

文献计量与内容分析^{*}

——文献群中隐含信息的挖掘

朱少强 邱均平

武汉大学信息管理学院 武汉 430072

〔摘要〕在对文献计量和内容分析进行特征归纳与比较的基础上,提出这两种研究方法所具备的一致性逻辑,即:通过借助于某种直观的或选定的形式化体系及定义在此形式化体系之上的运算集,对文献的某些外部特征与粗略内容特征进行量化统计,力图发现附着在大量文献群背后的隐含信息;探讨这两种方法在应用中必须满足的、隐含的预设前提;认为文献计量和内容分析都可以归并到信息计量与分析的类属下,发挥各自优势,进行理论、方法与应用的综合。

〔关键词〕文献计量 内容分析 隐性知识 形式化 比较研究

〔分类号〕G350

Bibliometrics and Content Analysis——Finding Latent Semantics behind Documents

Zhu Shaoqiang Qiu Junping

School of Information Management, Wuhan University, Wuhan 430072

〔Abstract〕Based on a comparative study of bibliometrics and content analysis, this paper presents the coherent relationship between these two methods, which aim to find some latent semantics by mapping vast documentary information into a formalized symbol system, namely indices, with a set of algorithms defined over the symbol system, and by statistic of those indices. The paper also explores the implicit prerequisites which are held by the two methods. It is believed that bibliometrics and content analysis can both be contained within the framework of informetrics or information analysis, and that synthesized theory, methods and applications are likely to emerge in near future.

〔Keywords〕bibliometrics content analysis tacit knowledge formalization comparative research

1 引言

文献计量和内容分析都是常见的科学研究及情报研究方法。文献计量起源于对科技文献数量特征的考察;内容分析最初是传播学研究中特有的方法,后来被推广到几乎所有社会科学领域,也少量地应用于对科技文献的分析。近年来,随着文献计量、内容分析各自理论与方法体系的成熟,其应用也在不断扩展,并有相互交叉融合的趋势,表现在对于某些特定的研究案例,无法简单地判定其究竟是属于文献计量方法还是内容分析方法。有人认为,内容分析是文献计量的一个分支,只是各自侧重点不同,但这种说法并未得到广泛认可。因此,如何认识这两种重要的研究方法的本质、相互关系以及两者整合应用的可能性,将是本文所力图探究的问题。

2 文献计量法与内容分析法之特性比较

2.1 概念

传播学家伯纳德·贝雷尔森给出了对内容分析法的权威定义:“一种对具有明确特征的传播内容进行的客观、系统和定量的描述的研究技术”。

本文作者之一——邱均平教授对文献计量学有如下定义:“以文献体系和文献计量特征为研究对象,采用数学、统计学等的计量方法,研究文献情报的分布结构、数量关系、变化规律和定量管理,并进而探讨科学技术的某些结构、特征和规律的一门学科”。

2.2 相似性

客观、系统和定量的特征描述,不仅适用于内容分析法,对文献计量法也同样适用。另外,笔者认为它们在方法上都有较为简单、直观、粗略和间接、非接触的特点。

^{*} 本文系教育部重点研究基地重大项目“文献计量与内容分析的综合研究”(项目编号:02JAZD870003)研究成果之一。

收稿日期:2005-03-14

2.2.1 客观 用事实和数据说话,是两者客观性的主要表现。所分析的对象,对于文献计量来讲,是十分显著的文献外部特征;对于内容分析来讲,则是有明确特征的传播内容。它们都从不凭空推测分析对象背后的可能含义,而依赖于固有的分析程序来得出结果;一旦研究目的与范围确定,就要尽量排除人为因素的影响,做到客观、无偏向;所选择的类目体系应科学,具有可验证的合理性;同样的对象和程序如果由不同的人来进行,应能得到相同的结果;对于编码后得出的统计数据,一般还要进行信度检验,确保数据偏差在正常范围内。

2.2.2 系统 一般而言,文献计量与内容分析的对象都是大量的、系统化的、具有一定历时性的文献;都要面对如何确定调查范围和取样的问题。系统化调查取样是进行数据统计的基本前提,必须有足够的数据来克服可能出现的随机偏差。除语言符号分析等特殊情形之外,单个的、少量的文献通常不能作为分析的依据。

