O'REILLY®

Badanie danych

Raport z pierwszej linii działań

UNIKALNE WPROWADZENIE DO NAUKI O DANYCH!

HELION

Rachel Schutt, Cathy O'Neil

Tytuł oryginału: Doing Data Science: Straight Talk from the Frontline

Tłumaczenie: Zdzisław Płoski

ISBN: 978-83-246-9626-0

Wydawnictwo HELION
ul. Kościuszki 1c, 44-100 GLIWICE
tel. 32 231 22 19, 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *http://helion.pl* (księgarnia internetowa, katalog książek)

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
http://helion.pl/user/opinie/badada
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Printed in Poland.

Spis treści

ROZDZIAŁ 10.

Sieci społeczne[1] i dziennikarstwo danych

W tym rozdziale zajmiemy się dwoma tematami, które nabrały szczególnej aktualności na przestrzeni 5 – 10 ostatnich lat: sieciami społecznościowymi i dziennikarstwem danych. Sieci społeczne (niekoniecznie tylko te *online*) są przedmiotem studiów w instytutach socjologii od dziesięcioleci, podobnie jak ich odpowiednik w instytutach informatyki, matematyki i statystyki — teoria grafów. Jednakże przez pojawienie się sieci społecznych online (sieci społecznościowych), takich jak Facebook, LinkedIn, Twitter i Google+, dysponujemy obecnie nowym, bogatym źródłem danych, które otwiera wiele problemów badawczych, zarówno z socjologicznego, jak i ilościowego lub technicznego punktu widzenia.

Najpierw usłyszymy o tym, jak pewna firma, Morningside Analytics, wizualizuje dane w sieciach społecznościowych i odnajduje w nich znaczenia, oraz poznamy pewne aspekty teorii sieci społecznościowych. Potem przyjrzymy się konstruowaniu opowieści, które da się wywieść z danych powstających w sieciach społecznościowych, co stanowi odmianę dziennikarstwa danych. Rozważanie profilów badaczy danych, mieszanki matematyki, statystyki, komunikacji, wizualizacji i programowania wymaganej do uprawiania nauki o danych lub dziennikarstwa danych — i w tym wypadku formuła genów jest odpowiednią analogią — jest nieco odmienne, lecz zasadnicze umiejętności są takie same. W centrum obu znajduje się zdolność do stawiania dobrych pytań, odpowiadania na nie za pomocą danych i komunikowania swoich odkryć. Mając to na uwadze, zapoznamy się pokrótce z dziennikarstwem danych z perspektywy Jona Brunera, redaktora wydawnictwa O'Reilly.

Analiza sieci społecznych w Morningside Analytics

Pierwszym współautorem tego rozdziału jest John Kelly z Morningside Analytics, który przybył, aby opowiedzieć nam o analizie sieci.

[1] Angielski termin *social networks* ma dwie interpretacje w języku polskim: sieci społeczne (szersza) i sieci społecznościowe (węższa, używana na określenie sieci społecznych zawiązujących się i istniejących w Internecie, w oryginale nazywane *online social networks*); dalej stosujemy oba terminy, zależnie od kontekstu — *przyp. tłum.*

Kelly ma cztery dyplomy z Columbia University: rozpoczął od stopnia BA uzyskanego w 1990 roku w Columbia College, następnie uzyskał tytuły magistra i MPhila[2] oraz obronił doktorat w School of Journalism, gdzie skoncentrował się na socjologii sieci i statystyce w naukach politycznych. Spędził również kilka semestrów w Uniwersytecie Stanforda, ucząc się projektowania przeglądów, teorii gier i innych przedmiotów związanych z analizą ilościową (ang. *quanty stuff*). Pracę magisterską napisał wraz z Markiem Smithem[3] z Microsoftu; jej temat dotyczył ewolucji debat politycznych jako sieci. Po ukończeniu college'u i przed studiami podyplomowymi Kelly zajmował się sztuką, wykorzystując komputery do projektowania dźwięków. Spędził trzy lata jako kierownik mediów cyfrowych w Columbia School of the Arts. Jest również programistą: nauczył się Perla i Pythona, będąc przez rok w Wietnamie ze swoją żoną.

Kelly uważa matematykę, statystykę i informatykę (łącznie z uczeniem maszyn) za narzędzia, którymi musi się posługiwać i które musi dobrze znać, aby robić to, czym naprawdę chce się zajmować. Niczym szef kuchni, potrzebuje dobrych garnków i patelni oraz ostrych noży, rzeczywistym produktem są natomiast potrawy.

Co zatem serwuje w swojej kuchni? Kelly chce zrozumieć, jak ludzie się zwołują, i kiedy to się stanie, jaki jest ich wpływ na politykę i zasady współżycia społecznego. Klientami jego firmy, Morningside Analytics, są think tanki[4] i organizacje polityczne. Chcą one zazwyczaj wiedzieć, w jaki sposób media społecznościowe oddziałują na politykę i ją kreują.

Keely zarabia pieniądze na komunikacji i prezentacjach — wizualizacje są nieodłączną częścią zarówno specjalistycznych ekspertyz, jak i komunikacji — toteż jego kwalifikacje są połączeniem tworzenia wizualizacji i wyciągania z nich wniosków. Morningside Analytics nie opłaca się przecież jedynie odkrywać ciekawych materiałów, firmie zależy na pomaganiu ludziom w ich spożytkowaniu.

Dane przypadek-atrybut a dane sieci społecznościowej

Kelly nie modeluje danych w standardowy sposób, za pomocą danych postaci przypadek-atrybut. Przypadek-atrybut odnosi się do sytuacji, kiedy masz do czynienia z ludźmi, którzy zasilają modele różnymi „przypadkami", dającymi się odnieść do ludzi lub zdarzeń, przy czym i ludzie, i zdarzenia mają różne „atrybuty" dotyczące na przykład wieku lub systemu operacyjnego, lub historii wyszukiwania.

Modelowanie w układzie przypadek-atrybut sięga lat 30. ubiegłego wieku i wczesnych badań rynku, wkrótce zastosowano je również w marketingu, a także w polityce.

Kelly podkreśla istnienie przemożnej skłonności do modelowania z danymi przypadek-atrybut. Możliwym jej wyjaśnieniem jest łatwość przechowywania danych przypadek-atrybut w bazach danych lub łatwość *gromadzenia* takich danych. Tak czy owak, Kelly uważa, że wskutek tego gubi się wiele zagadnień, na które poszukujemy odpowiedzi.

---

[2] Ang. *Master of Philosophy*; amerykański tytuł uniwersytecki — *przyp. tłum.*

[3] Zob. *http://videolectures.net/marc_smith.*

[4] Z założenia niezależne grupy (komitety) doradcze — *przyp. tłum.*

Wspomniał Paula Lazarsfelda i Elihu Katza, dwóch pionierskich socjologów, którzy przybyli z Europy i rozwinęli dziedzinę *analizy sieci społecznych* (ang. *social network analysis*), podejście oparte nie tylko na indywiduach, lecz także na relacjach między nimi.

