大数据与双边关系的量化研究*：以GDELT与中美关系为例

池志培　侯　娜

【内容提要】 大数据事件库的出现给量化国家间关系提供了一种新的可能路径。本文利用目前全球最大的事件数据库GDELT来测量1993—2016年的中美关系，并探讨五种不同的计算方法及其问题。本文讨论了如何判断测量的准确性问题，并将结果与清华大学的“中国与大国关系数据库”中同时段的中美关系测量值进行比较，证明基于大数据的测量具有一定的价值。本文同时分析了大数据测量方法存在的问题，并探讨了其解决方法。

【关键词】 GDELT　大数据　事件　双边关系测量　中美关系

【作者简介】 池志培，中央财经大学国防经济与管理研究院助理研究员，中央财经大学全球经济与可持续发展研究中心兼职研究员。
电子邮箱：baconchi@126.com
侯娜，中央财经大学国防经济与管理研究院副教授、副院长，中央财经大学全球经济与可持续发展研究中心兼职研究员。
电子邮箱：hounacufe@126.com

国际关系的研究对象是国与国之间的关系，但由于国与国之间的关系往往是多维和复杂的，因而对关系的判断通常只能依赖于研究者的经验、直觉和理论偏好，这也就意味着不同的研究者对于国与国之间关系的判断差异很大。要解决这个问题，一个方法是对于国家间的关系进行量化研究。

* 本文系中央财经大学全球经济与可持续发展研究中心专项研究课题“战略安全与国家动员能力建设”的成果。

量化是当代社会科学研究的趋势,通过准确的数据,我们才能进行跨时段、跨国的比较,从而发现趋势与规律,甚至提供某种预测。但是要将国家间关系转化成单一维度的数字来测量还面临着方法论和实际操作中的巨大困难。

目前衡量国家间关系的主要方法是分析事件数据。从理论上来说,如果能对国家之间发生的所有事件进行统一的分析,那么应该就能很好地把握它们之间的关系,因为关系必然要通过事件来体现。如果所有的事件都能通过一个统一的测量标准进行衡量,再将这些测量的结果进行汇总,那么就能对关系进行量化。由于双边关系是所有关系的基础,多边关系可以还原成多组双边关系,所以本文的研究将聚焦于双边关系的测量。

近年来,随着计算机技术和智能硬件技术的飞速发展和普及以及网络社会的形成和算法的进步,人类得以累积了海量数据,即所谓的大数据。虽然不同的学者或者机构对于大数据的定义有所区别,但是对于大数据的一些共性特征,各方还是有一些共识——大数据包含了 3 个 V[①]:量(volume),数据量非常巨大;种类(variety),即数据类型多样,从文本到图像、视频,等等;速度(velocity),即数据产生和处理的速度非常快。本文讨论所涉及的目前全球最大的社会科学数据库 GDELT(Global Database of Events, Language, and Tone)就是其中一个,它从全世界超过 100 种语言的媒体中收集信息,并通过特定的编码体系由计算机自动将其编码成一个个事件,时间跨度从 1979 年到今天,并持续每天更新。目前已经收集了超过 2.5 亿个事件的信息,包括事件的发起者、对象、地理位置、事件类型、信息来源等 32 个变量。就传统国际关系关注的国家之间的关系而言,这个数据库几乎涵盖了所有已经公开的事件。正如维克多·麦尔-荀伯格(Viktor Mayer-Schönberger)所提及的,抽样数据会变得过时,因为我们可以获得全部的数据[②]。如果能

① 关于大数据的不同定义,可以参见:Jonathan Stuart Ward and Adam Barker, "Undefined By Data: A Survey of Big Data Definitions," arxiv.org/abs/1309.5821. 关于大数据的简要历史,可参见:Gil Press,"A Very Short History Of Big Data," *Forbes*, May 9, 2013, https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#5db944eb65a1. 此外,亦有观点认为大数据还需要第四个 V,即 veracity,指数据的真实性。

② Kenneth Cukier and Viktor Mayer-Schoenberger, "The Rise Of Big Data: How It's Changing the Way We Think about The World," *Foreign Affairs*, Vol. 92, No. 3, 2013, pp. 28-40.

充分利用这些巨量的事件大数据，那么就应该可以相当准确地对双边关系的现状和趋势作出判断。那么，使用这些海量数据能否得到可靠的结论，与现有的方法相比如何呢？同时，使用这些大数据可能遇到的问题和可能的解决方式又有什么呢？本文将以中美关系为例，通过对 GDELT 事件数据的使用来探讨这些问题。

一、事件数据与双边关系：历史与文献回顾

虽然大数据事件库是近些年才出现的，但是人类通过记录事件数据来发现和研究社会已经有很长的历史了，一个例子就是对犯罪事件的记录。当代学术意义上的通过事件数据来考察两国之间的关系开始于 1960 年代，随着社会科学中的行为主义革命而产生。行为主义革命者希望采用能观察到的变量和计量方法来研究社会现象。对于国际关系、外交政策的分析而言，这个能被测量的对象就是事件（event）。[①] 查尔斯·麦克莱兰（Charles McClelland）最早从他对外交史的研究开始了这种尝试。最初的事件数据的生成采用的是人工手动编码的方式。由于人工编码需要大量的人力，研究者不得不在分析的广度和成本之间作取舍[②]。覆盖的来源越多、广度越大，

① 关于事件数据库的起源与行为主义革命的关系以及早期发展，可以参见：Stephen J. Andriole and Gerald W. Hopple, "The Rise and Fall of Event Data: From Basic Research to Applied Use in the US Department of Defense," *International Interactions*, Vol. 10, No. 3-4, 1984, pp. 293-309; John Lewis Gaddis, "Expanding the Data Base: Historians, Political Scientists, and the Enrichment of Security Studies," *International Security*, Vol. 12, No. 1, 1987, pp. 5-7; Philip A. Schrodt, "The Statistical Characteristics of Event Data," *International Interactions*, Vol. 20, No. 1-2, 1994: 35-53. 关于政治学中的行为主义革命的特征及其简要历史，可以参见：Robert Dahl, "The Behavioral Approach in Political Science: Epitaph for a Monument to A Successful Protest," *American Political Science Review*, Vol. 55, No. 4, 1961, pp. 763-772; David Easton, "Introduction: The Current Meaning of 'Behavioralism' in Political Science," in J. S. Charlesworth, ed., *The Limits of Behavioralism in Political Science* (Philadelphia: American Academy of Political and Social Science, 1962), pp. 1-25; David Easton, "Political Science in the United States: Past and Present," *International Political Science Review*, Vol. 6, No. 1, 1985, pp. 133-152.

② Richard L. Merritt, "Measuring Events for International Political Analysis," *International Interactions*, Vol. 20, No. 1-2, 1994, p. 6.

