沈石，宋长青，程昌秀，等. GDELT：感知全球社会动态的事件大数据. 世界地理研究，2020，29(1)：71-76.［SHEN Shi，SONG Changqing，CHENG Changxiu，et al. GDELT：Big event data for sensing global social dynamics.World Regional Studies，2020，29(1)：71-76.］
DOI：10.3969/j.issn.1004-9479.2020.01.2019800

GDELT：感知全球社会动态的事件大数据

沈 石[1,2,3]，宋长青[1,2,3]，程昌秀[1,2,3]，高剑波[2,3]，叶思菁[1,2,3]
（1. 北京师范大学地表过程与资源生态国家重点实验室，北京 100875；2. 北京师范大学地理科学学部，北京 100875；3. 北京师范大学地理数据与应用分析中心，北京 100875）

摘　要：正确解析国家间政治关系及其演化过程是开展地缘关系研究的重要基础。从大数据的角度开展地缘关系研究为该领域的探索提供了一种新的途径。国家或区域的局部政治倾向数据无法为地缘关系研究提供全面和翔实的数据支撑。论文介绍的一个全新的事件数据库 GDELT（Global Database of Event，Language ，Tone），它在诸多方面弥补了传统数据的不足。该数据不仅详细记录了全球范围事件的发生时间、地点、内容以及参与者信息，而且系统地对事件进行分类和评分。本文从数据内容、事件评分和分类体系三方面详细介绍 GDELT 数据，并总结了该数据的优势和潜在研究方向，以期为我国地缘关系研究等领域提供帮助和参考。

关键词：GDELT；地缘关系；事件数据；戈尔德斯坦量表

0　引言

事件数据作为一种“传感器”揭示了国家之间和国家内部的互动行为，为地缘关系、国际政治、恐怖主义活动等研究提供了关键支撑。自 Charles McClelland 等人于 20 世纪 60 年代首先提出了事件数据概念以来，国外许多学者或机构相继构建了不同主题、范围和类型的事件数据，例如 WEIS（World Event Interaction Survey）[1]，COPDAB（Conflict and Peace data banck）[2]，UCDP GED（Uppsala Conflict Data Program Georeferenced Event Dataset）[3]，KEDS（Kansas Event Data System）[4]和 ACLED（Armed Conflict Location and Event Dataset）[5]等。国内的事件数据则以阎学通等人构建的中国与大国关系事件数据为代表。虽然这些事件数据的来源不一而足，但是新闻报道一直是构建事件数据的核心和主要数据源。

然而，如何从世界范围内不同语种新闻中抽取事件内容及其时空信息仍然是构建大规模全球性事件数据面临的巨大挑战。随着信息和通信技术的发展，尤其是互联网的快速扩张，电子和网络新闻已成为新闻媒体的主流发表形式之一，且数量呈几何级数增长。因此传统人工抓取文章、分析内容、再分类和评分的方法已无法适用于大数据时代背景下事件的构建。为了实现对世界社会动态的感知和计算，Philip A. Schrodt 等提出了基于计算机和机器学习的自动化方法，旨在构建一个全新的大规模全球性事件数据，即 GDELT[6]。目前该数据

收稿日期：2019-09-10；　**修订日期：**2019-12-21
基金项目：第二次青藏高原综合考察研究资助（2019QZKK0608）。
作者简介：沈石（1990-），男，讲师，博士，主要研究方向：地理数据挖掘与分析，E-mail：shens@bnu. edu. cn。
通讯作者：宋长青（1961-），男，教授，主要研究方向：地理区域综合研究、全球化与地缘关系等，E-mail：songcq@bnu. edu. cn。

库依托谷歌公司的海量数据存储、处理和检索能力，时刻搜集世界各国超过100种语言的广播、刊印和网络新闻报道[7]。而且GDELT还通过机器学习模型实现了对新闻内容的自动识别、提取和分类，核心识别人员、位置、组织、数量、主题、数据源、情绪、报价、图片和每秒都在推动全球社会的事件，并依据戈尔德斯坦量表对不同类型事件进行评分。从1979年至今，GDELT涵盖了与政治、军事、外交、经济等社会动态相关的事件上亿条，仅2018年的数据量就已经超过了2. 5TB。