Aby nabrać wyobrażenia, dlaczego analiza sieci czasami jest ważniejsza od analizy według schematu przypadek-atrybut, zastanówmy się nad następującym przykładem. Rząd federalny sfinansował ankietowanie ludzi w Afganistanie. Chodziło o to, żeby się dowiedzieć, czego chcą mieszkańcy, aby przewidzieć, co stanie się w przyszłości. Jednak, jak wykazuje Kelly, to, co się wydarzy, nie jest prostą funkcją indywidualnych postaw; jest to natomiast pytanie o tych, którzy mają władzę, i to *ich* myślenie trzeba brać pod uwagę.

Wyobraźmy sobie również, że cofamy się w czasie i wykonujemy naukowy sondaż obywateli Europy w 1750 roku, aby określić przyszłą politykę. Gdybyśmy się znali na rzeczy, interesowałoby nas, kto kogo poślubił w rodzinach królewskich.

W pewnym sensie obecne skoncentrowanie na danych przypadek-atrybut jest problemem poszukiwania czegoś „pod latarnią" — rodzajem odchylenia obserwacyjnego sprowadzonego do sytuacji, w której ludzie zwykli postępować w pewien (zazwyczaj łatwiejszy) sposób i trzymają się go nawet wówczas, gdy nie dostają odpowiedzi na nurtujące ich pytanie.

Kelly dowodzi, że świat jest siecią znacznie, ale to znacznie bardziej skomplikowaną niż porcja przypadków z atrybutami. Jeżeli rozumiesz tylko zachowanie jednostek, jak masz powiązać sprawy w całość?

Analiza sieci społecznych

Analiza sieci społecznych wywodzi się z dwóch źródeł: teorii grafów, w której Euler rozwiązał problem siedmiu mostów w Królewcu, i socjometrii, zapoczątkowanej przez Jacoba Moreno w latach 70. XX wieku, w czasie kiedy wczesne komputery sprostały wykonywaniu wielkoskalowych obliczeń na dużych zbiorach danych.

Analiza sieci społecznych została zainicjowana przez Harrisona White'a, emerytowanego profesora Columbia University, w tym samym czasie co prace innego socjologa z tej uczelni, Roberta Mertona. Ich pomysł zasadzał się na założeniu, że działania ludzi muszą pozostawać w związku z ich cechami, lecz aby naprawdę je zrozumieć, trzeba również przyjrzeć się sieciom (tzn. systemom), które *umożliwiają im dane działania*.

Jak przenosimy ten pomysł do naszych modeli? Kelly chce, abyśmy rozważyli to, co on nazywa mikro kontra makro, czyli podział na to, co indywidualne, i to, co ogólnoustrojowe: w jaki sposób przerzucić most nad tymi podziałami? Czy raczej: jak łączyć te podziały w różnych kontekstach?

Na przykład w USA mamy formalne mechanizmy budowania pomostów między podziałem na mikro i makro, mianowicie rynki w przypadku podziału „kupowanie rzeczy" i wybory w przypadku podziałów politycznych. Jednak większość świata nie rozporządza tymi formalnymi mechanizmami, choć często mają tam fikcyjne cienie tych rozwiązań. W większości wypadków musimy dowiedzieć się dostatecznie dużo o faktycznej sieci społecznej, aby wiedzieć, kto sprawuje [w niej] władzę i ma wpływ na zmiany.

Terminologia z obszaru sieci społecznych

Podstawowe jednostki sieci są nazywane *aktorami* lub *węzłami* (ang. *actors*, *nodes*). Mogą to być ludzie lub witryny internetowe, lub nawet dowolne „rzeczy", które bierzesz pod uwagę; obiekty te są często reprezentowane przez jedną kropkę w wizualizacji. Zależności między aktorami są określane jako *powiązania* (ang. *relational ties*) lub *krawędzie* (ang. *edges*). Na przykład to, że się kogoś lubi lub jest znajomym, może być uwidocznione za pomocą *krawędzi*. Pary aktorów określamy jako *diady* (ang. *dyads*), a trójki — jako *triady* (ang. *triads*). Na przykład, jeśli mamy krawędź między węzłem A i węzłem B oraz krawędź między węzłami B i C, to *domknięcie triadyczne* (ang. *triadic closure*) oznaczałoby istnienie krawędzi między węzłem A i węzłem C.

Czasami rozważamy *podgrupy* (ang. *subgroups*), nazywane również *podsieciami* (ang. *subnetworks*), składające się z podzbioru całego zbioru aktorów wraz z ich powiązaniami. Oczywiście oznacza to, że rozważamy także samą *grupę*, przez co rozumie się całą „sieć". Zauważmy, że jest to koncepcja stosunkowo prosta w przypadku — powiedzmy — sieci Twittera, lecz staje się bardzo trudna w przypadku „liberałów".

Przez pojęcie *relacji* (ang. *relation*) rozumiemy na ogół sposób utrzymywania powiązań między aktorami. Na przykład lubienie innej osoby jest relacją, lecz jest nią również zamieszkiwanie z kimś. *Sieć społeczna* (ang. *social network*) jest kolekcją złożoną z pewnego zbioru aktorów i relacji.

W rzeczywistości istnieje kilka różnych typów sieci społecznych. W najprostszym przypadku masz porcję aktorów połączonych więzami. Tą konstrukcją możesz się posługiwać do uwidaczniania grafu Facebooka — dowolne dwie osoby są ze sobą zaznajomione albo nie i każde dwie mogą być teoretycznie znajomymi (przyjaciółmi).

W *grafach dwudzielnych* połączenia istnieją tylko między dwiema formalnie oddzielnymi klasami obiektów. Możesz więc mieć ludzi z jednej strony i firmy z drugiej i możesz połączyć osobę z firmą, jeśli należy ona do zarządu danej firmy. Albo możesz mieć ludzi i rzeczy, które ich potencjalnie interesują, i łączyć ich z nimi, jeśli naprawdę tak jest.

Na koniec są również *sieci ego* (ang. *ego networks*), zazwyczaj formowane jako „część sieci w otoczeniu jednej osoby". Mogłaby to być na przykład „podsieć moich znajomych na Facebooku", którzy w pewnych wypadkach mogą znać się także między sobą. Jak wykazują badania, ludzie o wyższym statusie socjoekonomicznym mają bardziej skomplikowane sieci ego, możesz więc wnioskować o poziomie czyjegoś statusu społecznego, przyglądając się jego sieci ego.

Miary centralności

Pierwsze pytanie, często zadawane przez ludzi w odniesieniu do sieci społecznej, brzmi: *kto tutaj jest ważny?*

Oczywiście znaczenia można nabrać różnymi sposobami i różne definicje, za pomocą których próbuje się uchwycić coś takiego jak ważność, prowadzą do różnych *miar centralności* (ang. *centrality measures*). Podamy tu kilka typowych przykładów.