就意味着研究成本的急剧上升,因此研究者不得不选择很有限的几个事件的数据来源。同时,人工编码也容易受个人身体状态(比如疲劳等)和主观判断的影响①,因此,事件分析方法的局限比较明显。早年的代表性数据库有鲁道夫·鲁美尔(Rudolph J. Rummel)的"国家的维度"(Dimensionality of Nations,DON; Rummel, 1972),查尔斯·赫尔曼(Charles Hermann)等的"国家事件比较研究"(Comparative Research on the Events of Nations, CREON; Hermann et al., 1977),爱德华·阿萨尔(Edward Azar)的"冲突与和平数据集"(Conflict and Peace Data Bank,COPDAB; Azar, 1980, 1982; Azar and Sloan, 1975),查尔斯·麦克莱兰(Charles McClelland)的"世界事件互动测量"(World Event/Interaction Survey,WEIS; Charles McClelland,1976)等。这些数据库被广泛应用于国际关系研究,尤其是关于冲突、动乱和战争的研究。② 其中 COPDAB 和 WEIS 的数据在经济相互依赖与冲突的研究中时常被使用。③ 在 1970 年代末和 1980 年代初,美国的政府机构如国务院和国防部也组织了类似的项目,主要用于冲突预警。

在计算机广泛应用于社会科学研究之后,有学者将计算机程序自动编码引入事件分析之中,以解决广度和成本的矛盾。从 1980 年代末和 1990 年代初开始,美国国家科学基金会(National Science Foundation, NSF)支持了"国际关系中的数据发展"项目(Data Development in International Relations, DDIR),利用计算机来自动编码新闻事件。在这项资金的支持下,堪萨斯大

---

① 在关于事件数据库的研究开始以后,研究者对于人工编码存在的各种可能的问题也进行了研究,包括利用同一数据来源来比较不同编码者之间的差异,不同的编码者和数据库的表现差异很大,有从 40%到 90%不等的可靠性。参见:Philip A. Schrodt and Christopher Donald, "Machine Coding of Events Data," paper presented at the International Studies Association meetings, Washington DC, April 1990, p. 6.

② Philip A. Schrodt, "The Statistical Characteristics of Event Data," *International Interactions*, Vol. 20, No. 1-2, 1994, pp. 35-6.

③ Solomon W. Polachek, "Conflict and Trade," *Journal of Conflict Resolution*, Vol. 24, No. 1, 1980, pp. 55-78; Mark Gasiorowski and Solomon W. Polachek, "Conflict and Interdependence: East-West Trade and Linkages in the Era of Détente," *Journal of Conflict Resolution*, Vol. 26, No. 4, 1982, pp. 709-729; Jon C. Pevehouse, "Interdependence Theory and the Measurement of International Conflict," *The Journal of Politics*, Vol. 66, No. 1, 2004, pp. 247-266.

学的美国政治学家德波拉·耶纳(Deborah J. Gerner)和菲利普·斯洛德特(Philip Shrodt)等人利用 WEIS 的事件编码系统，实现了自动事件编码处理，形成堪萨斯事件数据系统(Kansas Event Data System，KEDS)，成为当时最大的事件数据库。耶纳和斯洛德特等后来开发了新的编码系统——冲突和调停事件观察(Conflict and Mediation Event Observations，CAMEO)，并为后来的大型事件数据库所广泛使用。① 斯洛德特还开发了使用这个编码系统的程序——以强化替换说明进行的文本分析(Textual Analysis by Augmented Replacement Instructions，TABARI)。

2008 年，美国国防部下的国防高级研究计划局(Defense Advanced Research Projects Agency，DARPA)资助了一个事件数据库项目“综合冲突早期预警系统”(the Integrated Conflict Early Warning System，ICEWS)，通过收集事件数据来进行风险预警，主要是针对亚太地区。该项目资助开发了新的自动编码程序 BBN ACCENT，达到了相当高的准确率，与人工编码比照，大概在 80%左右。② 这个数据库在 3 年的试验期后转交给了美国海军。其数据涵盖了 1995 年到当下，并且大量数据已经放置于哈佛的开源数据库(Harvard Dataverse)中，不过开源部分的数据只能到当前年份的前一年。美国国家科学基金会在 2013 年也再次资助了事件库的开发和研究，目前有一个历史数据库(Cline Center Historical Phoenix Event Data)和一个实时数据库(Phoenix Near-Real-Time Data)，可以回溯到 1945 年，但是新闻来源相对有限。③ 目前最新的仍在进行的工作是重新开发新的编码系统——可验证事件数据的政治语言本体(Political Language Ontology for Verifiable Event Records，PLOVER)和编码软件——文本解析与相关编码

---

① 另一个不太常用的自动编码系统是 the Integrated Data for Event Analysis (IDEA)，为另一个自动事件数据项目 the Protocol for the Assessment of Nonviolent Direct Action (PANDA)所使用。

② Elizabeth Boschee, et al., “ICEWS Coded Event Data,” Harvard Dataverse, V22, 2015, https://doi.org/10.7910/DVN/28075; BBN ACCENT Event Coding Evaluation. updated v01.pdf

③ the New York Times (1945—2005), the BBC Monitoring's Summary of World Broadcasts (1979—2015) and the CIA's Foreign Broadcast Information Service (1995—2004).

体系的派森引擎(Python Engine for Text Resolution and Related Coding Hierarchy,PETRARCH-2),而这将是事件数据库发展的一个新阶段,将取代目前主要数据库所采用的 CAMEO 编码系统。同时这些最新的开发工作都是开源的,源代码都放在了开源社区 Github 中。

与前面述及的这些数据库相比,GDELT 要更为庞大。它是由卡勒夫·李塔鲁(Kalev Hannes Leetaru)和斯洛德特在 2012 年启动的一个项目。它也是使用计算机自动编码的方式,从全球超过 100 种语言的媒体中自动挖掘信息,将新闻信息编码成一个个事件的输入。其中,最重要的 3 个信息来源是法新社、美联社和新华社。时间范围是从 1979 年到现在,并将扩展到 1800 年。同时,计算机每天都会从世界各地的媒体中持续收集信息,日增约 10 万个事件。编码体系采用的是 CAMEO 系统。在 CAMEO 系统中总共有 20 个大类超过 300 种不同的事件类型。而 GDELT 又将 300 多类的事件最终分成 4 个大类,即言语合作(verbal cooperation)、现实合作(material cooperation)、言语对抗(verbal conflict)和现实对抗(material conflict)。GDELT 项目的每一个输入包含了许多不同的信息项目,从 1979 年到 2013 年 3 月 31 日止的输入包含 57 个信息,而 2013 年 4 月 1 日开始则增加到 58 个信息,增加了信息的来源以便核对。这些信息包括每个事件的时间,行为主体和行为对象的国家、名字、组织、类别,事件本身的性质、影响程度、在所有来源中被提及的次数,描述的语气,事件的地理信息(包括经纬度等)以及录入日期、信息来源,等等。目前的数据已达数以亿计,并可以免费获得。

