数据库的运用与比较:以提升学术评价质量

蔡明月
(台湾政治大学图书资讯档案研究所所长、教授)

10 月 27 日上午,台湾政治大学图书资讯档案研究所所长蔡明月教授作了题为"数据库的运用与比较:以提升学术评价质量"的学术报告。

蔡教授首先论述了科研绩效在学术评价中的重要地位。科研文献的出版是科研绩效的基础,它们可以根据论文数量与质量、期刊等级和合作研究的规模来衡量。而科研绩效的评价是世界大学评价的重要标准之一。在英国《泰晤士报高等教育》和上海交通大学的世界大学排名体系中,与论文产出、被引情况相关的指标的总权重均不低于 60%。目前的宏观学术评估倾向于采取定量和统计方法,尤其是在情况上发表的论文,其质量和所谓的影响力主要取决于他们发表的期刊的权威程度。另一方面,期刊的影响力,主要是根据其在 JCR、SCIE、SSCI、CSSCI 中的影响因子来确定。而上面提到的索引在社会科学和人文科学领域的评价中没有被普遍采用。

蔡教授以 SCIE、JCR、Scopus 为例,介绍和比较了一些重要的引文数据库等。来源索引数据库的功能是主题索引,其目标为客户提供一个给定学科的全面的文献检索。除了 SCIE 以外,还存在大量的来自不同学科的主题索引数据库。通常情况下,主题索引数据库拥有比 SCIE 更大的期刊收录量。而一篇文章需要先被收录,获得能见度,然后才能获得被引用的机会。当启动一项新的研究时,文学评论是必不可少的。为了遵守全面、综合的原则,以及避免重复已有的研究,最有效的检索和确认工具是主题索引数据库。考虑到在全球的知名度,被编入索引主题索引数据库中的文章也是一个重要的指标。作者生产力的定量和定性的调查,都需要参考书目和引文来源,如主题索引数据库和引文索引数据库等工具。由于不同数据库的政策、类型、格式、时间和空间都不相同,准确性和完整性需要予以考虑。合作作者权重的计算、通讯作者和所属机构也需要进一步明确。引文数据可以作为评价作者学术成果的直接和透明的标准。引文索引数据库能提供个人发文量、作者总被引频次、合作排名、引用文章数量、引用作者、引用文章、自引和 h 指数等数据。然而,期刊影响因子的应用是有局限性的,期刊引文索引是期刊的索引,而不是论文的索引。因此,在试图揭示引文全球竞争力时,论文引文索引的检索系统是有必要的。当涉及到"被引"时,引文分析存在的争议和问题的必须纳入考虑范围。

此外,蔡教授还指出了学术评价中存在的一些其他问题。第一,各研究机构的评价指标存在差异。在 2006 年出版的《差异的世界:大学排名的全球调查》(A World of Difference: A Global Survey of University League Tables)中,加拿大教育政策研究院研究了世界各地进行大学评价的 18 个研究机构,并提出了一个修正模型,包括 7 项指标:新生质量、教育投入——师资、教育投入——资源、教育产出、最终成果、科学研究和声誉。每个指标又包含多个

更详细的二级指标,如“科学研究”指标包含 21 个二级指标。第二,科研成果的类型包括期刊论文、专著、会议论文、专利、评论、艺术作品、网络出版物等多种形式。期刊论文已被普遍采用以评估学术生产力,但这可能会导致低估整体科研产出从而未能反映整体学术能力。虽然开放存取文献将成为未来的趋势,但其中很大一部分,如机构库文献,在网络环境中被忽视。第三,自然科学的评价和社会科学的评价之间存在差异。期刊论文已成为自然科学领域的关注的焦点之一。然而,对于人文科学和社会科学来说,研究范围的完整性和研究框架才是他们的主要关心的问题,该领域的学者倾向于以专著和研究报告的形式发表他们的研究成果。文学经典的翻译和注释的人文科学研究的基础,而这些成果未编入任何引文索引数据库。

最后,蔡教授总结了以下结论:首先是学术评价的复杂性,学术评价是从来都不是一件简单的事,需要考虑不同的阶段和因素,包括教学,科研,社会服务,学术资源等等。其次是引文分析具有假设前提,即引用一篇文献意味着使用该文献,反映了该文献的价值,引文与所引用相关,并且所有的引用都是平等的。第三,引文分析存在多种问题,包括多重作者、自引、同形异义字、同义词、数据源类型、隐参考文献、时间的推移波动、场变化、错误等。总之,基于引文分析的研究应当认真细致开展,而引文分析的局限、假设和存在的问题也必须充分考虑。

(曾 倩 翻译整理)

(上接第 6 页)

然无法跟上数据的增长速度。这些环境都催生了数据库知识发现或数据挖掘的出现。当我们处理科学数据的时候,一定要注意所有已知的事实都要纳入考虑。并且,在科学里,罕见的事实通常是至关重要的发现,不能和错误或噪声混为一谈。鲁索教授认为知识表示是情报学的核心使命。一个简单的知识表示的例子是图书馆目录。图书馆的众多工具如关键词表、同义词表、分类法和本体都是知识表示的形式。知识表示扮演着五种角色,它是一种代表(surrogate),一种本体选择,一种智能推理的偏好理论,一种有效计算的媒介和一种人类表达的媒介。

大数据技术发展如此之快,信息计量学家该如何回应呢?尽管人们已经知道数据挖掘和文献计量学尤其是引文分析的关系有一段时间了,但是将引文分析用于数据挖掘还是今年来才有的事。一种较为启发性的方法是用于链接预测技术来探测错漏的链接,另外在大数据领域引文分析可用于探测热点主题。通常来说可视化技术有助于数据的可视化,其中一项最近的研究是 Rafols 等人对交叉(overlay)的利用,近来的数据挖掘技术包括 Prabowo&Thelwall 的语义分析和概念挖掘,其中单词和他们所代表的概念相关联。由于信息计量学领域可定义为任意形式(不限于记录或书目)和任意社会群体(不限于科学家)的信息的定量方面的研究,所以所有形式的数据挖掘自然属于这个领域的研究范畴。

鲁索教授最后给出了一些很有启发性的结论。首先,大数据是个巨大挑战,其存在导致了数据越多反而越难获得的矛盾,因为要找的特定信息可能淹没在 TB 级的数据里。然而数据挖掘处理的方法和结论为这个领域的研究提供了广阔的前景,因为大数据促使更多人研究其社会和机构背景。其次,分析科学中的数据(或者也包含引文信息)有助于我们理解知识是如何获取的,知识随时间是如何变化的。数学模型在分析大数据时可能会派上用场,当模型很好地拟合时,可用于预测目的并用于指导政策制定。最后,大数据改变了知识的定义,刻意追求准确和客观可能会是误导性的。在大数据背景下,数据越多并不意味着更好。大数据一旦脱离了实际情景,就失去了意义。

作为一名对引文网络分析情有独钟的情报学家,鲁索教授希望信息计量学领域能在知识发现和数据挖掘里发挥应有的作用,带来一些激动人心的发展,减少科学研究的瓶颈并做出真正创新性的应用。

(余厚强 翻译整理)