2.2.3 量化 即都涉及某些定量化过程,通过将文献特征表示成一些数量指标来进行统计和推测。文献计量以几个经验定律为核心,直接对一个个的文献外部特征等予以计数,所使用的数学模型略微复杂一些;内容分析大多采用人工判读的方式将分析单元归入某个类目,进行简单的计数统计,再加以比较分析。

2.2.4 简单、粗略和直观 表现在分析单元的选取上,都比较粗略。文献计量中采用的一些计量指标以简单的文献数量作为核心,如文献增长率、老化率、借阅流通率、被引次数、科学生产率等。它并不关心更细致的内容之间的差别。内容分析虽然深入到了文献内容的层次,但也只限于“具有明确特征”的内容,如判读电视广告中的某个人物形象是否具有女性歧视意味等。在数学处理上,内容分析以直观的线性计数与统计为主;文献计量则用到了几个复杂一点的经验定律,也仍然以验证性的应用为主。与现代数据挖掘中所用到的某些复杂算法相比,文献计量与内容分析的量化处理具有简单、易用、明了的特点。

2.2.5 非接触 它们都是通过二手资料进行的间接、非接触式的研究方法,这一点与社会调查、访谈、实验等研究方法有着根本的不同。

2.3 差异性

体现在以下几个方面:

2.3.1 分析对象 文献计量的指标以文献的外部特征为主,故只适用于有实体形态的科学文献。而内容分析不仅适用于对科学文献内容的分析,而且也可用于非科学领域的一般文献的分析以及无文献实体的语言传播的分析,甚至可推广到动作、行为等非语言内容的分析。

2.3.2 应用领域及学科归属 文献计量学是图书馆学情报学的特殊研究方法,主要用于对科学文献的研究,对各门学

科都适用。内容分析发源于传播学研究,现已推广到社会科学各个领域,如政治、军事、教育等,在科技领域的应用则比较少。目前还很难说内容分析到底是哪一个学科的专有方法。

2.3.3 分析粒度与指标体系 所谓粒度,是指所选取分析单元的大小。显然,注重文献外部特征的文献计量法,其分析粒度要比注重文献内容特征的内容分析法粗。文献计量法中,量化的核心是以件、篇、个为单位计算的实体数目,其他所有计量指标如增长率、老化率、流通率、引文率、被引率等,都是在此基础之上的衍生。因此,文献计量法中的指标体系具有一元性特征,指标稳定且系统化。内容分析大多采用类似“有、无”的二度指标,或类似“正面、负面、中立”的三度、五度指标,少数采用七度指标,在此基础上计算各度指标出现的频次。不同的是,文献计量的类目体系是直观的,文献、作者、机构、期刊、学科、引文,都一目了然且是有限的。而内容分析所依据的类目体系则十分不稳定、不系统,经常要根据研究目的的不同而自行设定。类目体系的确立可以说是内容分析法的关键,但它却是一个相当主观的、人为的、定性的过程。所以内容分析法的指标体系具有多元性特征,不存在一个统一而稳定的指标系统。

2.3.4 量化程度 内容分析对数学方法的运用还处于初级阶段,属于半定量化的研究方法。相比而言,文献计量的量化程度更高一些,数学上的处理更为成熟。

2.3.5 目标 按照邱均平教授对文献计量学的概念定义,文献计量的目标在于“研究文献情报的分布结构、数量关系、变化规律和定量管理,并进而探讨科学技术的某些结构、特征和规律”,可见其主要致力于对文献群实体自身运动变化规律的研究,谋求改进文献信息的管理,提高文献信息交流效率。至于通过文献计量对科学技术规律加以推断,其实是间接的和辅助性的。这本是文献计量学的原义。而按照传播学研究的通常说法,内容分析的目标可以归纳为三点:①将传播内容与社会现实进行比较;②推断信息传播者的态度、倾向;③推断传播效果。可见,内容分析法从不以其所依据的文献信息本身作为研究目的。