数据规模的庞大以及容易获得使得 GDELT 一经推出就引起了广泛关注和讨论,虽然事件数据库主要用于冲突研究,但也已经有一些研究尝试利用 GDELT 来测量双边关系。如帕斯卡·阿卜(Pascal Abb)和盖尔·斯特伍(Georg Strüver)的文章利用 GDELT 来衡量中国与东盟各国的关系,即利用 GDELT 事件数据中每个事件的 Goldstein 分值来衡量双边关系。Goldstein 分值衡量了冲突或者合作事件的强度。他们将中国与东盟各个国家每年所有相关事件的 Goldstein 分值的平均值作为测量方式。①

① Pascal Abb and Georg Strüver, "Regional Linkages and Global Policy Alignment: The Case of China-Southeast Asia Relations", SSRN, March 15, 2015, https://ssrn.com/abstract=2600419 or http://sci-hub.tw/10.2139/ssrn.2600419.

国内利用事件数据库来测量双边关系开始得相对较晚。最早的研究来自李少军，他 2002 年的一篇文章从《人民日报》中选取了克林顿政府期间中美关系的 642 个事件，并通过给每个事件赋值然后再求和计算的方式来判断中美关系的冲突与合作水平。① 而最有影响力的研究是来自清华大学阎学通团队的系统工作。他们建立了一套自己的编码体系，然后采用了人工编码的方式②，主要基于《人民日报》和外交部的数据，系统地梳理了从 1953 年开始的最为重要的中国双边关系的月度变化数据，包括中国与美、日、俄、英、法、澳、越、印尼、巴基斯坦的关系。通过对每个月发生的双边关系事件的赋值打分和加总计算，展示了中国这些重要双边关系的演变，为实证研究双边关系的变化及其影响提供了非常重要的数据支持，产生了一系列重要的学术成果。但是，由于基于人工编码的方式，所涉及的工作量极大，也因而难以扩展到更多的信息来源，而这也是许多评论者认为还需要改进的方面。③

事件大数据的出现提供了一个新的机会来重新思考双边关系的度量，在这种全覆盖的方式下我们是否可以获得更为准确的数值呢？接下来本文将以中美关系的测量为例，探讨如何利用事件大数据来度量双边关系，并讨论其准确性以及潜在的问题和应对。

二、GDELT 与中美关系(1993—2016)的测量

(一) 计算方法

虽然 GDELT 包含了海量数据可供进一步分析，但是我们仍需要将这些

---

① 李少军：《“冲突—合作模型”与中美关系的量化分析》，《世界经济与政治》2002 年第 4 期，第 43—49 页。

② 见“事件分值基准表”，阎学通、周方银：《国家双边关系的定量衡量》，《中国社会科学》2004 年第 6 期，第 101—103 页。

③ 陈定定：《定量衡量的得与失——简评〈中外关系鉴览 1950—2005：中国与大国关系定量衡量〉》，《国际政治研究》2010 年第 4 期，第 163 页；董青岭：《从事件赋值走向关系赋值：双边关系的定量衡量——评〈中外关系鉴览 1950— 2005：中国与大国关系定量衡量〉》，《外交评论》2011 年第 2 期，第 155 页。

事件通过一定的算法进行处理来变成一个关系的量值。基于 GDELT 中能够获取的信息,有五种方式可以构建出关系量值。

第一,由于 GDELT 中每一个事件最终都可以归结成冲突事件或合作事件,所以计算每年合作数量与冲突数量的比值就可以衡量出双边关系的好坏。如果合作事件与冲突事件的比值很高,那么说明合作事件的数量要远多于冲突事件,那么就可以认为双边关系处于较为良好的状态。通过观察年度数据的变化,自然也就可以衡量出双边关系的变化。这种方式的问题在于没有将事件的强度考虑进去,有些事件可能比另外一些事件对双边关系的影响更大。

第二,也可以将合作事件数量减去冲突事件数量的差值作为双边关系的一个衡量。但是这个方法在当下使用会遇到严重的问题,即事件数量的逐年增加。由于过去几十年中技术的飞速进步、网络对社会生活渗透程度的巨大变化以及跨国交往的急剧增加,数据库收录的事件数量越往后越多,直接的差值也会随着年份往后而急剧变大,因而很难用来测量双边关系的变化。当然,如果将来技术和社会变革到了平台期,这种方式可能也可以较好地反映双边关系的变化。此外,这种方法也没有考虑到事件之间强度的差异。

第三种方法略有不同。由于 GDELT 也给出了每个事件的 Goldstein 分值,这个值衡量了事件的强度,因此也可以用这个值来构建双边关系的值。可以将所有事件的 Goldstein 分值相加,最后的结果即双边关系的赋值。这种测量方式的问题是事件的影响并不能相互抵消,即即便一个合作事件与一个冲突事件的影响的 Goldstein 值相同,在计算中会相互抵消,但是现实中对关系的影响也可能依然存在,尤其是在双边的复杂关系中。同时,一个 Goldstein 值为-10 的事件对关系的影响也很难同 10 个 Goldstein 值为-1 的事件的影响等同。此外,这种计算方式也无法反映出事件数量。而在双边关系的大部分事件为低烈度事件的时候,事件数量可能比强度能更好地把握双边关系。同时,这个方法也会受到事件数据的自然膨胀的影响,遇到与第二种方法类似的问题,所以如果要使用大数据事件库来测量计算的话,这个方法会有内在缺陷,导致它也不是很适合使用。李少军利用事件数据来衡量中美关系的论文实际上就是采用了这个计算方法,

当然，他的数据来源是单一的，不存在逐年自然扩张的问题，因此准确度要更高一些。

第四种方法基于 Goldstein 分值的计算，是在将所有事件的 Goldstein 分值加总以后，再除以事件数，得出所有事件的平均 Goldstein 值作为双边关系的一个度量。阿卜和斯特伍的文章对于中国与东南亚国家关系的度量即采用了这种方法。这种方法解决了事件数量随年份增加的问题，但是也没有解决事件之间相互抵消的问题。当然，这也是所有通过事件来衡量关系的方法都会遇到的问题，即一个正值事件与一个负值事件或者一个高值事件与多个低值事件之间的权衡问题。