目前基于GDELT数据的研究报道开始逐渐增多，主要集中在地缘关系[8]、国际关系[9-13]、恐怖主义活动[14]、冲突强度与影响[15,16]，社会态势与风险[17,18]、区域一体化[19]、原油价格预测[20]和新闻文化传播[21]等领域。但是受研究报道篇幅的限制，这些研究对于GDELT的描述和介绍不够清晰和深入。而且GDELT的原始说明内容庞杂，缺少相关中文资料。这些都在很大程度上阻碍了研究者对该数据的深刻认识和理解，进而影响分析数据和开展相关研究。因此本文主要介绍GDELT数据的组成结构和基本含义，分析事件分类体系和评分方法，旨在为国内相关研究提供借鉴和参考。

1 GDELT概览

GDELT是一个全面记录事件和事件参与者的时空数据集，其核心是对事件及其参与者信息的自动化识别、概化、分类和编码。2013年GDELT数据库公布的第一个版本包含自1979年以后的所有事件，更新频率为1天。2015年GDELT的第二个版本更新频率提高到了15分钟，但时间范围仅从2015年2月19日开始。截止2018年底，GDELT记录的事件总数超过6亿条（图1）。因为GDELT巨大的数量体积和非常高的更新频率，它主要依托于谷歌公司的BigQuery服务实现在线查询和检索数据。

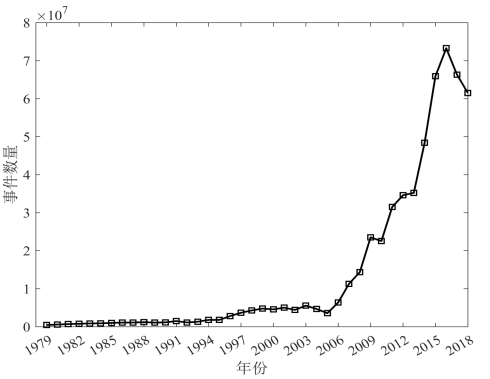

图1 GDELT事件数量年际变化（1979-2018）

Fig.1 Inter-annual variation of the number of events in GDELT during 1979-2018

每条GDELT数据记录主要由5个部分组成，分别是事件编号和日期、事件参与者、事件动作、事件地理信息以及数据管理。表1是GDELT数据记录的详细信息，包括数据字段名称、编号以及中文含义事件编号由一串递增的整数型编码组成，是每条事件记录的唯一标识。日期包括了事件发生的具体日期，并且在第二版中精确到15分钟。

事件参与者描述了事件两个参与者的属性。GDELT一般使用参与者1（Actor1）和参与者2（Actor2）分别指明事件的发起者和目标。参与者的属性主要包含3个内容，分别是参与者代码、参与者名称和参与者国家代码。GDELT采用CAMEO（Conflict and Mediation Event Observations）[22]中的人员分类和编码方法对参与者进行识别、简化和编码。参与者代码表征了参与者的“地理、阶层、民族、宗教信仰和其他身份信息（例如政治精英、军官、反对派等）”[6]。参与者名称记录了“政治领袖姓名、政治组织正式名称、有关国家的首都或主要城市名称”[6]。参与者国家代码标识了参与者所属国家的ISO三位编码，可以用于定位参与者所在国家。参与者属性值为空表明无法识别事件的参与者。

事件动作刻画了事件发起者对目标行动的各类属性,并通过戈尔德斯坦评分等方法评估事件的重要性及其造成的影响。事件动作主要包括4个内容,分别是事件类型、戈尔德斯坦量表分值、报道次数和平均语调。GDELT同样是基于CAMEO中的事件分类体系对事件性质进行归类和编码。事件类型表征了该事件的政治、军事和经济等性质,可以用以实现对目标事件的筛选。在对事件分类的基础上,GDELT借助戈尔德斯坦量表衡量事件的影响性质和程度。此外,事件动作还提供了事件报道次数信息,主要包括提及该事件的次数、提及该事件的消息来源数量、提及该事件的新闻文章数量。这些信息从新闻报道的角度刻画了事件的重要程度。最后平均语调给出了"提及该事件的所有文章语调的平均值"[6]。平均语调值介于-100到+100之间;正负分别表示积极的语调和消极的语调;绝对值表示语调强弱程度。平均语调可以用作衡量一个事件的重要性及其影响的辅助指标。

