Statystyczna Analiza Danych – laboratorium

Paradygmat tidy

Dorota Celińska-Kopczyńska

Uniwersytet Warszawski

Artykuł

Wickham, Hadley. 2014. *Tidy Data*. Journal of Statistical Software, vol. 59.

Motywacja

- Wyczyszczenie danych i przygotowanie ich do dalszej analizy pochłania wiele wysiłku
- Z kolei... niewiele wysiłku wkładane jest w badania jak to przygotowanie danych prowadzić efektywnie
- Każdy zabałaganiony zbiór danych jest zabałaganiony na swój własny sposób, ale oczyszczone dane powinny zachowywać się zgodnie z pewnymi zasadami

Dane uporządkowane (*tidy data*)

- Trzecia postać normalna Codda przeformułowana na potrzeby języka statystycznego
- Tutaj skupimy się na pojedynczym zbiorze danych (tabeli) niż na wielu połączonych w relacyjnej bazie danych

Zasady *tidy data*

- Każda zmienna tworzy kolumnę
- Każda obserwacja tworzy wiersz
- Każdy typ jednostki obserwacyjnej tworzy tabelę

Pięć najczęstszych problemów

- Nagłówki kolumn to ich wartości a nie nazwy zmiennych
- Wiele zmiennych trzymanych w jednej kolumnie
- Zmienne trzymane zarówno w wierszach jak i kolumnach
- Wiele typów jednostek obserwacyjnych trzymanych w tej samej tabeli
- Pojedyncza jednostka obserwacyjna trzymana w wielu tabelach

Nagłówki kolumna jako wartości zamiast nazw zmiennych

- Zwykle w danych tabelarycznych na potrzeby prezentacji
- Niekiedy może być użyteczne! W szczególności, gdy korzystamy z operacji macierzowych

religion	<$10k	$10–20k	$20–30k	$30–40k	$40–50k	$50–75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

Table 4: The first ten rows of data on income and religion from the Pew Forum. Three columns, $75–100k, $100–150k and >150k, have been omitted.

Dane surowe, wytop (*melting*) i stopione (*molten*) dane

- Zbiór danych z Tabeli 4 zawiera trzy zmienne: *religion*, *income* i *frequency*
- Aby ten zbiór stał się *tidy* musimy przeprowadzić **wytop** (*melting*) – kolumny muszą stać się wierszami
- Niektórzy mogli już zetknąć się z tą procedurą pod nazwą **transformacja zbioru danych z reprezentacji szerokiej (wide) na wąską (long)**.
- Wprowadzimy dwie nowe zmienne: jedną zawierającą nazwy kolumn, a drugą z powiązanymi wartościami liczbowymi
- Wynikiem będzie **stopiony** *molten* zbiór danych.

Surowe dane, wytop i stopione dane – mały przykład

row	a	b	c
A	1	4	7
B	2	5	8
C	3	6	9

(a) Raw data

row	column	value
A	a	1
B	a	2
C	a	3
A	b	4
B	b	5
C	b	6
A	c	7
B	c	8
C	c	9

(b) Molten data

Table 5: A simple example of melting. (a) is melted with one colvar, row, yielding the molten dataset (b). The information in each table is exactly the same, just stored in a different way.

Wiele zmiennych trzymanych w jednej kolumnie

- Wytop może skończyć się trzymaniem wielu nazw zmiennych w jednej kolunnie
- Aby taki zbiór stał się *tidy* z reguły należy użyć heurystyk (od zwykłego podziału po znakach aż po wyrażenia regularne)
- Zbiory danych zgodne z paradygmatem tidy umożliwiają łatwiejszą pracę z wartościami zmiennych (niższa liczba kombinacji)