第五种方法也是基于 Goldstein 分值，即分别计算冲突事件和合作事件的 Goldstein 分值的平均值，然后以两者比值的绝对值作为双边关系的衡量。[①] 取平均值是为了避免事件数量在不同年份的巨大差异。这种方式测量的双边关系值是合作事件和冲突事件的平均强度的比值。这里也会遇到与第三种方法相同的问题，即无法反映出事件的数量。同时也会遇到事件之间的折算问题，比如，一个 Goldstein 值为 5 的事件是否对双边关系的影响等于 5 个 Goldstein 值为 1 的事件。同时，在对计算值的解释上也会有一定的困难。双边关系中合作事件和冲突事件的平均强度的比值作为双边关系的衡量不够直观，在具体进行研究的时候会遇到对结果的阐释问题。

除计算方法外，这里还会涉及时间段的选择问题。用事件数据库的方法来进行测量，我们必须选择特定的时间段即一个时期内发生的事件，这个时期可以是一周、一个月、一年或者任意长度的时间段。阎学通和周方银对李少军以及一般的用事件数据来衡量关系的一个批评，是事件分析法没有将历史考虑进去，基于事件的分析总是从零开始计算关系。某个月没有发生冲突或者合作的事件，双边关系也并不会是零。如果某个月都是负值事件，也不表示两国关系就是以冲突为主。[②] 这一批评有一定的合理性，但是

① 由于冲突数据的 Goldstein 分值为负数，所以最后比值也为负数，故取绝对值。

② 阎学通、周方银：《国家双边关系的定量衡量》，《中国社会科学》2004 年第 6 期，第 92 页。

并不完全准确。这个问题的实质是时间段选择的问题。一个月内发生的事件可能具有一定的偶然性,一个月内可能确实不会发生大的事情或者都是同一方向的,但是如果选取的时间段足够长,那么对该时间段内所有双边事件的考察应该足以反映出双边关系。在较长的一个时间段内,两国关系的种种不同方面应该都会通过事件反映出来,毕竟历史本身也就只是一个时间段。笔者认为,在用事件衡量双边关系时,一年是一个比较合适的时间区间。一年内发生的事件应该足以体现双边关系的历史性基础,因为人类社会的计划周期也一般以一年为基础,比如年龄计算、预算周期、工作报告,等等。而在政治世界中,年度周期也是比较明显的,比如年度峰会、年度会晤、年度报告,等等。考虑到当今世界双边关系的互动程度,在一年之内没有发生合作或者冲突事件或者只发生一种类型事件的可能性是可以忽略的。还有一个考量是基于与现实的其他数据库的匹配,政治类的数据库多以年度更新为主,故以年度事件数据为依据计算也便于结合其他的数据进行分析。因此,在以事件数据来计算双边关系时,采用年度变化值是比较合适的,基本能够避免测量非历史性的批评。

(二)测量的准确性

讨论完测量方法之后的另一个问题,就是如何知道我们的测量是否可靠地、比较真实地反映了双边关系的事实,而这也是测量最为重要的一个方面。

最为直接和常见的方法当然是与我们的常识相比较,如果度量的结果过于反常识,可能就意味着测量或者计算的方法有一定的问题。比如,如果两国之间发生战争,双边关系的值反而显示出双边关系更为友好,那么对双边关系的测量显然是有严重的偏误。比常识更为严格的检验是专业人士的共识,毕竟专家们能够掌握更多更全面的信息,对双边关系这种复杂事实的把握要比常识更为可靠,所以我们的测量值也可以同专家们比较一致的意见相比较来考察其准确度。但是这两种方法都是定性的考察,而我们的测量则是定量的。要想较为准确地比较定性和定量的分析存在方法论上的困难,毕竟定性语言的含混性与定量语言的准确性之间有难以弥合的鸿沟。

定性语言的区分度要更低一些，从数值 3 到 3.5 或者 4 的变化在定性研究者看来也许都属于友好关系，难以再进行进一步的区分。

如果我们使用定量的方法来检验，那么我们可以使用的方法包括与现成的数据库相比较，又或者通过替换数据源的方式。具体到用 GDELT 的数据来衡量中美关系，我们可以与清华大学中国与大国关系数据库中的中美关系值相比较。由于后者已经给出了中美关系的月度变化值，所以将计算结果与之比较就能考察利用事件大数据进行双边关系测量的准确性，并且由于清华大学的数据来自于人工筛选和计算，这种比较就显得更有价值。需要说明的是，由于基本方法论和计算方式上的差别，直接的数值比较并没有意义，能够进行比较的应该是对趋势的把握，即双边关系的变化趋势在两种测量方式下的比较。

当然，定性和定量的方法在思考测量的准确性时可以同时使用，并非相互排斥。

（三）中美关系（1993—2016）

基于前面的论述，这一部分将具体讨论用 GDELT 来测量双边关系。由于中美关系的重要性，现有的通过事件数据来测量双边关系的研究都特别关注中美关系，也选取中美关系进行测量，而这也有利于将本文的研究置于对中美关系的普遍关注和讨论中。另一方面，GDELT 中，时间越往后，数据的数量越多，所以本文选择的时间范围是冷战结束以后，即克林顿执政（1993 年）以后。由于用于比较的数据可获得性的原因，本文数据截止于 2016 年底。因此，本文研究的时间范围是 1993 年至 2016 年。在这段时期内，中美关系整体上处于一个较为稳定的状态。

笔者首先从 GDELT 提供的数据中选取了从 1993 年到 2016 年中美政府间的事件，将其按照 4 个大类计算；同时由于区分了施动者，所以又可以分为中国对美国的事件和美国对中国的事件，简单的总和计数如表 1 所示。可以发现，合作事件要远远多于冲突事件，而言语合作事件又要多于言语冲突事件。并且中美各自发起的冲突和合作的事件数量也大体相同，体现了两个大国之间较为对等的关系。合作事件要远远多于冲突事件，这应该也

是符合常识和多数专家意见的,即中美两国虽然时有冲突,但是还是合作更多。①

表 1 中美两国合作与冲突事件的数量(1993—2016)

言语合作(中→美)	现实合作(中→美)	言语合作(美→中)	现实合作(美→中)	言语冲突(中→美)	现实冲突(中→美)	言语冲突(美→中)	现实冲突(美→中)
14866	786	15985	504	1010	448	908	364

前文已经讨论了 5 种将事件转化为关系的算法,在现有的 GDELT 中,不同年度收录的事件数据有着比较大的差异,并且越往后的年度数据越多,因此基于差值计算的方法将会面临年度数据越往后越大的问题,很难反映出是关系的变化还是数据收集技术的变化带来的差别。本文采用合作与冲突事件次数的比值和年度平均 Goldstein 分值的方法进行计算。