Po pierwsze, istnieje pojęcie *stopnia* (ang. *degree*). Bierze się tu w rachubę liczbę osób mających z Tobą połączenie. Tak więc w mowie Facebooka jest to liczba posiadanych przez Ciebie znajomych.

Dalej mamy pojęcie *bliskości* (ang. *closeness*). Mówiąc inaczej, jeśli jesteście „bliscy" wszystkim, to powinniście mieć najwyższy wynik bliskości.

Aby wyrazić to ściślej, potrzebujemy pojęcia odległości między węzłami w *grafie spójnym* (ang. *connected graph*), co w przypadku sieci znajomych oznacza, że każda osoba jest połączona z każdą inną za pośrednictwem łańcucha wspólnych znajomych. Odległość między węzłami x i y, zapisywana jako $d(x, y)$, jest definiowana po prostu jako długość najkrótszej ścieżki między dwoma węzłami. Posługując się tą notacją, możesz zdefiniować bliskość węzła x jako sumę:

$$C(x) = \sum 2^{-d(x,y)}$$

wziętą po wszystkich węzłach y różnych od x.

Istnieje też miara centralności zwana *wewnętrznością* (umiejscowieniem pomiędzy, ang. *betweenness*), określająca stopień, w którym ludzie w Twojej sieci znają się za pośrednictwem Ciebie lub — nieco precyzyjniej — czy najkrótsze ścieżki między nimi przechodzą przez [węzeł reprezentujący] Ciebie. Pomysł jest tutaj taki, że jeśli masz dużą miarę wewnętrzności, to informacje prawdopodobnie przechodzą przez Ciebie.

Aby to uściślić, dla każdych dwóch węzłów x i y w tej samej, spójnej części sieci definiujemy $\sigma_{x,y}$ jako *liczbę najkrótszych ścieżek między węzłem x i węzłem y* i $\sigma_{x,y}(v)$ jako *liczbę najkrótszych ścieżek między węzłem x i węzłem y, które przechodzą przez trzeci węzeł v*. Wówczas miara wewnętrzności jest zdefiniowana jako suma:

$$B(v) = \sum \frac{\sigma_{x,y}(v)}{\sigma_{x,y}}$$

wzięta po wszystkich osobnych *parach* węzłów x i y różnych od v.

Ostatnia miara centralności, którą zajmiemy się szczegółowo w podrozdziale „Reprezentacje sieci i centralność wartości własnej", po wprowadzeniu pojęcia macierzy incydencji, nosi nazwę *centralności wektora własnego*. Innymi słowy, osoba, która jest *popularna wraz z popularnymi dziećmi*[5], ma dużą centralność wektora własnego. Przykładem takiej miary centralności jest PageRank w Google.

Branża miar centralności

Jest ważne, aby nie przyjmować bez zastrzeżeń stosowania poprzednich miar centralności. Otóż „ludzie od pomiarów" tworzą branżę, w której każdy próbuje sprzedawać się jako *autorytet*. Doświadczenie mówi nam jednak, że każda [miara] ma swoje wady i zalety. Przede wszystkim należy wiedzieć, że przyglądamy się właściwej sieci lub podsieci.

Na przykład, jeśli poszukujesz bardzo wpływowego blogera wśród Bractwa Muzułmańskiego, i sporządzisz listę 100 największych blogerów w pewnym dużym grafie blogerów, po czym, idąc od jej szczytu w dół, zaczniesz poszukiwać blogera z Bractwa Muzułmańskiego, to nie osiągniesz zamierzonego celu. Znajdziesz kogoś, kto jest wpływowy zarówno w dużej sieci, jak i bloguje dla Bractwa Muzułmańskiego, lecz nie będzie to osoba wpływowa *wśród* Braci Muzułmańskich, raczej wśród ponadnarodowych elit w większej sieci. Innymi słowy, musisz mieć na uwadze lokalne sąsiedztwo w grafie.

[5] W rozumieniu węzłów w grafie — *przyp. tłum.*

Inny problem z miarami centralności: z doświadczenia wynika, że różne konteksty wymagają różnych narzędzi. Coś może się nadawać do blogów, lecz gdy pracujesz z danymi Twittera, może Ci być potrzebne coś zupełnie innego.

Jedną z przyczyn są różnice w danych, inną — różne sposoby, za pomocą których ludzie *grają* miarami centralności. Na Twitterze na przykład ludzie tworzą 5000 twitterowych [ro]botów, które podążają jeden za drugim i za innymi strategicznie wyselekcjonowanymi (prawdziwymi) ludźmi, aby sprawiać, że wyglądają oni na wpływowych według pewnych miar (być może według centralności wektora własnego). Z oczywistych powodów nie jest to trafne; to tylko efekt grania przez kogoś miarami.

Istnieją już pewne pakiety sieciowe, które potrafią obliczać różne, wymienione uprzednio miary centralności. Można tu wskazać NetworkX[6] lub igraph[7], jeśli używasz Pythona, lub statnet[8] dla R, albo NodeXL[9], jeśli wolisz Excela, a na koniec rzuć okiem na mający się ukazać pakiet w języku C autorstwa Jure'a Leskoveca z Uniwersytetu Stanforda[10].

Eksperyment myślowy

Należysz do elity, dobrze opłacanego think tanku w DC. Możesz wynająć ludzi i masz do wydania 10 milionów dolarów. Twoim zadaniem jest doświadczalne przewidzenie przyszłej politycznej sytuacji w Egipcie. Jakie partie polityczne dojdą tam do głosu? Jak Egipt będzie wyglądał za 5, 10 lub 20 lat? Masz dostęp tylko do dwóch następujących zbiorów danych dotyczących wszystkich Egipcjan: sieci facebookowej lub twitterowej, kompletnego zapisu, kto z kim chodził do szkoły, tekstów rozmów telefonicznych każdej osoby i jej adresu lub danych sieciowych dotyczących członków wszystkich formalnych organizacji politycznych i prywatnych przedsiębiorstw.

Nim podejmiesz decyzję, zważ, że rzeczy zmieniają się z biegiem czasu: ludzie mogą wynosić się z Facebooka, a rozmowy polityczne mogą wymagać zawoalowania, jeśli blogowanie odbywa się zbyt jawnie. Sam Facebook daje mnóstwo informacji, lecz czasami ludzie będą próbowali stać się niewidoczni — być może ci sami, którzy pozostają w sferze Twoich największych zainteresowań. Z tego powodu lepszą reprezentacją mogą być zapisy telefoniczne.

Jeśli myślisz, że ten scenariusz jest na wyrost, wiedz, że jest on już realizowany. Na przykład niemiecki Siemens sprzedał Iranowi oprogramowanie do monitorowania ich krajowej sieci telefonów komórkowych. W rzeczywistości rządy — mówiąc ogólnie — wkładają więcej energii w zapełnianie tego obszaru swoimi sprzymierzeńcami, a mniej w jego osłabienie: Pakistan wynajmuje Amerykanów, aby blogowali na rzecz Pakistanu, a Rosjanie pomagają Syryjczykom.