表1　GDELT数据字段说明

Tab. 1　Instruction for the GDELT data fields

字段名称	字段编号	中文含义
Event ID	1	事件编号
Date	2-5	日期
Actor1	6-15	参与者1
Actor2	16-25	参与者2
Event Action	26-35	事件性质
Actor Geography	36-49	参与者地理信息
Action Geography	50-56	事件相关地理信息
Date Added	57	事件录入日期
Source URL	58	来源URL

事件地理信息描述了参与者与事件的地理位置信息,包括参与者和事件的"位置信息类型,位置全称、国家编码、一级行政区编码,经维度及GNS(GEOnet Names Server)或GNIS(Geographic Names Information System)地名辞典标识符"[6]。研究者借助事件地理信息可以对事件及其参与者进行更精确的空间定位,进而实现GDELT数据的空间化及空间分析。

GDELT还提供了事件添加时间和来源链接两个数据管理信息。事件添加时间是指一条事件记录被添加入GDELT数据库的的日期。来源链接记录了报道事件的新闻文章来源的URL(Uniform Resource Locator,统一资源定位符)信息。该字段一般出现在2013年4月1日之后的事件记录中。需要明确的是,即使存在多篇文章提及同一事件,GDELT仍然仅给出其中一条文章的URL链接。添加事件和来源链接两个数据字段可以帮助研究者定位到事件的原始消息源,从而为感兴趣的研究者提供有关事件的更多信息。

2　戈尔德斯坦量表

戈尔德斯坦量表是由戈尔德斯坦提出用以刻画两国之间冲突和合作关系的评分体系[23]。该量表基于国际关系相关概念进行设计,并得到了国际关系学者的验证。利用事件数据的实证研究也表明戈尔德斯坦量表的结果更加符合国际关系的实际情况。

戈尔德斯坦量表构建了一组介于在-10到+10之间的评分,从理论上度量事件对国家产生的潜在影响。负值代表负面影响,正值代表正面影响,0表示中性事件。分值的绝对值表示影响的程度。绝对值越大则表示影响程度越深,反之亦然。GDELT中的每一条事件记录都对应一个戈尔德斯坦分值。需要特别注意的是,戈尔德斯坦量表是根据事件类型对事件进行评分而非事件具体内容。例如不同人数参与的抗议或暴乱事件都具有相同的戈尔德斯坦分值。解决该问题的可能途径之一是结合平均语调、提及次数、消息源数量等其他信息进行综合判断。

3 CAMEO分类体系

GDELT采用的CAMEO编码方法是是对“堪萨斯事件数据库(Kansas Event Data System，KEDS)”的事件和参与者编码方法的扩展和改进。CAMEO的事件类型分为20大类和300多个小类(详细介绍可以参考GDELT或全球新闻事件数据共享平台)。20个大类分别以01-20数字编码进行唯一标识，并且每一大类又分为若干个子类。此外，每一小类编码的由所属的父类编码和自身顺序编号组成。例如为争取权利的暴动(Riot for rights)的事件类型编码为1453，其中最后一位数字3表示该类型在子类中排序，145表示该其父类为暴动(Riot)子类，14表示其所属大类为抗议(Protest)。与KEDS一样，GDELT还对每一类事件赋予了戈尔德斯坦评分。GDELT中的20大类事件类型和对应的戈尔德斯坦评分如表2所示。

表2 GDELT中20大类事件类型及对应戈尔德斯坦评分

Tab.2 Twenty root event types and corresponding Goldstein Scales score in GDELT

类型代码	事件类型	评分
01	Make public statement（公开声明）	0.0
02	Appeal（呼吁）	3.0
03	Express intent to cooperate(表达合作意向)	4.0
04	Consult（商量）	1.0
05	Engage in diplomatic cooperation（开展外交合作）	3.5
06	Engage in material cooperation（开展实质合作）	6.0
07	Provide aid（提供援助）	7.0
08	Yield（妥协）	5.0
09	Investigate（调查）	-2.0
10	Demand（要求）	-5.0
11	Disapprove（不赞成）	-2.0
12	Reject（拒绝）	-4.0
13	Threaten（威胁）	-6.0
14	Protest（抗议）	-6.5
15	Exhibit force posture（炫耀军事力量）	-7.2
16	Reduce relations（减少关系）	-4.0
17	Coerce（胁迫）	-7.0
18	Assault（攻击）	-9.0
19	Fight（对抗）	-10.0
20	Use unconventional mass violence（使用非传统的大规模暴力）	-10.0