首先是用合作和冲突事件次数比值的方法。这种计算方法暗含的假定是所有事件的权重相同。阎学通和周方银的文章已经证明,即便只计算事

① 清华大学的中国与大国关系数据库显示,中美关系在这个区间内也是正值要远多于负值(按该数据库的定义说明,关系在普通区间以上,即合作多于冲突)。学术界的普遍看法是,冷战后从克林顿政府开始,美国对华采取了接触政策(engagement)为主的战略,强调以合作塑造中国的行为,对这一政策的反思近年来在美国学界和政界颇多,可参见:Kurt Campbell and Ely Ratner,"The China Reckoning: How Beijing Defied American Expectations," *Foreign Affairs*, Vol. 97, March/April Issue, 2018, pp. 60-70. 这篇 *Foreign Affairs* 上的文章引发了一些最为知名的中美关系学者对美国对华政策以及中美关系的讨论,参见:Wang Jisi et al., "Did America Get China Wrong?: The Engagement Debate," *Foreign Affairs*, July/August Issue 2018, https://www.foreignaffairs.com/articles/china/2018-06-14/did-america-get-china-wrong. 对美国政府对华政策的内部考量,还可参见:Michael Green, *By More Than Providence: Grand Strategy and American Power in the Asia Pacific since* 1783 (New York: Columbia University Press, 2017). 另一方面,虽然承认中美之间在利益上有差异和冲突,但国内学者也认为合作是冷战后的中美关系的基本基调,参见:王缉思:《浅论中美关系的大环境和发展趋势》,《美国研究》2006 年第 1 期,第 89—96 页;金灿荣、段皓文:《当前中美关系的问题与出路》,《国际观察》2014 年第 1 期,第 71—83 页。当然,主流的说法是"中美关系好不到哪里去,也坏不到哪里去",强调对分歧的管控。

件的个数(假定权重赋值都相同),得出的结果依然是相当准确的。[①] 笔者随后计算了中美从 1993 年到 2016 年每年的合作和冲突事件的比值(图 1)。可以看到,合作事件远多于冲突事件,两者的比值较大。但是随着时间的推移,虽然具体年份值的高低并不是一个线性的变化过程,但是这个值有减少的趋势,也就是说冲突事件的比例在上升,这意味着双边关系有逐渐紧张的趋势。需要说明的是,虽然有这种趋势变化,但中美之间仍是合作远多于冲突。这符合我们一般的对中美关系的认知,即中美关系总体是和平合作的,而随着中国国力的逐渐增强,中美之间竞争的一面会越来越强。阎学通将中美关系总结为"假朋友关系",两者之间关系不稳定,时有冲突,长期来看对抗的趋势不可避免。另外一个有意思的现象是,在 GDELT 测量值中,一般在总统选举年中美关系的值都处于低值(除了 1996 年),这大概与选举年美国的政客需要表现强硬有关系,双方的言语冲突会更多一些。

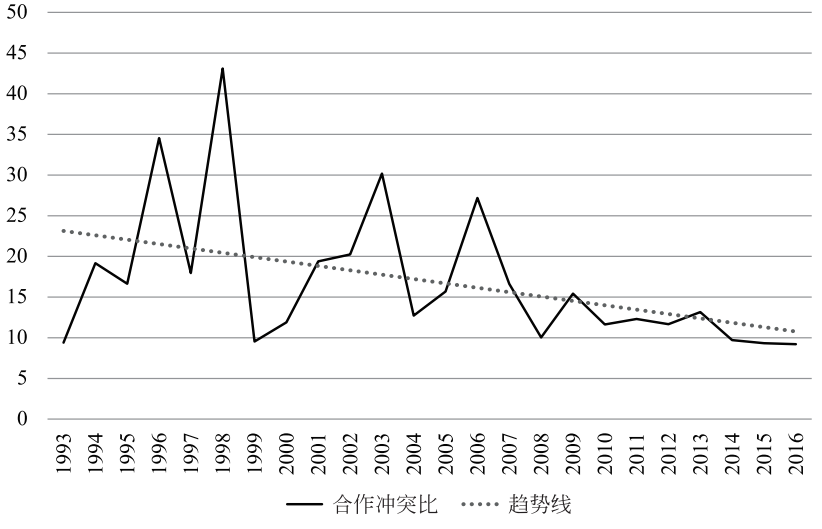

图 1　中美关系的量化测度(合作冲突比)(1993—2016)

如前所说,GDELT 的所有事件都区分了施动者和受动者,所以我们也可以将中美对对方的政策给出类似的测量,用来衡量政策的合作和冲突的

① 阎学通、周方银:《国家双边关系的定量衡量》,《中国社会科学》2004 年第 6 期,第 96 页。

程度(图 2)。与双边关系的总体情况类似,两国的政策也是合作为主,但是竞争的一面随着时间的推移在缓慢地增加。并且相对而言,中国的政策稳定性可能要更强一些(趋势线的斜率要略小于美国),但是近年来两国的政策越来越趋近。① 这种测量是相对于清华大学中国与大国关系数据库的一个优势,在清华数据库中并不能区分出行为方向,而是从一个第三者的角度去观察。在某些研究中,这种对双方政策的描述会有重要的用途,比如考察中美两国对对方的政策如何影响其他国家的相关政策。

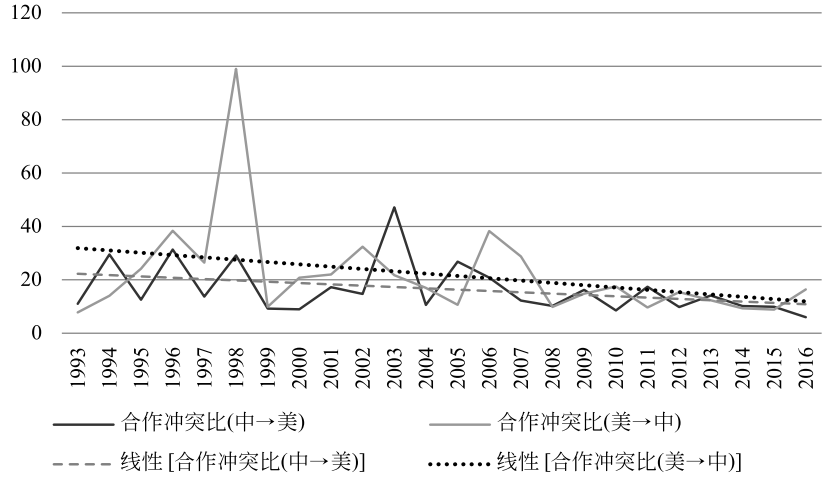

图 2 中美对对方政策的测量(1993—2016)

从常识和专家意见来看,通过 GDELT 大数据的直接测量可以抓住双边关系的一些基本事实。为了进一步考察其准确性,笔者将其与清华大学中国与大国关系数据库中的中美关系值作了比较。由于计算方法不同,所以两者的分值不能直接比较,这里比较的是趋势变化。由于 GDELT 中计算出来的值较大,为了便于观察,从 GDELT 中计算的值都除以 10。同时,对清华大学的月度数据分别采用取年平均值和年末月度值的方式来转化成年度值。由于年末的数据反映了当年前 11 个月的双边关系运行的结果,所以第 12 个月也可以作为当年的关系值。图 3 和图 4 即结果。