Ostatnia uwaga: musisz rozważyć zmianę typowego kierunku myślenia. Mnóstwo ludzi zadaje pytanie: czego możemy się dowiedzieć z tych czy innych danych? Pomyśl o tym inaczej: co by to znaczyło, móc przewidywać kierunki polityki w społeczeństwie? I jakie rodzaje danych są Ci potrzebne, aby tego dokonać?

---

6 Zob. *http://networkx.github.io*.

7 Zob. *http://igraph.org/redirect.html*.

8 Zob. *http://statnet.org*.

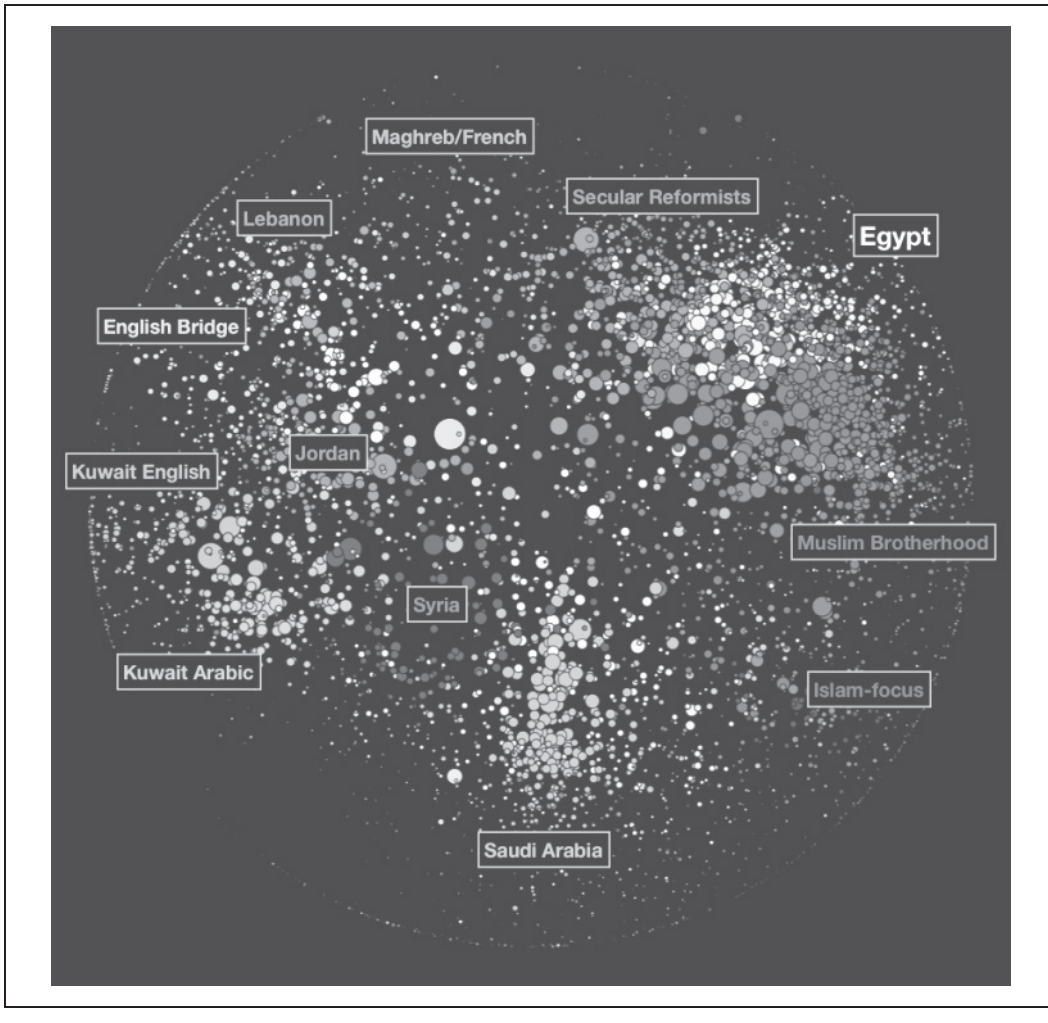
9 Zob. *http://research.microsoft.com/en-us/projects/nodexl*.

10 Zob. *http://cs.stanford.edu/people/jure*.

Innymi słowy, najpierw znajdź pytania, a potem szukaj danych, które pomogą na nie odpowiedzieć.

Metody analityczne w Morningside

Kelly pokazał nam mapę sieci 14 największych światowych blogosfer. Aby zrozumieć te rysunki, wyobraź sobie siłę, na przykład taką jak wiatr, która wypycha węzły ku brzegom, a jednocześnie drugą siłę przeciwstawną — odsyłacze (linki) między blogami — która je spaja. Na rysunku 10.1 pokazano przykład arabskiej blogosfery.



Rysunek 10.1. Przykład arabskiej blogosfery (patrz kolorowa wkładka)

Różne kolory reprezentują kraje i skupiska blogów. Rozmiar każdej kropki wyraża centralność mierzoną według stopnia, tj. liczbę odsyłaczy do innych blogów w sieci. Fizyczna struktura blogosfery może dać nam do myślenia.

Jeżeli analizujemy tekst, używając przetwarzania języka naturalnego (ang. *natural language processing* — NLP), to myślenie o blogowych postach jak o górze lub rzece tekstu powoduje, że dostrzegamy tylko obraz mikro lub makro — tracimy przekaz najważniejszy. Gubi się analiza sieci społecznych (ang. *social network analysis* — SNA), która jest pomocna w odwzorowywaniu i analizowaniu wzorców interakcji. Na przykład 12 różnych blogosfer międzynarodowych wygląda inaczej. Może to nas prowadzić do wniosku, że różne społeczeństwa mają różne zainteresowania, co powoduje odmienne wzorce.

Dlaczego jednak one się różnią? W końcu są reprezentacjami czegoś więcej wymiarowego rzutowanego na dwa wymiary. A może po prostu zostały różnie narysowane? Owszem, lecz możemy wykonać analizę mnóstwa tekstów, która przekonuje, że te obrazy rzeczywiście coś ukazują. Wkładamy wysiłek w jakościowe interpretowanie treści.

I tak na przykład w blogosferze francuskiej widzimy grono dyskutujące o smacznym gotowaniu. W Niemczech spotykamy różne grona dyskutujące o polityce i rozmaitych zwariowanych hobby. W blogach angielskich zauważamy dwa duże skupienia [Cathy/mathbabe wtrąca swoje trzy grosze: porno dla gejów i zwykłe porno?]. Okazuje się, że są to blogi konserwatystów i liberałów.

W Rosji sieci blogowania wykazują tendencję do wywierania nacisku na pozostawanie w sieciach, dlatego widzimy dobrze określone, porozdzielane skupienia.

