincaid grade level（5） Automated readability index（3） Flesch reading ease（2） Coleman–Liau index（2） Fog scale level（2） SMOG Index（2） Forecast readability（1）	9	准确性（21）	编辑次数（9） 编辑者数量（9） 讨论数量（4） 词条年龄（4） 积极编辑次数（1） 管理员编辑次数（4） 注册编辑者编辑次数（1） 管理员数量（1） 编辑者编辑行为类型比例（1）
3	可验证性（16）	参考文献数量（14） 搜索引擎结果排名（4）	10	及时性（11）	更新频率（7） 最近更新时间（4）
4	一致性（8）	管理员编辑次数（4） 不同版本差异程度（4） 管理员数量（1）	11	稳定性（4）	版本回滚次数（4） 版本回滚间隔时间的中值（3） 页面小修改比例（1）
5	客观性（4）	无具体评价指标	12	相关性（3）	无具体评价指标
6	简洁性（2）	无具体评价指标	13	可信性（1）	无具体评价指标
7	可用性（1）	词条每天浏览量（1）	14	增殖性（1）	无具体评价指标

其次，与第一类研究相比，评价指标更多的是间接指标。对于“准确性”而言，在第一类研究中，“错误数量”是采用最多的指标。而在第二类研究中，“编辑次数”“编辑者数量”是使用最

多的指标，其背后的逻辑是，编辑者数量越多，编辑次数越多，错误产生的可能越小。在“完整性”测量方面，句子数量和段落数量可以说是文字数量的扩展，都代表了词条内容数量。“名词所占比例”是指名词在所有词类中的比重，在一定程度上反映了事实数量密度[42]。这一指标与第一类研究中“遗漏事实数量”的含义相反，即名词所占比例越大，词条内容完整性越强。“编辑者数量”和“编辑次数”两个指标背后的逻辑是，编辑者数量越多，编辑次数越多，词条可能涵盖的内容越多，词条内容完整性越强。在“权威性”测量方面，“外链数量”和“外链密度”与第一类研究中的“搜索引擎排名”逻辑一致[43]，具体是指该词条被其他网页引用越多，词条越权威。“注册编辑者比例”“编辑者编辑次数差异”“编辑者协作网络中心度”“编辑者所编辑版本留存时间”都是编辑者权威性的体现[44]，背后的逻辑是，编辑者越权威，词条越权威。在“一致性”测量方面，“管理员编辑次数”和“管理员数量”是间接指标，反映了管理员参与词条编辑的程度，其背后的逻辑是，管理员参与度越高，词条存在分歧的可能性越大，词条一致性越差[33]。在可读性、可验证性与及时性测量方面，第二类研究与第一类研究采用的主要指标基本一致。

网络百科通常都制定了基于用户评价的词条质量评价机制。以维基百科为例，词条质量分为特色、甲级、优良、乙级、丙级、初级和小作品七类。这种评价机制的存在，不但可以识别出优秀词条供用户更好地使用，也可以对相应的用户进行褒奖激励，还可以识别出问题词条以更好地编辑修改。但相对于词条的巨大规模，人工评价分类的效率太差，而且由于用户差异较大，主观性较强，分类误差较大。因此，网络百科对词条进行自动评价势在必行。

综合以上分析，可以发现此类研究基本上也是从“数据”视角和“用户”视角出发而展开。其中，数据视角下的评价标准包括完整性、准确性、及时性、稳定性、客观性、相关性、一致性、增殖性；用户视角下的评价标准包括权威性、可读性、可验证性、简洁性、可用性、可信性。与第一类研究不同，因没有评价人的介入，评价指标以间接指标为主；部分标准的评价指标存在交叉，如“准确性”和“完整性”的衡量都用到了“编辑次数”和“编辑者数量”两个指标；超过 1/3 的评价标准并未给出对应评价指标。因此，评价标准如何被间接指标准确衡量，是值得探究的问题。

4 总 结

以维基百科和百度百科为代表的网络百科已成为网络时代重要的知识源，其词条质量的重要性不言而喻，对其词条质量进行评价研究兼具实践与学术双重意义。既有的对网络百科词条质量评价研究的情况与特点可以归结如下。

（1）在网络百科词条质量人工评价研究方面，首先，现有研究结果显示网络百科词条质量整体偏低；其次，词条质量评价结果不一致的情况具有一定普遍性，其背后的原因可能与评价范围、样本数量等方面的差异有关；再次，评价对象较为单一，评价范围相对集中，样本规模较小，网络百科词条质量尚缺乏全面评价；最后，因为网络百科词条始终处于动态变化的状态，因此，此类研究还会持续下去，但研究采用的对比资源应该尽量一致，评价标准、评价范围和样本数量应该相差不大，只有如此，才能保证评价结果的权威性。

（2）在网络百科词条质量自动评价研究方面，评价准确率与分类数量基本成反比，相对于分类数量、样本数量以及研究投入，技术方法和评价指标的选择对评价准确率的影响可能更为关键。因此，未来研究应该两条腿并行。一方面积极探索先进技术方法，特别是对机器学习算法的深入研究；另一方面，继续加深对网络百科词条质量内涵的研究。在指标选择时，综合考虑准确性和可操作性。

【参考文献】

［1］MORELL M F. Good faith collaboration: The culture of Wikipedia［J］. Journal of Communication, 2013(1):146-147.