① 但是这个结果受到 1998 年的异常值影响较大。如果在数据中删除 1998 年的值,则中美两国的趋势线则几乎是平行的。

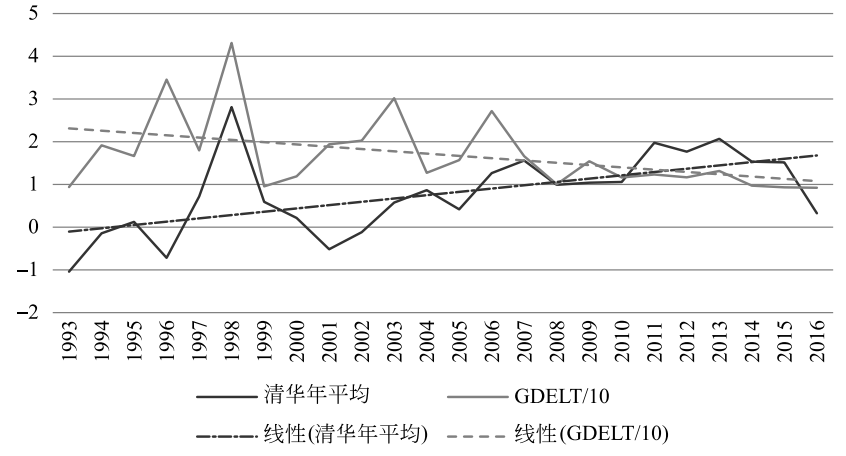

图 3　中美关系 1993—2016(清华值为年平均值)

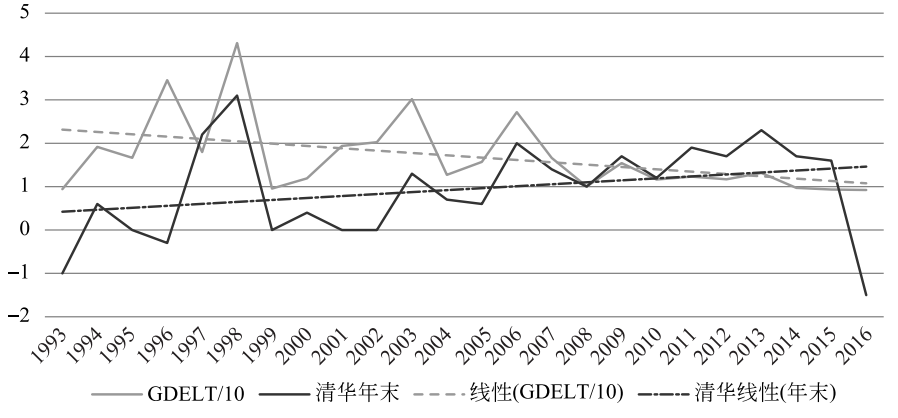

图 4　中美关系 1993—2016(清华值为年末月度值)

简单来看，两种方式计算的值的变化大部分比较相似，比如，不管哪种测量，都认为冷战后中美关系的峰值在 1998 年，但是谷值却有所不同。其中的差异在于，如果某年发生了极为负面的事件，在清华大学的数据里面影响要更大一些，而由于只是事件计数，如果当年也发生了很多合作性的事件，那么从 GDELT 中计算出来的数值就不会是极低值。比如在 1996 年，虽然台海危机导致两国关系非常紧张，但是在当年 3 月危机之后，两国关系即过了最低点，紧接着 1997 年和 1998 年双方元首互访，并将关系定位于“战略伙伴关系”，因此 1996 年在 GDELT 的测量值中不会出现极低值。类似的是

2001 年,虽然发生了南海撞机事件,但是“9·11”事件之后两国关系的迅速改善也使得当年的值不会是极低值。这种差别也贯穿在其他年份中,比如 1997 年和 2004 年。也需要看到,清华大学数据本身的年平均值与年末值也有差异。GDELT 计算的数据与清华大学的数据在 2005 年之后的变化曲线要更为相似一些,这可能与 GDELT 的数据来源有关,事件越早,GDELT 中的数据越少,所以偏误的可能性也更大,这在后文讨论大数据的缺陷和问题时再进一步讨论。

当然,最根本的区别在于趋势线,不管是用年平均值还是用年末值来计算,清华大学中国与大国关系数据库的年度数据都有上升的趋势,即双边关系有逐步改善的倾向,只是在 2016 年奥巴马执政的最后一年有一个急剧的下滑。具体来看,清华大学数据显示,奥巴马时期的中美关系总体要好于小布什时期,除了最后两年的急剧下滑。但是 GDELT 数据显示,奥巴马时期总体不如小布什时期,最后两年的下滑也没有那么戏剧化。这种差别很可能源于 GDELT 对于南海中美冲突事件的抓取要多于清华大学,毕竟清华大学的主要源数据来自于外交部和《人民日报》,相关内容的报道比 GDELT 数据源要少。总体来看,通过计算 GDELT 中合作和冲突事件的比值来度量中美双边关系具有一定的学术价值,值得进一步考察。

另一种测量方法是计算 Goldstein 年平均值(图 5、图 6)。这个值波动幅度很小,在整个时间段内都在 2～3 之间浮动,整体趋势略微向下。如果我们

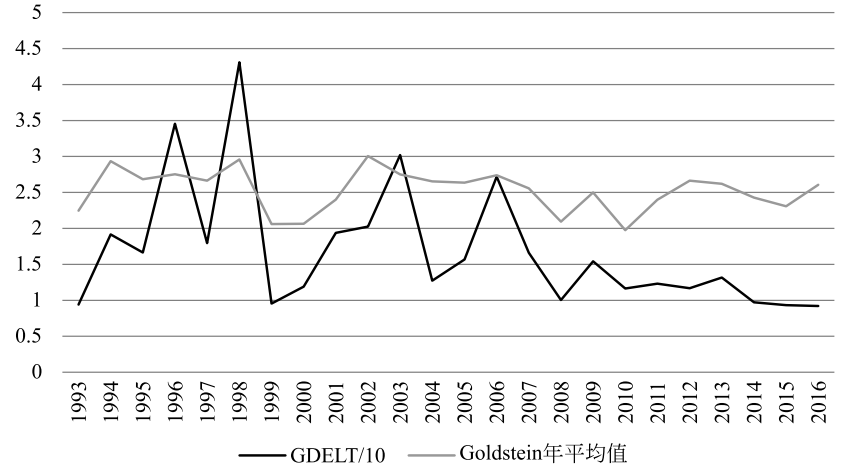

图 5 中美关系合作冲突比与 Goldstein 年平均值计算结果的比较(1993—2016)