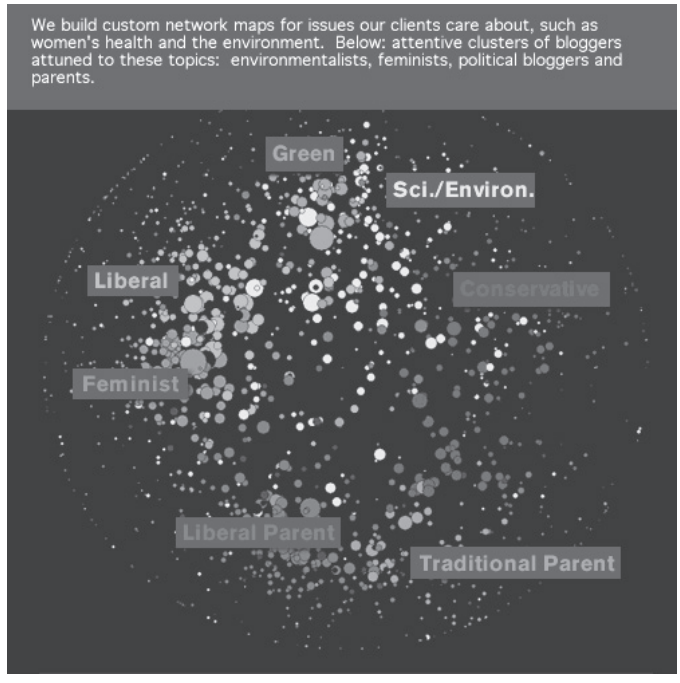
Grupowanie tego, co w pobliżu (ang. *proximity clustering*), jest wykonywane za pomocą algorytmu Fruchtermana-Reingolda, w którym przebywanie w tym samym sąsiedztwie oznacza, że Twoi sąsiedzi są podłączeni do innych sąsiadów, zatem odzwierciedla ono naprawdę zjawisko wywierania zbiorowego wpływu. Następnie interpretujemy segmenty. Na rysunku 10.2 przedstawiono przykład blogów w języku angielskim.

Jak wizualizacje pomagają znaleźć ławice ryb

Każda z firm zajmujących się mediami społecznościowymi czerpie z tego, że albo dysponuje danymi, albo zestawem narzędzi — opatentowanym mechanizmem [sondowania] opinii (ang. *sentiment engine*) lub czymś w tym rodzaju, *maszyną do robienia hałasu*. Bądźmy jednak świadomi, że media społecznościowe są w dużym stopniu wytworem organizacji zainteresowanych w nadawaniu biegu sprawom, czyli *graniem maszyną do robienia hałasu*. Aby uwierzyć w to, co widzisz, musisz trzymać rękę na pulsie, to znaczy musisz rozszyfrować zasady gry, zrozumieć, na czym ona polega. A to oznacza, że potrzebujesz wizualizacji.

Przykład. Jeśli przymierzasz się do wyborów, obejrzyj blogi z mamusiami lub miłośnikami sportu. To będzie bardziej komunikatywne niż blogi zwolenników, których odpowiedź już znasz.

Inny przykład. Po podzieleniu blogosfery na koszyki Kelly przedstawił nam analizę różnych typów odsyłaczy (linków): do wideonagrań zwolenników, jak na przykład z przemówieniem Martina Luthera Kinga „I have a dream" oraz profesjonalnego filmu z kampanii Romneya. W przypadku filmu z przemówieniem MLK widzimy jego żywiołowe rozsyłanie w okresie wyborczym po całej blogosferze, lecz w przypadku wideokampanii Romneya obserwujemy zgodne wysiłki konserwatywnych blogerów polegające na wysyłaniu wideozapisu na komendę.



Rysunek 10.2. Blogi w języku angielskim[11] *(patrz kolorowa wkładka)*

Należy przez to rozumieć, że gdyby oglądać tylko wykres odsyłaczy — samą ich liczbę — to mogłoby wyglądać, że wideozapis z Romneyem jest rozsyłany wirusowo, jeśli jednak spojrzymy na to z uwzględnieniem segmentacji blogosfery, to staje się jasne, że była to zaplanowana operacja mająca na celu granie na miarach „wirusowości"[12].

Kelly współpracuje również z harvardzkim Berkman Center for Internet and Society. W 2008 roku i ponownie w 2011 przeanalizował irańską blogosferę, znajdując wiele podobieństw pod względem skupień: młodzi antyrządowi demokraci, poezja (ważna część irańskiej kultury) i prorządowe grona konserwatywne dominowały w obu latach.

Jednakże do 2011 roku ostało się tylko 15% blogów z tych, które istniały w 2008 roku.

Tak więc, choć ludzie często skupiają się wobec *jednostek* (model przypadek-atrybut), pojedyncza ryba jest mniej ważna niż *ławice* ryb. Wykonując analizę sieci społecznościowych, poszukujemy ławic, ponieważ w ten sposób dowiadujemy się o tym, co szczególnie nurtuje społeczeństwo i na jakiej zasadzie dążenia te utrzymują się mimo upływu lat.

Morał z tego jest taki, że powinniśmy skupiać się na wypośrodkowanych wzorcach (mezzowzorcach), a nie na wzorcach z poziomu mikro czy makro.

---

[11] Tekst w nagłówku rysunku głosi: „Budujemy sieci odzwierciedlające zamówienia klientów, dotyczące na przykład zdrowia kobiet lub zagadnień środowiska. Poniżej uwidoczniono zasługujące na uwagę skupiska blogerów reprezentujących grupy zainteresowań: ekolodzy, środowiska feministyczne, blogerzy polityczni i rodzice" — *przyp. tłum.*

[12] Ang. *virality*; w znaczeniu informacji rozprzestrzeniającej się lotem błyskawicy — *przyp. tłum.*

Szersze tło statystyczne analizy sieci społecznych

Jednym z podejść do analizy sieci społecznych (ang. *social network analysis* — SNA) są rozważania dotyczące samej sieci traktowanej jako obiekt losowy, na podobieństwo liczby losowej lub zmiennej losowej. Sieć można rozpatrywać jako wynik procesu losowego, czyli wynikającą z odpowiedniego rozkładu prawdopodobieństwa. Jest do pomyślenia próba złożona z [wielu] sieci, w odniesieniu do której można by zadawać takie pytania jak: Co charakteryzuje sieci dające się przyrównać do Twittera? Czy dana sieć odzwierciedla przyjaźnie w rzeczywistym świecie? Co w ogóle oznaczałoby udzielenie na to pytanie odpowiedzi twierdzącej lub przeczącej?

To są jedne z głównych pytań dyscypliny określanej jako analiza sieci społecznych, która wyłoniła się z takich akademickich dziedzin jak matematyka, statystyka, informatyka, fizyka i socjologia, mającej rozległy zakres zastosowań w jeszcze liczniejszych dziedzinach, w tym w badaniach fMRI[13], epidemiologii i studiach nad sieciami społecznościowymi (sieciami społecznymi online), takimi jak Facebook lub Google+.