Wiele zmiennych trzymanych w jednej kolumnie – przykład

country	year	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f014
AD	2000	0	0	1	0	0	0	0	—	—
AE	2000	2	4	4	6	5	12	10	—	3
AF	2000	52	228	183	149	129	94	80	—	93
AG	2000	0	0	0	0	0	0	1	—	1
AL	2000	2	19	21	14	24	19	16	—	3
AM	2000	2	152	130	131	63	26	21	—	1
AN	2000	0	0	1	2	0	0	0	—	0
AO	2000	186	999	1003	912	482	312	194	—	247
AR	2000	97	278	594	402	419	368	330	—	121
AS	2000	—	—	—	—	1	1	—	—	—

Table 9: Original TB dataset. Corresponding to each 'm' column for males, there is also an 'f' column for females, `f1524`, `f2534` and so on. These are not shown to conserve space. Note the mixture of 0s and missing values (—). This is due to the data collection process and the distinction is important for this dataset.

Wiele zmiennych trzymanych w jednej kolumnie – przykład

country	year	column	cases
AD	2000	m014	0
AD	2000	m1524	0
AD	2000	m2534	1
AD	2000	m3544	0
AD	2000	m4554	0
AD	2000	m5564	0
AD	2000	m65	0
AE	2000	m014	2
AE	2000	m1524	4
AE	2000	m2534	4
AE	2000	m3544	6
AE	2000	m4554	5
AE	2000	m5564	12
AE	2000	m65	10
AE	2000	f014	3

(a) Molten data

country	year	sex	age	cases
AD	2000	m	0–14	0
AD	2000	m	15–24	0
AD	2000	m	25–34	1
AD	2000	m	35–44	0
AD	2000	m	45–54	0
AD	2000	m	55–64	0
AD	2000	m	65+	0
AE	2000	m	0–14	2
AE	2000	m	15–24	4
AE	2000	m	25–34	4
AE	2000	m	35–44	6
AE	2000	m	45–54	5
AE	2000	m	55–64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

(b) Tidy data

Table 10: Tidying the TB dataset requires first melting, and then splitting the `column` column into two variables: `sex` and `age`.

Zmienne trzymane zarówno w wierszach jak i kolumnach

- Najbardziej skomplikowana forma zabałaganionych danych

id	year	month	element	d1	d2	d3	d4	d5	d6	d7	d8
MX17004	2010	1	tmax	—	—	—	—	—	—	—	—
MX17004	2010	1	tmin	—	—	—	—	—	—	—	—
MX17004	2010	2	tmax	—	27.3	24.1	—	—	—	—	—
MX17004	2010	2	tmin	—	14.4	14.4	—	—	—	—	—
MX17004	2010	3	tmax	—	—	—	—	32.1	—	—	—
MX17004	2010	3	tmin	—	—	—	—	14.2	—	—	—
MX17004	2010	4	tmax	—	—	—	—	—	—	—	—
MX17004	2010	4	tmin	—	—	—	—	—	—	—	—
MX17004	2010	5	tmax	—	—	—	—	—	—	—	—
MX17004	2010	5	tmin	—	—	—	—	—	—	—	—

Table 11: Original weather dataset. There is a column for each possible day in the month. Columns d9 to d31 have been omitted to conserve space.

Zmienne trzymane zarówno w wierszach jak i kolumnach

id	date	element	value
MX17004	2010-01-30	tmax	27.8
MX17004	2010-01-30	tmin	14.5
MX17004	2010-02-02	tmax	27.3
MX17004	2010-02-02	tmin	14.4
MX17004	2010-02-03	tmax	24.1
MX17004	2010-02-03	tmin	14.4
MX17004	2010-02-11	tmax	29.7
MX17004	2010-02-11	tmin	13.4
MX17004	2010-02-23	tmax	29.9
MX17004	2010-02-23	tmin	10.7

(a) Molten data

id	date	tmax	tmin
MX17004	2010-01-30	27.8	14.5
MX17004	2010-02-02	27.3	14.4
MX17004	2010-02-03	24.1	14.4
MX17004	2010-02-11	29.7	13.4
MX17004	2010-02-23	29.9	10.7
MX17004	2010-03-05	32.1	14.2
MX17004	2010-03-10	34.5	16.8
MX17004	2010-03-16	31.1	17.6
MX17004	2010-04-27	36.3	16.7
MX17004	2010-05-27	33.2	18.2