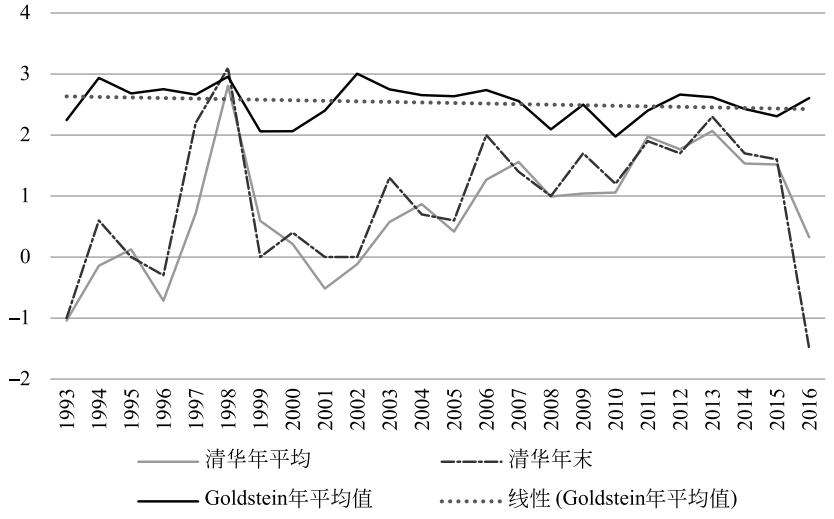

图 6　中美关系 Goldstein 年平均值及与清华大学数据的比较(1993—2016)

关注其折线变化的话，可以发现其波动变化趋势与合作冲突比计算结果的变化几乎是一致的，只是幅度要小得多，而与清华数据也有一定程度的相似性。最为明显的差别是在 2015 年、2016 年，Goldstein 分值的平均值计算显示双边关系有好转，与其他测量都是相反的结果。此外，与其他测量值相比，Goldstein 平均值这种计算方法似乎没有很好地反映出双边关系的波动性，其年度之间的变化强度比其他度量要更小。可能的原因是在当前中美关系的情势下，事件数量比权重的考虑更为重要。在低烈度竞争状态下，将所有事件的权重近似为相同得出的结果可能更能把握双边关系的实质，与前文所引的阎学通和周方银的结论类似。而这可能也是用 GDELT 事件数量比值的测量更能把握双边关系趋势的一个原因。

三、大数据测量双边关系的问题与解决方法

虽然基于如 GDELT 这种大数据事件库的测量为双边关系度量带来一种新的覆盖面更广也更为简便的实现方法，但是它也存在着很多问题需要解决。需要说明的是，这些问题很多也存在于人工编码的事件数据库中，并且人工编码还涉及不同的编码者之间不一致的问题，所以并非人工编码就

一定优于机器编码。接下来这一部分将侧重探讨机器编码事件大数据库可能遇到的问题及其可能的解决途径。

(一)数据来源的质量问题

对于所有的大数据研究来说,数据来源的可靠性问题都是最为重要的。如果数据的源头被污染,那么所有结论都会被质疑。具体对于如 GDELT 这种事件数据库而言,其数据质量问题也一直是有争议的,很多学者也主张数据在使用之前需要进行"数据清洗"①。数据质量问题的原因有很多种,比较严重的包括报道视角、重复和新闻质量。

在所有的新闻报道中,报道者都有其视角,因此必然会包含其对问题有意识或者无意识的特定看法,这无疑会导致由此生成的事件数据也带有同样的偏见,进而难以用来客观地测量双边关系。当然,这不仅仅限于大数据事件库,在人工数据库中,由于涉及大量的劳动力和高强度的重复工作,往往只能有选择性地偏重于特定数据来源。比明显的偏见更为严重的问题则涉及更为根本的哲学争论,即对事件的报道并不是一个完全客观的问题,报道者及阐释者都参与建构了事实。事件数据库往往只选择了特定解释、"唯一的"解释、假想的"上帝视角"客观无偏差地看待一件事的发生。② 现实情况是同一个事件存在着多个不同版本的叙事,而对于计算机编码而言,被编码成不同事件的可能性很大。这就涉及收录新闻事件重复的问题。如果是对同一事实不同角度的表述,那么是否属于重复就需要进一步探讨。

除了新闻事件中的多重视角和建构问题,重复错误也是大数据事件库中普遍存在的一个问题。在今天媒体活跃的时代,同一个事件在一定时间内会被反复地、类似地描述和报道,并被大量的媒体转载报道。在这种情况下,计算机可能会自动抓取这些重复数据,使得数据库中收录的事件数据要(大大)多于实际所发生的数据。沃德(Ward)等人则将 GDELT 与另一个机

---

① 董青岭:《反思国际关系研究中的大数据应用》,《探索与争鸣》2016 年第 7 期,第 91—94 页。

② Gavan Duffy, "Events and Versions: Reconstructing Event Data Analysis," *International Interactions*, Vol. 20, No. 1-2, 1994, pp. 147-167.

器自动编码的事件数据库 ICEWS 相比较，他们发现 GDELT 编码的事件与 ICEWS 编码有相当的出入，前者要比后者数量庞大得多。他们认为 GEDLT 编码的事件比实际要多[①]，而 ICEWS 则比实际少。在某些受事件数量影响的特定算法下，这可能会严重干扰双边关系测量的准确性。

事件数据库还存在着源新闻质量的问题，即新闻错误甚至是假新闻，尤其是在媒体和自媒体更为发达的今天，在这个极为强调媒体时效性的时代，新闻出现错误甚至是故意制造假新闻的情况都不罕见。2016 年的美国总统大选就饱受假新闻困扰，假新闻甚至被认为对最后结果产生了巨大影响。如果没有甄别出来这种新闻错误或者假新闻，事件数据库的客观性必然会受到影响，进而影响到对双边关系的客观测量。

针对这些问题，也有特定的途径来解决或者减少它们对于最终数据和结果的影响。首先，扩大数据源和增加更多数据。更多数据和更广覆盖范围可以使得偏见问题得到部分解决，如果我们认为人的偏见是内在的、难以避免的，那么包含了所有偏见的数据比只包含特定偏见的数据就要更为准确和客观。同时，如果我们能够增加更多可靠的数据源，那么假新闻和新闻错误的影响也会被相应地缩小。当数据量极大的时候，即便存在一部分错误和重复，其对最终结果也不会有影响。其次，利用人工智能识别假新闻和新闻错误。深度学习很适合用于发现数据中的特定模式，因此，通过训练能让机器自动过滤很多假新闻和新闻错误。目前，很多大科技公司已经在利用类似的算法来识别和过滤假新闻[②]。最后，适时的人工干预。对最终数据进行人工筛查依然是最为可靠的解决方式之一。当然，这种方法随着数据量的增加会越来越不现实和低效，但是这种工作可以依靠计算机辅助，有选择性地对特定数据进行筛查以提高效率。

---

① Michael D. Ward et al., "Comparing GDELT and ICEWS Event Data," *Analysis*, Vol. 21, No. 1, 2013, pp. 267-297.