Reprezentacje sieci i centralność wartości własnej

W niektórych sieciach krawędzie między węzłami są skierowane: mogę postępować za Tobą na Twitterze, podczas gdy Ty za mną nie postępujesz, toteż krawędź będzie prowadzić *ode* mnie *do* Ciebie. Natomiast inne sieci mają tylko symetryczne krawędzie: albo się wzajemnie znamy, albo nie. Te ostatnie sieci są określane jako *nieskierowane*.

Sieć nieskierowaną o N węzłach można przedstawić w postaci macierzy $N \times N$ złożonej z jedynek i zer, w której element (i, j) jest równy 1 wtedy i tylko wtedy, gdy węzły i i j są połączone. Macierz tę nazywamy *macierzą sąsiedztwa* (ang. *adjacency matrix*) lub *macierzą incydencji* (ang. *incidence matrix*)[14]. Zauważmy, że możemy to w istocie zdefiniować także dla sieci skierowanych, lecz w przypadku sieci nieskierowanych macierz taka jest zawsze symetryczna.

Alternatywną reprezentacją sieci jest lista list: dla każdego węzła i wypisujemy wykaz węzłów, z którymi węzeł i ma połączenie. Nazywa się to listą *incydencji* i zauważmy, że nie zależy ona od tego, że sieć jest nieskierowana. Przedstawianie sieci w ten sposób oszczędza pamięć — węzły mogą mieć atrybuty reprezentowane w postaci wektora lub listy. Na przykład, jeśli węzły oznaczają ludzi, to atrybutami mogą być informacje demograficzne lub informacje dotyczące ich zachowań, obyczajów czy upodobań.

Krawędziom można również przypisywać wartości, czyli wagi (wektory), wyrażające informacje o charakterze powiązań węzłów, między którymi występują. Wartości te można zapamiętać w macierzy $N \times N$ zamiast jedynek i zer, które reprezentują tylko obecność lub nieobecność powiązania.

Korzystając z pojęcia macierzy sąsiedztwa A, możemy na koniec zdefiniować *centralność wartości własnej* (ang. *eigenvalue centrality*), o której wspomnieliśmy już w punkcie „Miary centralności". Definiuje się ją zwięźle jako jednoznaczne wektorowe rozwiązanie x równania

[13] Czyli w obrazowaniu metodą rezonansu magnetycznego; akronim pochodzi od *functional Magnetic Resonance Imaging* — *przyp. tłum.*

[14] Według innych definicji macierz incydencji odnosi się do grafów skierowanych — *przyp. tłum.*

$$Ax = \lambda x$$

takie, że

$$x_i > 0,\ i = 1 \cdots N$$

Jak się okazuje, ostatni warunek jest równoważny wyborowi największej wartości własnej λ. Zatem w rzeczywistym algorytmie należy znaleźć pierwiastki równania $det\left(A - tI\right)x$ i uporządkować je według rozmiaru, biorąc największy i określając go jak λ. Następnie znajdujemy x, rozwiązując układ równań:

$$\left(A - \lambda I\right)x = 0$$

W ten sposób otrzymujemy x, wektor wyników centralności wektora własnego.

Zauważmy, że nie mówi nam to *zbyt* wiele o centralności wartości własnej, mimo że stanowi sposób na jej obliczenie. Aby wyrobić sobie o niej głębsze pojęcie, należy rozważyć ją jako granicę prostego schematu iteracyjnego, choć wymagałoby to dowodu, który możesz znaleźć na przykład tutaj[15].

Otóż zaczynamy od wektora, którego elementy są po prostu stopniami węzłów[16], być może przeskalowanymi tak, aby suma elementów wynosiła 1. Same stopnie nie dają nam jednak prawdziwej wiedzy o sposobie połączenia danego węzła, toteż w następnym powtórzeniu dodajemy stopnie wszystkich sąsiadów danego węzła, znowu je skalując. Powtarzamy to postępowanie, dodając za każdym razem stopnie o krok dalszych sąsiadów. W granicy — jako że ten proces iteracyjny się nie kończy[17] — otrzymujemy wektor centralności wartości własnej.

Pierwszy przykład grafów losowych: model Erdősa–Rényiego

Przeróbmy prosty przykład, w którym sieć można rozpatrywać jako jednostkowy efekt procesu stochastycznego. Mianowicie taką, w której dany węzeł występuje zgodnie z rozkładem prawdopodobieństwa, a *wszystkie węzły są traktowane niezależnie*.

Sieć Bernoulliego

Nie wszystkie sieci o N węzłach są jednakowo prawdopodobne w tym modelu. Prawdopodobieństwo zaobserwowania sieci, w której wszystkie węzły mają połączenie ze wszystkimi innymi, wynosi p^D, natomiast prawdopodobieństwo zaobserwowania sieci, w której wszystkie węzły są rozłączone wynosi $\left(1-p\right)^D$. I oczywiście między tymi skrajnościami jest do pomyślenia wiele innych sieci. Model Erdősa-Rényiego jest również określany jako *sieć Bernoulliego*. W literaturze matematycznej, model Erdősa-Rényiego jest traktowany jako obiekt matematyczny o interesujących własnościach, umożliwiający dowodzenie twierdzeń.

[15] W pracy Leo Spizzirriego: *http://www.math.washington.edu/~morrow/336_11/papers/leo.pdf* — *przyp. tłum.*

[16] Tj. liczbami krawędzi sąsiadujących z poszczególnymi węzłami — *przyp. tłum.*

[17] Liczba węzłów w grafie jest skończona, ale mogą w nim występować pętle znajomości — *przyp. tłum.*

Powiedzmy, że zaczynamy od N węzłów. Istnieje więc $D = \binom{N}{2}$ par węzłów, czyli *diad*, które mogą być połączone krawędzią (nieskierowaną) albo nie. Wobec tego jest 2^D sieci możliwych do zaobserwowania. Najprostszy rozkład, według którego można ulokować poszczególne węzły, nosi nazwę *modelu Erdősa-Rényiego*. Zakłada się w nim, że krawędź między każdą parą węzłów (i, j) istnieje z prawdopodobieństwem p.

Drugi przykład grafów losowych: wykładniczy model grafu losowego

Teraz zła nowina: sieci społeczne możliwe do zaobserwowania w rzeczywistości nie przypominają sieci Bernoulliego. Na przykład sieci znajomych lub sieci współpracy akademickiej na ogół przejawiają takie cechy jak: *przechodniość* (tranzytywność, ang. *transitivity*) — tendencję, że jeśli A zna B i B zna C, to A zna C, skupianie (ang. *clustering*) — tendencję do gorzej lub lepiej zdefiniowanych niewielkich grup istniejących w ramach większych sieci, wzajemność, czyli obopólność (w sieci skierowanej jest to tendencja do śledzenia B przez A, jeśli A śledzi B), i wewnętrzność (położenie pośrodku, tendencja polegająca na istnieniu pewnych osób, przez które przepływają informacje).