3 形式化及其演算:内在的一致性逻辑

通过以上对比,笔者基于文献计量与内容分析法的相似性推断,这种相似性有可能反映了它们的某种共同本质,即有可能采用某种一致性的程序框架予以承载。这种框架可以描述为:借助某种形式化体系及其运算集,将客观实在经过简化映射为抽象的概念符号,从而可以量化运算与推导,发现隐含于大量文献群背后的客观知识。至于两者的差异性问题,笔者将其归结为该框架中的某些可控变量及由此变量变化而引致的结果。

3.1 文献群中隐含信息的挖掘——目的一致性

文献计量与内容分析所据以分析的对象本身是十分明确的,即文献外部特征或具有明显特征的传播内容。其目的则是通过分析得出之前我们所不知道或无法确证的某个隐含事实,例如某学科文献信息量近十年来的增长趋势、小说人物形象中的种族歧视意味等。这些信息不可能事先以显性文本的方式表达出来,而只能通过对大量文献群的定量分析才能予以事后揭示。在时间线上,这种定量研究只能是借助于文献这种二手资料所进行的事后分析;因为在此之前,隐含信息尚未完整地生成,任何分析都是接近无效的。

若将决定文献信息分布与变化的社会存在视作一个隐含信息表达的“主体”,则整个过程可以视作一个完整的语义生成、表达、传递与获取的过程:以待分析的文献群整体即“世界3”作为载体,作为“世界1”的社会存在得以表达和反映,最终为“世界2”所感知。对这一过程可略作如下描述。

3.1.1 隐含语义的生成 根据因果律,既然待分析文献信息呈现出一定的分布、变化状态,而不是另外一种状态,则必有决定该状态呈现的动因。假如某报纸记者中存在种族歧视观念,则极可能在报纸内容中表达出来,使得对黑人的负面报道远远多于正面。同样,如果某个学科正处于新兴阶段,则关于该学科的文献数量就会急剧增长。这种决定大量文献信息呈现状态背后的社会存在,即为隐含语义生成的源泉。语义生成属于“世界1”的范围。

3.1.2 隐含语义的表达 由于世界联系的普遍性和多样性,作为原因的某个事件可能引发多个不同的后果,乃至发生一系列的连锁反应。也就是说,来自“世界1”的隐含语义在生成之后可能沿多个路径表达和传播。但对于文献计量和内容分析方法而言,它们所关心的仅仅是其中的一个路径,即社会存在决定了文献信息分布所呈现的某个特定状态。隐含语义通过影响“世界3”而得以表达。值得注意的是,语义并非通过“世界3”中的实际存在的文本得到显性表达,而是通过文献自身有规律的分布来隐含地表示。

3.1.3 隐含语义的传递 因为语义是隐含的、未知的,故我们通过文献计量、内容分析等方法来解读该语义,其实质是通过研究文献信息自身的分布规律来进行间接的推断。

3.1.4 隐含语义的获取 即语义解读完毕,为人们所认知从而进入“世界2”的过程。此时,该隐含语义可能获得显性表达,形成客观知识文本,并以新的文献实体形式进入“世界3”。

3.2 形式化演算的基本步骤——方法一致性

以下框架可以作为文献计量与内容分析以及几乎所有定量分析过程的一致性描述。为此,需要用一些较为抽象的名词术语,如“粒度”、“实体”、“形式化体系”、“符号化”等来予以容纳。

3.2.1 确定分析对象与范围 首先,要明确研究目标,即对

问题或假设做清楚明白的表述,避免将一些无关因素扯入其中;其次,要划定研究范围,对问题所涉及的时间域、空间域、主题域、收集资料的范围等作进一步限定;最后,按照研究目标与范围的要求搜集资料。如果样本总量太大,则采取一定的方法进行抽样。这些准备程序实际上对于任何信息分析流程来说,都是一样的。

3.2.2 选取分析粒度 粒度粗细主要是指分析单元的大小。分析单元可以是整个文献层次上的,如文献的一切外部特征以及单篇文献的主题、引文等。更小的分析单元有文章段落、句子、单词、符号,某个特定的事实、数据,或与分析类目有关的某个意思表示。选取粒度并非越细越好,也不是越粗越好,而要根据研究目标的不同来具体规定,因为不同的粒度是用来适应不同的目标的。