② Josh Constine, "Facebook Chose To Fight Fake News With AI, Not Just User Reports," *Tech Crunch*, Nov 15, 2016, https://techcrunch.com/2016/11/14/facebook-fake-news/.

(二) 编码系统与编码程序

事件数据库面临的第二个挑战来自自动编码系统和编码程序。目前自动编码系统生成的事件数据库使用的基本都是 CAMEO,而编码程序则有斯洛德特的 TABARI 和 ICEWS 自己开发的 BBK-ACCENT。根据与人工编码的数据比较,目前 BBK-ACCENT 编码程序的准确率可以达到 80%左右,但这也意味着仍然有数量相当巨大的错误编码事件存在。错误编码事件自然会影响事件数据库的准确性以及最终测量的双边关系的准确性。目前的编码程序采用的是基于字典的稀疏句法分析(dictionary driven sparse parsing)①,即基于新闻中的有效信息与字典的匹配进行事件分类。这种方法导致计算机对于复杂语言的分析能力并不强,分析语境或者分析使用修辞语的复杂语言的能力很有限。比如,很可能分不清一场拳击赛与一场冲突之间的区别,又或者大量使用比喻等。当然,在主要的媒体新闻尤其是国际新闻中,句法一般不会太复杂。

但进一步而言,这个问题还需要区分随机错误与系统错误。随机错误即错误的出现是无规律的。就 GDELT 中的测量而言,错误出现在合作和冲突事件库中的可能性是一样的,这种情况对于最终结果的影响相对要小,尤其是当数据量极大的时候,这种影响甚至可以忽略。而如果是系统性的错误,那么对结果的影响就将是显著的。比如,如果编码程序的错误导致其偏好合作事件,那么最后测量的关系值必然也会偏向合作。系统性的错误无法通过增加数据量来解决。

对于编码程序的质量和错误问题,一方面是要增加数据量,解决大数据问题的一个重要方法永远是更多更大的数据,以此来使得其中的错误微不足道,不会影响到对关系的测量。另一个途径则是增加方法和数据的透明度,以便学界能够清楚地知道数据的产生过程,并能共同改进。在 2013 年 4 月 1 日以后的事件中,GDELT 都提供了信息来源,以方便研究者进行进一

① Philip Schrodt and Jay Yonamine, "A Guide To Event Data: Past, Present, And Future," *All Azimuth: A Journal of Foreign Policy and Peace*, Vol. 2, No. 2, 2013, pp. 12-13.

步的数据筛查，而这将有助于事件数据准确性的提高。最后一个也是最基本的方法，是改进编码系统和编码程序的分析能力。目前学界已经在积极地改进编码系统和编码程序，新一代系统（Patriarch 2）正在开发之中。这一发展将进一步使得利用大数据进行双边关系研究更具有可信度。目前，飞速发展的人工智能深度算法分析语言和句法的能力大大增强，在智能手机上广泛使用的语音识别工具（如苹果的 SIRI），其所表现出的语言识别能力已经非常惊人，将它与事件分析相结合，将能很大程度解决目前编码程序遇到的问题。①

（三）事件的性质与数量的关系

关于双边关系的量化度量，一致存在的争论是关于事件性质与数量的取舍。对于双边关系来说，对单个事件赋值与只考虑事件的数量，哪个能更加有效地测量双边关系值呢？前文对这些问题已经有所讨论，很大程度上，对事件权重与数量的取舍取决于研究问题和对于双边关系的一般判断。一般而言，当事件数量很少时，就不得不考虑事件的权重；而当事件数量非常巨大时，单个事件的权重就不那么重要了。

另一个问题是，言语和行为需要区别对待吗？在政府与政府的交往中，言语行为与实际行为往往一样重要，它们都对双边关系有着巨大影响。比如，双方领导人关于双边关系定位的言论对双边关系起指导作用。因此，言语本身是测量双边关系的一个重要指标，不需要将它与实际行为分离考虑。

当然，在实际进行测量的时候，依然需要针对具体研究问题来设计，设计取决于我们更关注重大事件还是整体图景对双边关系的影响。在设计时，要根据研究问题来选择变量的测量。

（四）数据来源的转换翻译问题

目前的事件数据库都基于英语，而自动编码程序使用的字典也是英文的，所以其他语言的新闻是通过谷歌翻译然后再用程序进行抓取。这会直

---

① John Beieler，“Generating Politically-relevant Event Data，” *arXiv*，2016，*preprint arXiv*：1609.06239.

接导致一个问题,即翻译过程中的意义损失。没有任何翻译是完美的,尤其是机器翻译。虽然随着人工智能的发展,机器翻译的质量有了飞速提高,但是由于语言对于社会背景与传统的依赖,不同语言之间的翻译依然意味着很多意义可能流失,进而影响对事件的理解。

这个问题的解决方法是给不同语言的新闻编辑以同种语言的字典。关于中文句法分析,在 Stanford Universal Dependencies 项目下已经有了中文字典,但是事件数据自动处理中的中文字典还没有。如果将来中国要建立类似 GDELT 的事件数据库,那么编撰自动编码系统的中文字典是首先需要完成的工作之一。

四、结 论

大数据事件库的出现为国际关系研究提供了新的数据来源,也为量化研究和实证检验国际关系理论提供了计算基础。通过对从 GDELT 中数据生成的中美关系值与清华大学中国与大国关系数据库中的中美关系值以及相关的定性研究对比,本文证明利用大数据度量双边关系具有一定的学术价值与进一步发展的潜力。而双边关系作为更为复杂的多边关系的基础,应用在双边关系上的方法也可以很容易地扩展到多边关系的研究上。因此,大数据对整个国际关系的量化具有潜在的重大贡献。本文是这种努力的一个初步尝试。

当然,没有数据和方法是完美的,大数据的事件依然存在种种问题,但是这些问题大都能在某种程度上得到缓解。随着新一代自动编码系统、编码程序以及人工智能的发展,事件数据库的质量也将进一步提升,而利用事件数据库的大数据来进行国际关系研究的前景将更为广阔。研究者在利用这些新工具时,也需要时刻意识到它可能存在的缺陷并通过研究设计来尽量避免这些不足。在大数据时代,最终决定研究质量的仍将是研究者的研究设计和分析方法。①

---

① Gary King:《大数据与数据无关》,钟杨主编:《实证社会科学》(第三卷),上海交通大学出版社,2017 年,第 10—16 页。