Niektóre z tych właściwości obserwowanych w rzeczywistych sieciach są dość proste do przełożenia na język matematyki. Na przykład przechodniość można ująć za pomocą liczby trójkątów w sieci.

Wykładnicze modele grafów losowych (ang. *exponential random graph models* — ERGM-y) służą do ujmowania tych cech sieci rzeczywistego świata i są powszechnie stosowane w socjologii.

Ogólne podejście do ERGM-ów polega na wybraniu odpowiedniej *statystyki grafu*, jak liczba trójkątów, liczba krawędzi i liczba *dwugwiazd* (ang. *2-stars*, czyli podgrafów składających się z węzła z dwiema szprychami (ramionami); tak więc węzeł stopnia 3 ma skojarzone ze sobą trzy *dwugwiazdy*) przy danej liczba węzłów i to wszystko może być traktowane jako zmienne z_i Twojego modelu. Wtedy trzeba tak dobrać towarzyszące im współczynniki θ_i, aby były dostrojone do pewnego typu zachowania, które obserwujesz lub chcesz symulować. Jeśli na przykład z_1 odnosi się do liczby trójkątów, to dodatnia wartość θ_1 oznaczałaby tendencję do większych liczb trójkątów.

Dodatkowe statystyki grafu, które się wprowadza, obejmują k-gwiazdy (podgrafy składające się z węzła z k szprychami — zatem węzeł stopnia $k+1$ ma $k+1$ k-gwiazd), stopień, czyli *naprzemienne k-gwiazdy*, zbiorczą statystykę liczb k-gwiazd dla różnych k. Aby dać pojęcie o tym, jak ERGM może wyglądać od strony wzorów, prezentujemy jeden z nich:

$$Pr(Y = y) = \left(\frac{1}{\kappa}\right)\left(\left(\theta_1 z_1(y) + \theta_2 z_2(y) + \theta_3 z_3(y)\right)\right)$$

Tutaj stwierdzamy, że prawdopodobieństwo zaobserwowania konkretnej realizacji grafu losowego lub sieci Y jest funkcją statystyki grafu, czyli cech, które właśnie oznaczyliśmy jako z_i.

W tych ramach sieć Bernoulliego jest szczególnym przypadkiem ERGM-u, takim, w którym mamy tylko jedną zmienną, odpowiadającą liczbie krawędzi.

Wnioskowanie w ERGM-ach

W warunkach idealnych, choć w pewnych przypadkach w praktyce nierealistycznych, można zaobserwować próbę kilku sieci Y_1, ..., Y_n reprezentowanych przez macierze sąsiedztwa, powiedzmy — dla ustalonej liczby węzłów N.

Mając te sieci, moglibyśmy zamodelować je jako niezależne i jednakowo rozproszone obserwacje z tego samego modelu prawdopodobieństwa. Moglibyśmy wtedy wywnioskować parametry tego modelu.

Oto pierwszy przykład. Jeśli weźmiemy pod uwagę sieć Bernoulliego, charakteryzowaną przez prawdopodobieństwo p istnienia dowolnego węzła, to możemy obliczyć szansę dowolnej sieci z naszej próby, wynikającą z danej sieci Bernoulliego jako

$$L = \Pi_i^n p^{d_i} \left(1-p\right)^{D-d_i}$$

gdzie d_i jest liczbą obserwowanych krawędzi w i-tej sieci, a D — jak poprzednio — jest sumaryczną liczbą diad w sieci. Następnie możemy wyciągnąć estymator dotyczący p jako:

$$\hat{p} = \frac{\sum_{i=1}^{n} d_i}{nD}$$

W praktyce w literaturze wykładniczych modeli grafów losowych (ERGM-ów) jest obserwowana tylko jedna sieć, przez co rozumie się, *że pracujemy na próbie rozmiaru jeden*. Na podstawie tego jednego przykładu estymujemy parametr modelu prawdopodobieństwa, który „wygenerował" tę sieć. W przypadku sieci Bernoulliego nawet z jednej sieci moglibyśmy wyestymować (oszacować) p jako odsetek krawędzi w łącznej liczbie diad, co wydaje się oszacowaniem rozsądnym.

Jednak dla bardziej skomplikowanych ERGM-ów estymowanie parametrów na podstawie jednej obserwacji sieci jest trudne. Jeśli dokonuje się go za pomocą czegoś określanego jako *procedura estymacji pseudowiarygodności*, skutkuje to niekiedy produkowaniem wartości nieskończonych (zob. artykuł Marka Handcocka *Assessing Degeneracy of Statistical Models of Social Networks* z 2003 roku[18]). Jeśli natomiast jest ono robione za pomocą tzw. *metod MCMC*[19], to dolega mu coś, co jest określane jako *degeneracja dedukcyjna*, kiedy to algorytm zbiega się do zdegenerowanych grafów — takich, które są pełne albo puste — lub algorytm nie zbiega się konsekwentnie (to również znajdujemy w artykule Handcocka).

Dalsze przykłady grafów losowych: modele przestrzeni ukrytych, sieci małych światów

Badacze, zmotywowani problemami degeneracji modelu i niestabilności w wykładniczych modelach grafów losowych, wprowadzili *modele przestrzeni ukrytych* (ang. *latent space models*); zob. *Latent Space Approaches to Social Network Analysis* Petera Hoffa.

W modelach przestrzeni ukrytych dąży się do uchwycenia następującego zagadnienia: obserwujemy pewną rzeczywistość, lecz istnieje też pewna związana z nią rzeczywistość utajona, której nie możemy zaobserwować. Możemy na przykład obserwować powiązania między

[18] Streszczenie pod adresem *http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.81.5086 — przyp. tłum.*

[19] Akronim pochodzi od ang. *Markov Chain Monte Carlo*, czyli: łańcuchy Markowa i metody Monte Carlo — *przyp. tłum.*

ludźmi na Facebooku, lecz nie możemy ustalić, gdzie oni mieszkają, ani innych właściwości, które powodują, że mają skłonność do zaprzyjaźniania się ze sobą.

Inni badacze zaproponowali *sieci małych światów* (ang. *small-world networks*); por. modele zaproponowane w artykule Wattsa i Strogatza w 1998 roku. Modele te zajmują w spektrum miejsce pomiędzy grafami całkowicie losowymi i całkowicie regularnymi, próbuje się w nich uchwycić zjawisko rzeczywiste na sześciu oddzielnych poziomach. W krytyce tych modeli podkreśla się, że produkują one sieci homogeniczne pod względem stopnia, natomiast sieci obserwowalne w rzeczywistym świecie wykazują tendencje do swobodnej skalowalności i niehomogeniczności, jeśli chodzi o stopień.

Oprócz opisanych modeli inne ich klasy obejmują losowe pola Markowa, modele bloków stochastycznych, modele mieszanej przynależności i modele mieszanej przynależności bloków stochastycznych — w każdym z nich dane relacyjne są modelowane w różny sposób i dąży się do włączania cech nieobecnych w innych modelach. (Zob. na przykład artykuł *Mixed Membership Stochastic Block Models* Eduardo Airoliego i innnych).