(b) Tidy data

Table 12: (a) Molten weather dataset. This is almost tidy, but instead of values, the `element` column contains names of variables. Missing values are dropped to conserve space. (b) Tidy weather dataset. Each row represents the meteorological measurements for a single day. There are two measured variables, minimum (`tmin`) and maximum (`tmax`) temperature; all other variables are fixed.

Wiele typów w jednej tabeli

- Zbiory danych mogą zawierać wartości zebrane na różnych poziomach agregacji, pochodzące z różnych typów jednostek obserwacyjnych (np. dane na poziomie województwa, powiatu, czy jednostki)
- Rozwiązanie tego problemu bezpośrednio wiąże się z normalizacją na potrzeby relacyjnego modelu baz danych
- **Istnieje niewiele narzędzi analitycznych pozwalających na bezpośrednią pracę z danymi w formacie relacyjnym!**

Jeden typ w wielu tabelach

- Występuje, gdy dane o jednostce obserwacji są rozrzucone po różnych tabelach lub plikach
- Rozwiązanie:
  - Wczytaj pliki do listy tabel
  - Do każdej z tabel dodaj nową kolumnę przechowującą oryginalną nazwę pliku – to często jest wartość ważnej zmiennej
  - Połącz wszystkie tabele w jedną

Manipulacja danymi

- **Filtrowanie** – wyciąganie podzbiorów lub usuwanie obserwacji na podstawie pewnego warunku
- **Transformacja** – dodanie lub modyfikacja zmiennych
- **Agregacja** – łączenie wartości wielu zmiennych w jedną wartość
- **Sortowanie** – zmiana uporządkowania obserwacji

Manipulacja danymi – narzędzia

- **Filtrowanie** – `base::subset()`
- **Transformacja** – `base::transform()`
- **Agregacja** – `plyr::summarise()`
- **Sortowanie** – `plyr::arrange()`
- **Praca z podzbiorami** – `base::by()`, `plyr::ddply()`
- **Łączenie zbiorów** – `base::merge()`, `plyr::join()`

Wizualizacja

- Narzędzia służące wizualizacji wymagają jedynie, żeby dane wejściowe były *tidy* – wynik działań jest graficzny
- Wizualizacja jako mapowanie pomiędzy zmiennymi i estetycznymi cechami wykresu
- Istnieją narzędzia do wizualizacji zabałaganionych danych

Wizualizacja – narzędzia

- Dane *tidy*: `base::plot()`, `lattice`, `ggplot2`.
- Zabałaganione dane: `base::barplot()`, `base::matplot()`, `base::mosaicplot()`...

Modelowanie

- Dane *tidy* przypominają wewnętrzny model danych, jaki jest używany w analizie regresji
- W zależności od struktury danych, niektóre z problemów mogą być rozwiązane poprzez korzystanie z innych technik analizy

Do przemyślenia

- Proszę przeczytać i przemyśleć Sekcję 5 – Case study.
- Czy wydaje się Państwu, że dane *tidy* ułatwiłyby Państwa pracę jako analityków? W jaki sposób?
- Proszę pomyśleć o ograniczeniach lub trudnościach związanych z korzystaniem z danych *tidy*
- Proszę przypomnieć sobie zbiory danych, z którymi już Państwo się zetknęli – czy były *tidy*? Jeśli nie, to co należałoby zrobić, aby takie się stały?

Dyskusja

- Dane *tidy* mogą nie być najbardziej efektywną formą przechowywania danych
- Jest możliwe, że paradygmat wymaga redefinicji na potrzeby analizy wielowymiarowej
- Restrukturyzacja to tylko część oczyszczania danych – jak usprawnić pozostałe działania?