［2］READ B. Middlebury college history department limits students' use of Wikipedia［J］. Chronicle of Higher Education, 2007,53(24): A29.

［3］GILES J. Internet encyclopedias go head to head［J］. Nature, 2005(7070):900-901.

［4］LAVSA S M, CORMAN S L, CULLEY C M, et al. Reliability of Wikipedia as a medication information source for pharmacy students［J］. Currents in Pharmacy Teaching and Learning, 2011,3(2):154-158.

［5］CLAUSON K A. Scope, completeness, and accuracy of drug information in Wikipedia［J］. The Annales of Pharmacotherapy, 2008, 42(12):1814-1821.

［6］李欣奕 . 网络百科条目质量评价研究［D］. 长沙 : 国防科学技术大学 , 2014:15.

［7］RASSBACH, LAURA, TREVOR Pincock, BRIAN Mingus. Exploring the feasibility of automatically rating online article quality［C］. Proceedings of the 2007 International Wikimedia Conference (WikiMania), Taipei, Taiwan. 2007，66.

［8］Wikipeida. Wikipedia: Content assessment［EB/OL］.［2019-11-13］. https://en.wikipedia.org/wiki/Wikipedia:Content_assessment.

［9］Chi E H , PENDLETON B , SUH B . Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie［C］// 2010.

［10］KOPAK R. On the measurability of information quality［J］. Journal of the American Society for Information Science and Technology, 2014(1):89-99.

［11］黄令贺 , 张宁 . 网络百科词条质量影响因素研究综述［J］. 情报理论与实践 , 2019,42(7):171-176.

［12］ZHANG Y, SUN Y, XIE B. Quality of health information for consumers on the web: a systematic review of indicators, criteria, tools, and evaluation results［J］. Journal of the Association for Information Science and Technology, 2015(10):2071-2084.

［13］中国社会科学院语言研究所词典编辑室 . 现代汉语词典 第 7 版［M］. 2016: 209.

［14］中国社会科学院语言研究所词典编辑室 . 现代汉语词典 第 7 版［M］. 2016: 1546.

［15］CABRERA-HERN á ndez L M, WANDEN-BERGHE C, CASTRO C C, et al. The presence and accuracy of food and nutrition terms in the Spanish and English editions of Wikipedia: in comparison with the Mini Larousse encyclopedia［J］. Nutricion Hospitalaria, 2015,31(1):488-493.

［16］WOLFE D. Accuracy and completeness of drug information in Wikipedia: A comparison with standard textbooks of pharmacology［J］. Amia Annu Symp Proc, 2008,6(9):912.

［17］CASEBOURNE I, DAVIES C, FERNANDES M, et al. Assessing the accuracy and quality of Wikipedia entries compared to popular online encyclopedias: A comparative preliminary study across disciplines in English, Spanish and Arabic［R］. University of Oxford, 2012.

［18］NICHOLSON, DAREN T . An evaluation of the quality of consumer health information on Wikipedia［J］. Iwmi

Books Reports, 2006, 232(3):685–689.

[19]FRIEDERIKE O. Does Wikipedia provide evidence–based health care information? A content analysis [J]. Zeitschrift Für Evidenz Fortbildung Und Qualitt Im Gesundheitswesen, 2008,102(7):441–448.

[20] 丁敬达 . 维基百科词条信息质量启发式评价框架研究 [J]. 图书情报知识 , 2014(2):11–17.

[21]DEVGAN L, POWE N, BLAKEY B, et al. Wiki–Surgery? Internal validity of Wikipedia as a medical and surgical reference [J]. Journal of the American College of Surgeons, 2007,205(3): S76–S77.

[22]RAJAGOPALAN M S, KHANNA V K, LEITER Y, et al. Patient–oriented cancer information on the internet: a comparison of Wikipedia and a professionally maintained database [J]. Journal of Oncology Practice, 2011,7(5):319–323.

[23]RANSBOTHAM S, KANE G C. Membership turnover and collaboration success in online communities: Explaining rises and falls from grace in Wikipedia [J]. Mis Quarterly, 2011,35(3):613–627.

[24]ROSENZWEIG R. Can history be open source? Wikipedia and the future of the past [J]. Journal of American History, 2006,93(1):117–146.

[25]CZARNECKA–KUJAWA K, ABDALIAN R, GROVER S C. The quality of open access and open source Internet material in gastroenterology: is Wikipedia appropriate for knowledge transfer to patients? [J]. Gastroenterology, 2008,134(4):325–326.