3.2.3 构建形式化体系 所谓形式化体系,是指对现实的物及其运动变化规律的概念抽象,并由一套符号系统予以表示。具体对于文献计量或内容分析来说,它包括三个方面:类目、指标和运算集(操作序列)。类目表达定性方面的属性和规律,如类属关系、包含关系、同一关系、并列关系、互斥关系等。指标则用以表达定量方面的属性,有统计前指标、统计后指标。运算集不仅可以包括数值运算,也可以包括归纳、演绎、判断、比较等逻辑运算。特定的运算操作只能在特定的类目和指标集上进行,不可能相互之间随意套用。类目、指标与运算集所构成的形式化体系整体,应是一个与现实状况无关、逻辑上自洽的纯理论体系,但却反映了客观事物运动规律的关键方面。

3.2.4 实体的符号化 实体是指文献群。文献群根据选定的分析粒度被拆分,形成分析单元。分析单元在概念上与预先划定的类目有着——对应关系。从实体的角度来看,“类目”可以看作是该实体在某个维度上的属性;而“指标”则是该属性的值。因为类目本身是可以用符号来表示的,因而分析单元就被在多个维度上表示为“属性-值”的形式。理论上,“值”可以取包括文本在内的所有形式,为了便于量化统计和计算,我们通常预先通过类目设计,使得该值只能取“0、1”二度指标或“-1、0、1”三度指标等数值形式。整个文献群的所有分析单元都被表示成一组“属性-值”对,就完成了实体的符号化。

3.2.5 统计汇总和运算 将多个分析单元的同一类指标值进行累计,再计算各类指标总数值之间的关系,是最常用的统计汇总运算方法之一。但实际情况可能远比这复杂,例如,每个指标值而非该类指标本身可以形成一个单独的分析维度。类目与类目、指标与指标之间可以具有复杂的层次关系和其他联系,其运算集是预定义的。就像实体的符号化是实体向类目、指标体系的形式化映射一样,对已符号化的文献群所进行的一切统计汇总和运算操作,也是向定义在该类目、指标体系之上的运算集的映射。

3.2.6 检验信度并解释结果 经过运算后所得出的结果,应至少具有形式上的一致性,这就是检验信度的过程。如果该结果是有意义的,即可视作完成了文献群中隐含信息的挖掘。

3.3 隐含前提的讨论

无论是内容分析还是文献计量,都不可避免地涉及到两个必须作为前提的隐含假设。

3.3.1 文献群中以我们可以预料的方式存在着某些隐含信息 对于客观世界某种不可知的存在事实,通过影响文献群的分布而间接建构出来的隐性知识,我们有能力通过文献计量或内容分析等形式化、程序化的分析方法予以解构和重构,从而使其为人们所认知。

应该说,万事万物有因必有果,有果必有因,这是自然的。因而,我们也完全有理由相信,文献群之所以表现出此种分布状态,而不是彼种状态,必定有其背后的动因,也就是说必定传达了某种隐含的信息。但问题在于,隐含信息透过文献信息特征进行表达的方式,未必与我们所预想的方式一致。以 $10-1=9$ 而论,我们也许以为因素-1 是使我们获得结果9的关键因素,而忽略了因素10,导致了南辕北辙。

3.3.2 形式化体系与客观存在之间存在着严格的对应关系 用概念、类目来划分事物是可行的;归入同一类目的事物具有完全相同的质,并用同样的符号来代表。所选用的指标体系则完全表达了同质事物之间一切量的区别。同时,定义于该形式化体系之上的运算集,是与该形式化体系相对应的客观事物运动规律的真实反映。

以 $1+1=2$ 为例,这只适用于两种绝对同质的事物,如两只苹果相加。苹果与苹果之间也许有好坏之分,但这种差别被我们忽略,它们被赋予绝对相同的形式并同等地参与计量。如果仔细考察文献计量或内容分析的过程,完全可以发现类似的情形。