4 总结与展望

GDELT数据具有数据量大、更新速度快和覆盖面广三个特点。由于实现了对全球社会动态的监测，且全球新闻报道的数量急剧增加，GDELT的数据量十分庞大。尤其是第二版GDELT数据的更新频率提高到15分钟，成为一种近乎实时的事件数据。而且因为GDELT涵盖了全球超过100种语言的新闻报道，其覆盖面也是到目前为止事件数据中最广泛的。

GDELT数据的分类体系和评分相对更科学。GDELT中事件和参与者的编码是基于更加系统和科学的CAMEO编码方法。该方法能够更加全面覆盖事件数据类型和参与者，简洁而细致地刻画出两个参与者，尤其是国家的角色以及国际互动的行为特征。此外，戈尔德斯坦量表与CAMEO分类方法地联合使用能够比较客观、真实地量化参与者之间互动关系的性质和程度。

当然，GDELT数据并不是完美无缺的。首先，GDELT数据设计初衷和面向的用户更多关注于区域或全球尺度的国际政治、军事和经济关系研究，对文化和体育等人文领域相关的

互动重视不足。其次,戈尔德斯坦量表对于事件的评分主要是基于美国国际关系学者的冲突与合作观念和理论。因此对该分值的使用和解释需要多加注意。最后由于新闻报道的方式和媒介的限制,GDELT数据呈现显著的非线性增长过程。任何基于GDELT中事件数量的分析都应该对这一点进行预处理或说明。

实现对全球社会动态感知仅是认识和理解人类世界以及人地系统的开始。虽然目前国内外学者已经开始注意到GDELT数据蕴含的丰富价值和为解决全球可持续发展的潜力,但是仍需更多的研究者充分利用数据。根据已有研究和GDELT数据特点,我们认为未来可以在以下三个反面更进一步开展GDELT的分析和研究工作:1)基于GDELT的区域或全球性的国家间地缘关系和互动行为研究;2)在国家、区域以及全球尺度上,结合气候变化、难民、传染病等数据探讨跨国问题对社会和政治的影响。3)利用GDELT数据开展我国对外关系和形象研究,为塑造我国国际社会形象,推动"一带一路"倡议等提供支撑。

参考文献(References):

[1] McClelland, C A. World event/interaction survey codebook, in Ann Arbor MI: Inter-university Consortium for Political and Social Research. 1976, ICPSR Ann Arbor, MI.

[2] Azar, E E. The Conflict and Peace Data Bank Project. Journal of Conflict Resolution, 1980,24(1): 143-152.

[3] Sundberg R, Melander E. Introducing the UCDP Georeferenced Event Dataset. Journal of Peace Research, 2013, 50(4): 523-532.

[4] Schrodt P A, Davis S G, Weddle J L. Political Science: KEDS—A Program for the Machine Coding of Event Data. Social Science Computer Review, 1994, 12(4): 561-587.

[5] Raleigh C, Linke A, Hegre H, et al. Introducing ACLED: An Armed Conflict Location and Event Dataset:Special Data Feature. Journal of Peace Research, 2010, 47(5): 651-660.

[6] GDELT Project, 2013. https://www. gdeltproject. org, 2019-12-30.

[7] Leetaru K, Schrodt P A. GDELT: Global data on events, language, and tone, 1979-2012, in International Studies Association annual conference. 2013, San Francisco, CA.

[8] 陈小强,袁丽华,沈石,等, 中国及其周边国家间地缘关系解析. 地理学报, 2019,74(08):1534-1547. [Chen X, Yuan L, Shen S, et al. Analysis of the geo-relationships between China and its neighboring countries. Acta Geographica Sinica, 2019, 74(8): 1534-1547.]

[9] 沈石, 袁丽华, 叶思菁, 等, 近40年中美地缘政治关系波动及背景解析. 地理科学, 2019, 39(07):1063-1071. [Shen S, Yuan L, Ye S, et al. The Fluctuation and background analysis of geopolitical relations between China and the United States During the Last 40 Years. Scientia Geographica Sinica,2019,39(7):1063-1071.]

[10] 庞珣, 刘子夜. 基于海量事件数据的中美关系分析——对等反应、政策惯性及第三方因素. 世界经济与政治, 2019 (05): 53-79+157-158. [Pang X, Liu Z. China-U. S. relations in massive machine-coded event data: influence of reciprocity, Policy Inertia and a Third Power. World Economics and Politics, 2019(05): 53-79+157-158]

[11] 秦昆, 罗萍, 姚博. GDELT数据网络化挖掘与国际关系分析. 地球信息科学学报, 2019, 21(01):14-24. [Qin K, Luo P, Yao B. Networked mining of GDELT and international relations analysis. Journal of Geo-information Science, 2019, 21(1):14- 24.]

