Event data in forecasting models: Where does it come from, what can it do?

Philip A. Schrodt

Parus Analytics
Charlottesville, Virginia, USA
schrodt735@gmail.com

Paper presented at the Conference on Forecasting and Early Warning of Conflict, Peace Research Institute, Oslo
April 22, 2015

Why is event data suddenly attracting attention after 50 years?

- Rifkin [NYT March 2014]: The most disruptive technologies in the current environment combine network effects with zero marginal cost
- Key: zero *marginal* costs even though open source software is still "free-as-in-puppy"
- Examples
  - Operating systems: Linux
  - General purpose programming: gcc, Python
  - Statistical software: *R*
  - Encyclopedia: Wikipedia
  - Scientific typesetting and presentations: LaTeX

EL:DIABLO
Event Location: Dataset in a Box, Linux Option

- Open source: `https://openeventdata.github.io`
- Full modular open-source pipeline to produce daily event data from web sources. `http://phoenixdata.org`
- Scraper from white-list of RSS feeds and web pages
- Event coding from any of several coders: TABARI, PETRARCH, others
- Geolocation: "Cliff" open source geolocater
- "One-A-Day" deduplication keeping URLs of all duplicates
- Designed for implementation in inexpensive Linux cloud systems
- Supported by Open Event Data Alliance `http://openeventdata.org`

An incident must first generate one or more texts

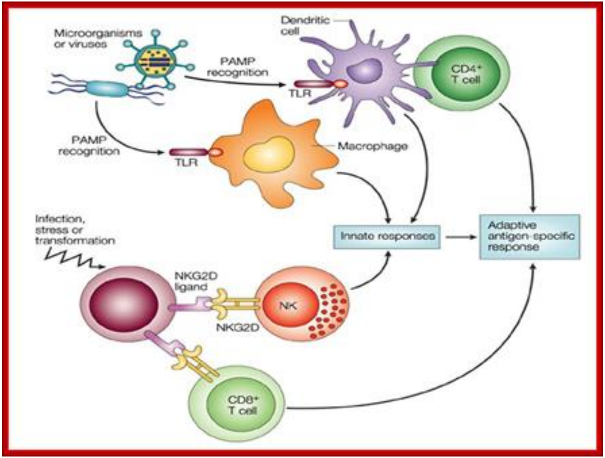
This is the biggest challenge to accuracy. At least the following factors are involved

- A reporter actually witnesses, or learns about, the incident
- An editor thinks incident is "newsworthy": This has a bimodal distribution of routine incidents such as announcements and meeting, and high-intensity incidents: "when it bleeds, it leads."
- Report is not formally or informally censored
- Report corresponds to actual events, rather than being created for propaganda or entertainment purposes
- News coverage is biased towards the coverage of certain geographical regions, and generally "follows the money"
- Reports will be amplified if they are repeated in additional sources

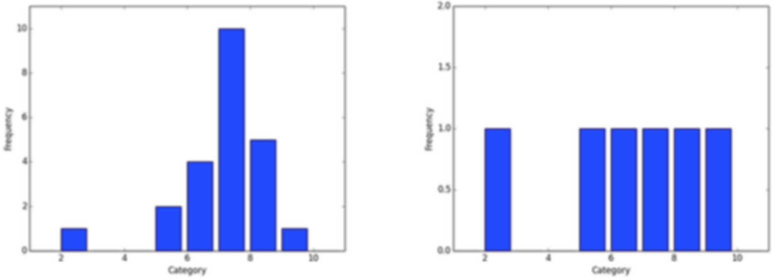
Humans use multiple sources to create narratives

- Redundant information is automatically discarded
- Sources are assessed for reliability and validity
- Obscure sources can be used to "connect the dots"
- Episodic processing in humans provides a pleasant dopamine hit when you put together a "median narrative": this is why people read novels and watch movies.