因此,若因文献计量、内容分析或任何定量化研究方法,运用了某种逻辑上自洽的、严密的形式化体系,进行了相当复杂的数学计算或证明,得出一个精确度很高的计算结果,就认为这个结果就绝对正确,显然是没有注意到进行定量分析所必须依赖的假设前提。实际上,正如邱均平教授在《文献计量学》一书中所指出的那样,任何定量分析都只能是相对精确和客观的,所以在应用时应持慎重态度。如果忽视这些隐含前提,不恰当地运用定量分析手段,也极有可能产生南辕北辙的后果。

3.4 形式化演算过程中的可控变量

3.4.1 类目 文献计量较注重文献的外部特征,因此类目体系较为稳定,有一整套既定的、成熟的类目与指标系统;基本上不需要自定义类目,只需要确定分析主题和分析范围即可做到高度程序化、机械化。文献内容范围广、变化大,其研究主题很可能要求作基于语用的分析,而语用表达的可能性

几乎是无穷的。因而想用一套统一的、不变的概念系统来囊括所有可能类目基本是不可能的。因而,内容分析大多要根据研究目标和样本数据来自定义类目。

3.4.2 运算集 目前,文献计量与内容分析均以线性的统计运算为主。在此基础上,文献计量可以运用和验证几个经验公式;而内容分析通常只做一些简单的对比分析和直观推测。总之,运算集(我们对现实关系进行模拟的数学模型)是有限的。因此,就运算集而言,尚有很大的开发潜力:一则可以通过对现实的观察和科学发现,提出更多、更有效的经验公式,形成知识库;二是线性近似未必是对客观事实的最佳模拟,应以有效度为标准,开发一些非线性的运算模型;三是可以从单纯的数值运算发展到数值运算与逻辑运算相结合。

3.4.3 分析粒度 粒度有粗细之分。在文献计量中,文献的一切外部特征,都是与文献整体相关的,这时分析粒度是较粗的。但词频、主题、引文等则是相对较细的分析粒度。内容分析根据研究目的或方法的不同,可以以单篇文献整体判读作为一个意义表示,这时是粗粒度的;更多的是以文献中某个与类目相关的事实、段落、句子等作为一个意义表示,总的来说粒度较细。实际上,粒度的粗细主要取决于类目划分的详尽程度。

3.4.4 分析对象 文献计量由于主要关注文献外部特征,所以其分析对象不能脱离文献实体。而内容分析则可以脱离文献实体,分析未形成文字材料的语言传播内容或行为等,可分析的对象范围较广。

3.4.5 映射过程 从文献实体向形式化体系映射形成符号式的“属性-值”对的过程,必须保证满足隐含前提,即实体与概念之间的联系是确切的、真实的;类目、指标的划分必须近似反映事物运动的关键方面,并不因未计入因素而发生大的偏差。

3.4.6 人为介入 与文献计量相比,内容分析中定性分析的过程占主要方面,需要较多的人为介入。其关键过程,即确定和划分类目以及将分析单元归入既定类目的过程,都是人为的,机器很少能够代劳。机器擅长数据记忆、数值计算、语法分析,但语义和语用理解、联想、猜测等则是其弱项。

4 结 语

本文试图给予文献计量与内容分析以理论上的综合,提出一个一致性的解释,即将文献群实体映射为经过简化、抽象的概念符号,进行形式化演算,从而力图发现附着在文献群背后的客观隐性知识。以此为依据,笔者认为,文献计量与内容分析都可以纳入情报学的分支“信息计量与分析”的范畴内,进行综合应用。至于有关方法与应用究竟应如何综合的问题,尚有待进一步探讨。

参考文献:

- 1 邹菲. 内容分析法的理论与实践研究. [学位论文]. 武汉:武汉大学,2004
- 2 邱均平,黄晓斌,段宇锋等著. 网络数据分析. 北京:北京大学出版社,2004
- 3 邱均平. 文献计量学. 北京:科技文献出版社,1988
- 4 苏新宁,杨建林,邓之鸿等著. 数据挖掘理论与技术. 北京:科技文献出版社,2003
- 5 卢泰宏. 信息分析. 广州:中山大学出版社,1998
- 6 邱均平,邹菲. 国外内容分析法的研究概况及进展. 图书情报知识,2003(6):6-8
- 7 邱均平,邹菲. 我国内容分析法的研究进展. 图书馆杂志,2003(4):5-8
- 8 简晟峰,陈秀涵. 内容分析法. [2004-11-12]. http://blue.lins.fju.edu.tw/~su/rm91/res_ca.htm
- 9 刘胜骥. 内容分析法. [2004-11-12]. <http://iir.nccu.edu.tw/Liuse/%B2z%BD%D7%BBP%ACF%B5%A6/%A4%BA%AEe%A4%C0%AAR%AAk.doc>
- 10 冯郁青. 媒介内容分析的相关理论. 新闻与传播研究,1998(3):66-73
- 11 卜卫. 试论内容分析法. 国际新闻界,1997(4):55-59,68
- 12 马费成,姜婷婷. 信息构建对当代情报学发展的影响. 图书馆论坛,2003(6):20-25
- 13 马文峰. 试析内容分析法在社科情报学中的应用. 情报科学,2000(4):346-349
- 14 陈维军. 文献计量法与内容分析法的比较研究. 情报科学,2001(8):884-886
- 15 罗金增. 内容分析法与图书馆学. 情报杂志,2003(4):51-53

〔作者简介〕朱少强,男,1975年生,博士研究生,发表论文7篇。

邱均平,男,1947年生,武汉大学中国科学评价研究中心主任,武汉大学图书情报研究所所长,教授,博士生导师,发表论文200余篇。

(上接第18页)

趋势》,成功预见了网络和全球经济一体化等现象,从而使这一方法受到世人瞩目。

参考文献:

- 1 申凡,威海龙. 当代传播学. 武汉:华中理工大学出版社,2000
- 2 卜卫. 试论内容分析方法. 国际新闻界,1997(4):55-59,68
- 3 李本乾. 描述传播内容特征 检验传播研究假设:内容分析法简介(上). 当代传播,1999(6):39-41
- 4 李本乾. 描述传播内容特征 检验传播研究假设:内容分析法简介(下). 当代传播,2000(1):47-49,51
- 5 吴世忠. 内容分析方法论纲. 情报资料工作,1991(2):37-39,47
- 6 邱均平. 文献计量学. 北京:科学技术文献出版社,1988
- 7 陈维军. 文献计量法与内容分析法的比较研究. 情报科学,2001,19(8):884-886
- 8 任学宾. 信息传播中内容分析的三种抽样方法. 图书情报知识,1999(3):29-30
- 9 范并思. 论社科情报研究的方法体系突破口. 情报资料工作,1995(2):4-7
- 10 Allen B, 宴平安. 图书情报学研究中的内容分析法. 国外情报科学,1993,11(1):27-30
- 11 罗良道,李晓红. 内容分析在图书馆学和情报学研究中的运用. 情报学刊,1993,14(2):130-132
- 12 Krippendorf K. Content Analysis: An Introduction to Its Methodology. Beverly Hills: SAGE Publications,1980
- 13 West M D. Applications of Computer Content Analysis. Ablex Pub., 2001
- 14 Haggarty L. What is ... content analysis? Medical Teacher, 1996,18(2):99-101
- 15 Bos W, Tarnai C. Content analysis in empirical social research. International Journal of Educational Research,1999:1
- 16 Neuendorf K A. The Content Analysis Guidebook. Thousand Oaks.: Sage Publications, 2002

〔作者简介〕赵蓉英,女,1966年生,副教授,博士研究生,发表论文30余篇。

邹菲,女,1978年生,硕士研究生,发表论文10余篇。

(上接第8页)

- 5 缪其浩. 观察国际图书馆学术前沿及其发展:内容分析. 中国图书馆学报,2002(3):5-8
- 6 范宇中. 智能信息系统中的知识获取研究. [学位论文]. 武汉:武汉大学信息管理学院,2004
- 7 Haggarty L. What is content analysis? Medical Teacher, 1996,18(2):99-101
- 8 Soloman M. An opportunity for the analytical librarian. Searcher, 2002,10(9):62-65

〔作者简介〕邱均平,男,1947年生,武汉大学中国科学评价研究中心主任,武汉大学图书情报研究所所长,教授,博士生导师,发表论文200余篇。

余以胜,男,1975年生,博士,发表论文6篇。