[12] 池志培, 侯娜. 大数据与双边关系的量化研究:以GDELT与中美关系为例. 国际政治科学, 2019, 4(02): 67-88. [Chi Z, Hou N. Big Data and quantitative research on bilateral relations][Take GDELT and Sino-US relations as an example. Quarterly Journal of International Politics, 2019, 4(02): 67-88.]

[13] Yuan Y, Liu Y, Wei G. Exploring inter-country connection in mass media: A case study of China. Computers, Environment and Urban Systems, 2017. 62: 86-96.

[14] Gao J, Fang P, Liu F. Empirical scaling law connecting persistence and severity of global terrorism. Physica A: Statistical Mechanics and its Applications, 2017, 482:74-86.

[15] Levin N, Ali S, Crandall D. Utilizing remote sensing and big data to quantify conflict intensity: The Arab Spring as a case study. Applied Geography, 2018, 94:1-17.

[16] Levin N, Ali S, Crandall D, et al. , World Heritage in danger: Big data and remote sensing can help protect sites in conflict zones. Global Environmental Change, 2019,55: 97-104.

[17] 马明清,袁武,葛全胜,等. "一带一路"若干区域社会发展态势大数据分析. 地理科学进展, 2019,38(07):1009-1020. [Ma M, Yuan W, Ge Q, et al. LI. Big data analysis of social development situation in regions along the Belt and Road. Progress In Geography, 2019, 38(7): 1009-1020]

[18] Zhang C, Xiao C, Liu H. Spatial Big Data Analysis of Political Risks along the Belt and Road. Sustainability, 2019, 11 (8): 2216.

[19] 刘毅, 王云, 杨宇, 等. 粤港澳大湾区区域一体化及其互动关系. 地理学报, 2019, 74(12): 2455-2466. [Liu Y, Wang Y, Yang Y, et al. Regional integration and interaction of the Guangdong-Hong Kong-Macao Greater Bay Area. Acta Geographica Sinica, 2019, 74(12): 2455-2466.]

[20] Elshendy M, Colladon A F, Battistoni E, et al. Using four different online media sources to forecast the crude oil price. 2018, 44(3): 408-421.

[21] 龚为纲, 朱萌, 张赛, 等. 媒介霸权、文化圈群与东方主义话语的全球传播——以舆情大数据GDELT中的涉华舆情为例. 社会学研究, 2019, 34(05):138-164+245. [Gong W, Zhu M, Zhang S, et al. . Media hegemony, cultural circle and the global dissemination of the orientalist discourse: taking public opinion on China in GDELT as an example. Sociological Studies, 2019, 34(05):138-164+245]

[22] Gerner D J, Jabr R, Schrodt P A, Conflict and mediation event observations (CAMEO): A new event data framework for the analysis of foreign policy interactions. in International Studies Association annual conference, 2002, New Orleans.

[23] Goldstein J S. A conflict-cooperation scale for WEIS events data. Journal of Conflict Resolution, 1992, 36(2):369-385.

GDELT: Big event data for sensing global social dynamics

SHEN Shi[1,2,3], SONG Changqing[1,2,3], CHENG Changxiu[1,2,3], GAO Jianbo[2,3], YE Sijing[1,2,3]

(1. State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing 100875, China; 2. Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China; 3. Center for Geodata and Analysis, Beijing Normal University, Beijing 100875, China)

Abstract: Properly analyzing the international political relations and their evolution process is an essential for geo-relationships research. Geo-relationships research from the perspective of big data provides a new approach for the exploration in this field. Data reflecting international political trends by traditional methods is unable to provide comprehensive and informative data support for regional or global research. A new event database GDELT (Global Database of Event, Language, Tone) introduced in this article has offset the deficiency of traditional data in many aspects. The data not only records the date-time, place, content and participant of events worldwide but also systematically classifies and scores events. This article introduces GDELT data in detail from three aspects of data content, event scoring, and classification system, and summarize the advantages and potential research directions of the data, intending to provide help and reference for China's geo-relationships research and other fields.

Key words: GDELT; geo-relationships; event data; Goldstein scale