Machines latch on to anything that looks like an event



This must be filtered



(a) What was generated (b) What remains

Figure 2: Effect of One-A-Day filtering

Implications of one-a-day filtering

- Expected number of correct codes from a single incident increases exponentially but is asymptotic to 1
- Expected number of incorrect codings increases linearly and is bounded only by the number of distinct codes

Tension in two approaches to using machines [Isaacson]

- "Artificial intelligence" [Turing, McCarthy]: figure out how to get machines to think like humans
- "Computers are tools" [Hopper, Jobs]: Design systems to optimally *complement* human capabilities

Does this affect the common uses of event data?

- Trends and monitoring: probably okay, at least for sophisticated users
- Narratives and trigger models: a disaster
- Structural substitution models: seem to work pretty well because these are usually based on approaches that extract signal from noise
- Time series models: also work well, again because these have explicit error models
- Big Data approaches: who knows?

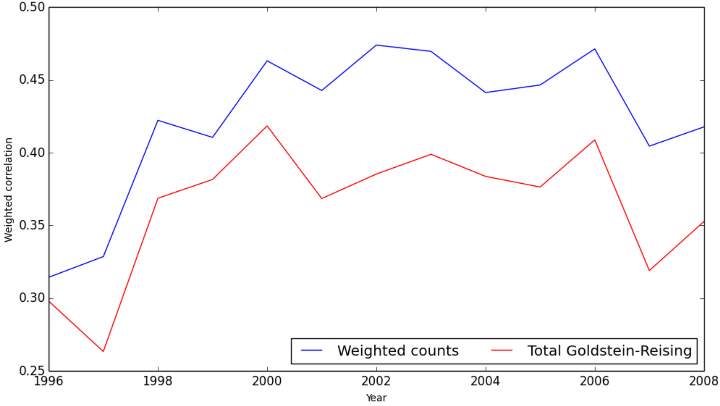
Weighted correlation between two data sets

$$wtcorr = \sum_{i=1}^{A-1} \sum_{j=i}^{A} \frac{n_{i,j}}{N} r_{i,j} \tag{1}$$

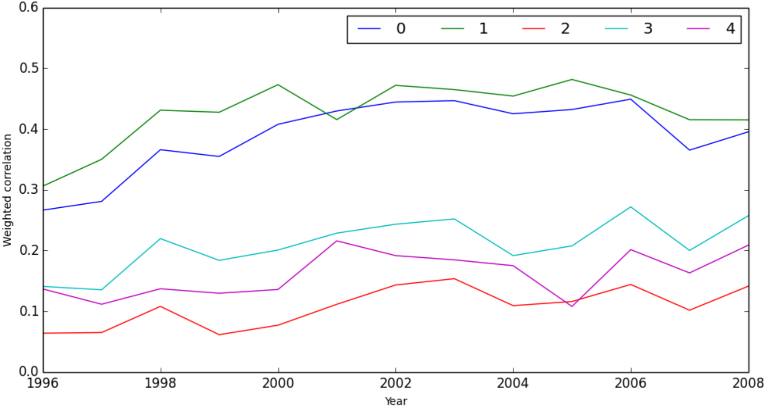
where

- A = number of actors;
- $n_{i,j}$ = number of events involving dyad i,j
- N = total number of events in the two data sets which involve the undirected dyads in A x A
- $r_{i,j}$ = correlation on various measures: counts and Goldstein-Reising scores