[26]KOROSEC L, LIMACHER P A, Lüthi H P, et al. Chemical information media in the chemistry lecture hall: A comparative assessment of two online encyclopedias [J]. Chimia International Journal for Chemistry, 2010,64(5):309–314.

[27]REAVLEY N J, MACKINNON A J, MORGAN A J, et al. Quality of information sources about mental disorders: a comparison of Wikipedia with centrally controlled web and printed sources [J]. Psychological Medicine, 2012,42(8):1753–1762.

[28]HOLMAN Rector L. Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles [J]. Reference Services Review, 2008,36(1):7–22.

[29]LAURENT M R, VICKERS T J. Seeking health information online: does Wikipedia matter? [J]. Journal of the American Medical Informatics Association, 2009,16(4):471–479.

[30]AZER S A. Evaluation of gastroenterology and hepatology articles on Wikipedia: are they suitable as learning resources for medical students? [J]. European Journal of Gastroenterology & Hepatology, 2014,26(2):155–163.

[31] 金燕 . 基于用户体验的协同内容创建系统质量保证措施——以百度百科为例 [J]. 情报理论与实践 , 2016,39(3):6–9.

[32] 刘冰 . 基于用户体验视角的信息质量反思与阐释 [J]. 图书情报工作 , 2012, 56(6): 76–80, 91.

[33]FERRETTI E, SORIA M, CASSEIGNAU S P, et al. Towards information quality assurance in Spanish: Wikipedia [J]. Journal of Computer Science and Technology, 2017,17(1):29–36.

[34] 仝召娟 , 许鑫 . 百度百科网页质量的自动化评价 [J]. 信息资源管理学报 , 2015(2):65–71.

[35] 陈学平 . 基于维基百科的 Web 网页数据质量评估系统 [D]. 南京 : 南京邮电大学 , 2014.

[36] 裘江南 , 翁楠 , 徐胜国 . 基于 C4.5 的维基百科页面信息质量评价模型研究 [J]. 情报学报 , 2012,31(12):1259–1264.

[37]DE LA CALZADA G. A strategy oriented, machine learning approach to automatic quality assessment of Wikipedia articles [D]. State of California: California Polytechnic State University, 2009.

[38]COZZA V, PETROCCHI M, SPOGNARDI A. A matter of words: NLP for quality evaluation of Wikipedia medical articles [C]//Proceedings of the 13th international conference on Web Engineering, Aalborg, Denmark, Springer, 2016: 448–456.

[39]STVILIA B, TWIDALE M B, SMITH L C, et al. Assessing information quality of a community–based

Encyclopedia［C］//Proceedings of the 2005 International Conference on Information Quality, Maryland, USA, ACM, 2005: 442-454.

［40］DALIP D H, GONÇALVES M A, CRISTO M, et al. On multi-view based meta-learning for automatic quality assessment of wiki articles［C］//Proceedings of the Second International Conference on Theory and Practice of Digital Libraries, Heidelberg, Germany, ACM, 2012:234-246.

［41］金燕 , 周婷 , 詹丽华 . 基于层次分析法的协同内容创建系统质量评价体系研究——以百度百科为例［J］. 图书馆理论与实践 , 2015(7):41-45.

［42］XU Y, LUO T. Measuring article quality in Wikipedia: Lexical clue model［C］//2011 3rd Symposium on Web Society, Port Elizabeth, South Africa, IEEE, 2011:141-146.

［43］SOONTHORNPHISAJ N. Assessing the quality of Thai Wikipedia articles using concept and statistical features［M］//New Perspectives in Information Systems and Technologies. 2014:513-523.

［44］PETROCCHI M. Maturity assessment of Wikipedia medical articles［C］//Proceedings of the 2014 IEEE 27th International Symposium on Computer-Based Medical Systems, Vancouver, Canada, IEEE, 2014:281-286.

Review of the Evaluation of Article Quality in Internet Encyclopedia

HUANG Linghe　LI Mingze　YU Jinping

(Hebei University Management College, Baoding 071002,China)

Abstract: **[Purpose/significance]** This paper reviews the previous researches of the evaluation of articles quality in internet encyclopedia, summarizes the basic situation, how these researches have been carried out, and the problems. **[Method/process]** Using the method of literature retrieval, qualitative and quantitative analysis, the paper reviews the researches of the evaluation of internet encyclopedia from the aspects of manual evaluation and automated evaluation. **[Result/conclusion]** Compared with authoritative information resources, the quality of online encyclopedia entries is generally low. The inconsistence of the evaluation results of term manual evaluation are universal. The accuracy of automated evaluation of entry quality basically inversely proportional to the number of categories. Although the purpose of research is different, the cognition and evaluation ideas of entry quality are basically consistent in both types of research.

Keywords: Internet encyclopedia; Manual evaluation; Automated evaluation; Evaluation standard; Evaluation index

（本文责编：王秀玲）