Correlations over time: total counts and Goldstein-Reising totals



Correlations over time: pentacode counts



Dyads with highest correlations

Table 1: Fifty dyads with highest average correlation on total counts

RUS-CHN 0.76	CHN-ZAF 0.72	CHN-EGY 0.67	CHN-PAK 0.66	CHN-DEU 0.66
CHN-SYR 0.66	CHN-HRV 0.65	CHN-JPN 0.64	RUS-JPN 0.63	UKR-HRV 0.63
RUS-IRN 0.61	CHN-FRA 0.60	CHN-ROU 0.60	CHN-IND 0.59	CZE-HRV 0.59
CHN-GBR 0.59	CHN-MEX 0.59	RUS-PSE 0.59	CHN-LKA 0.59	CHN-VNM 0.59
HRV-ROU 0.58	CHN-PSE 0.58	RUS-IND 0.58	RUS-DEU 0.57	TUR-POL 0.57
CHN-TUR 0.57	IRN-PAK 0.56	CHN-IRN 0.56	IRN-TUR 0.56	RUS-VNM 0.56
IRN-SYR 0.56	CHN-BRA 0.55	CHN-ESP 0.55	RUS-GBR 0.55	TUR-UKR 0.55
DEU-ROU 0.54	USA-CHN 0.54	RUS-CAN 0.54	CHN-AUS 0.54	RUS-EGY 0.54
CHN-ARG 0.54	RUS-ISR 0.54	TUR-ROU 0.54	RUS-SYR 0.54	RUS-POL 0.54
UKR-SVK 0.54	TUR-GEO 0.53	RUS-ROU 0.53	PSE-PAK 0.53	RUS-KOR 0.53

Dyads with lowest correlations

Table 2: Fifty dyads with lowest average correlation on total counts

MEX-SAU -0.0090	AUS-ITA -0.0086	GBR-VEN -0.0060	ISR-BGD -0.0060	AFG-SYR -0.0050
BRA-POL -0.0047	AFG-LKA -0.0045	SAU-NZL -0.0043	AUS-CZE -0.0042	CZE-LKA -0.0038
IDN-AZE -0.0037	ITA-NZL -0.0031	PRK-SAU -0.0030	IRQ-ZWE -0.0030	IND-ARG -0.0029
NPL-CAN -0.0028	PHL-LKA -0.0028	BRA-ITA -0.0027	VNM-SAU -0.0025	ESP-MYS -0.0025
NGA-LBN -0.0025	NGA-ITA -0.0025	PHL-ARG -0.0024	PSE-GEO -0.0024	IRN-NPL -0.0023
AZE-MYS -0.0022	GEO-SYR -0.0022	EGY-MEX -0.0022	BGD-SYR -0.0021	CAN-NZL -0.0020
TWN-EGY -0.0020	PRK-KEN -0.0019	COL-BGD -0.0018	PRK-LBN -0.0018	EGY-VEN -0.0018
CZE-VEN -0.0016	KOR-GEO -0.0016	KOR-VEN -0.0015	TUR-VEN -0.0015	NGA-VNM -0.0015
PHL-KEN -0.0015	SVK-SAU -0.0015	AFG-BRA -0.0015	SVK-ZWE -0.0015	AFG-VEN -0.0015
GEO-SAU -0.0015	KOR-ZWE -0.0015	SYR-ARG -0.0015	PSE-MEX -0.0014	ZAF-NZL -0.0014

What is to be done: Part 1

- Open-access gold standard cases, then use the estimated classification matrices for statistical adjustments
- Systematically assess the trade-offs in multiple-source data, or create more sophisticated filters
- Evaluate the utility of multiple-data-set methods such as multiple systems estimation
- Systematic assessment of the native language versus machine translation issue
- Extend CAMEO and standardize sub-state actor codes: canonical CAMEO is too complicated, but ICEWS substate actors are too simple

What is to be done: Part 2

- Automated verb phrase recognition and extraction: this will also be required to extend CAMEO. Entity identification, in contrast, is largely a solved problem (ICEWS: 100,000 actors in dictionary)
- Establish a user-friendly open-source collaboration platform for dictionary development
- Systematically explore aggregation methods: ICEWS has 10,742 aggregations, which is too many
- Solve—or at least improve upon—the open source geocoding issue
- Develop event-specific coding modules

Thank you

Email:
schrodt735@gmail.com

Slides:
`http://eventdata.parusanalytics.com/presentations.html`

Data: `http://phoenixdata.org`

Software: `https://openeventdata.github.io/`

Papers:
`http://eventdata.parusanalytics.com/papers.html`