Oto kilka podręczników do dalszych lektur:

- *Networks, Crowds, and Markets* (Cambridge University Press) autorstwa Davida Easleya i Jona Kleinberga z Instytutu Informatyki Uniwersytetu Cornella.
- Rozdział o wydobywaniu grafów sieci społecznych w książce *Mining Massive Datasets* (Cambridge University Press) Ananda Rajaramana, Jeffa Ullmana i Jure'a Leskoveca z Instytutu Informatyki Uniwersytetu Stanforda.
- *Statistical Analysis of Network Data* (Springer) Erica D. Kolaczyka z Boston University.

Dziennikarstwo danych

Naszym drugim mówcą wieczoru był Jon Bruner, redaktor z O'Reilly, który przedtem pracował jako redaktor danych w „Forbesie". Ma szerokie umiejętności: prowadzi rozległe badania danych i pisze o wszystkim, co się z nimi wiąże.

Kilka słów o historii dziennikarstwa danych

Dziennikarstwo danych istnieje od pewnego czasu, lecz do niedawna doniesienia wspierane komputerowo były domeną użytkowników mocnych w Excelu. (Nawet obecnie, jeśli umiesz napisać program w Excelu, jesteś elitą).

Ostatnio coś się w tym zmienia. Coraz więcej danych jest nam udostępnianych za pośrednictwem API, nowych narzędzi i przy mniejszym zużyciu mocy obliczeniowej, toteż prawie każdy może analizować całkiem duże zbiory danych na laptopie. Umiejętności programowania są obecnie szeroko upowszechnione, możesz więc znaleźć ludzi dobrych zarówno w piórze, jak i w programowaniu. Wiele osób biegłych w angielskim wie na tyle o komputerach, aby dać sobie z nimi radę; z drugiej strony można znaleźć znawców informatyki potrafiących pisać.

W dużych pismach, jak „New York Times", dziennikarstwo danych jest uprawiane z podziałem na obszary: grafika a właściwości interakcyjne, badania, inżynierowie baz danych, roboty internetowe (ang. *crawlers*), budowniczowie oprogramowania i autorzy piszący na tematy specjalistyczne. Jedni odpowiadają za stawianie właściwych pytań, które jednak przekazują innym do analizy. Na przykład Charles Duhigg z „New York Timesa" badał jakość wody w Nowym

Jorku, w związku z czym złożył wniosek o swobodny dostęp do informacji[20] stanu Nowy Jork — wiedział wystarczająco dużo, by zdawać sobie sprawę, co winno być we wniosku FOIA i jakie zadawać pytania, lecz faktyczną analizę wykonał kto inny.

W mniejszych strukturach sprawy mają się zupełnie inaczej. Podczas gdy „New York Times" ma do dyspozycji na „piętrze" redakcyjnym 1000 osób, „The Economist" ma coś ze 130, a „Forbes" na w swoich pokojach redakcyjnych — 70 lub 80 ludzi. Jeśli pracujesz gdzie indziej niż w krajowej gazecie codziennej, to zajmujesz się wszystkim: zadajesz pytania, na własną rękę gromadzisz dane, dokonujesz analiz i piszesz o tym. (Oczywiście w miarę możności możesz też korzystać z pomocy i współpracy swoich koleżanek i kolegów).

Uprawianie dziennikarstwa technicznego — rady eksperta

Jon był głównym matematykiem w college'u na Uniwersytecie w Chicago, potem podjął pracę w „Forbesie", gdzie powoli zwrócił się ponownie do zagadnień ilościowych. Zajął się na przykład zastosowaniem teoretycznych narzędzi grafowych w badaniu subsydiowania polityków przez miliarderów.

Wyjaśnił słuchaczom termin „dziennikarstwo danych" (ang. *data journalism*), przedstawiając własny profil naukowy.

Przede wszystkim obejmuje ono *dużo* wizualizacji danych, ponieważ jest to szybka droga do opisywania najważniejszych wniosków ze zbioru danych. Znajomość informatyki również ma dość duże znaczenie w dziennikarstwie danych. Istnieją wąskie, nieprzekraczalne terminy i dziennikarze zajmujący się danymi muszą dobrze znać swoje narzędzia oraz radzić sobie z niechlujnymi danymi, jako że nawet dane federalne są zanieczyszczone. Trzeba znać arkana przeróżnych formatów, co często oznacza konieczność rozbioru materiału w Pythonie. Sam Jon używa między innymi JavaScriptu, Pythona, SQL-a i języka MongoDB.

Statystyka, jak mówi Bruno, uczy sposobu myślenia o świecie. Inspiruje Cię do opisywania rzeczy, na przykład *przeciętna osoba* na Twitterze jest kobietą i ma z 250 fanów, lecz *osoba wynikająca z mediany* ma 0 wyznawców — dane są wyraźnie tendencyjne. To jest w sam raz temat na artykuł.

Bruno przyznaje, że jest nowicjuszem w dziedzinie uczenia maszynowego. Zaznacza jednak, że jest specjalistą od krytycznego podejścia do dziennikarstwa danych: z wyjątkiem ludzi, którzy mogą specjalizować się w jednym temacie, powiedzmy, w biurze rządowym lub wielkonakładowym dzienniku, w mniejszej gazecie musisz być wszechstronny i musisz szybko nabywać sporo specjalistycznych umiejętności.

Oczywiście komunikacja i prezentacje mają dla dziennikarzy bez wątpienia znaczenie zasadnicze. Ich podstawową umiejętnością jest *translacja* — zdolność wyciągania ze skomplikowanych historii znaczenia zrozumiałego dla czytelników. Muszą również uprzedzać pytania, obracać je w eksperymenty ilościowe i przekonująco na nie odpowiadać.

Oto jakiej rady udziela Jon wszystkim wchodzącym na dziennikarską ścieżkę: *nie upieraj się przy swoim zdaniu, zanim nie porozmawiasz ze znawcą tematu*. Zacznij od luźnego wyobrażenia tego, czego poszukujesz, i miej gotowość zmiany swojego myślenia i stanowiska, jeśli eksperci zaprowadzą Cię w nowym i ciekawym kierunku. Brzmi trochę jak eksploracyjna analiza danych!

[20] Na mocy *Freedom of Information Act*, amerykańskiej regulacji prawnej o uzyskiwaniu dostępu do informacji; por. *http://foia.state.gov/Request/Submit.aspx — przyp. tłum.*

Skorowidz

N

O

P

R

O'REILLY®

Badanie danych

Raport z pierwszej linii działań

W dzisiejszych czasach najcenniejszym dobrem jest informacja. Ogromne ilości danych są przechowywane w przepastnych bazach danych, a kluczem do sukcesu jest ich umiejętna analiza i wyciąganie wniosków. To dynamicznie rozwijająca się dziedzina wiedzy, w której do tej pory brakowało solidnych podręczników, pozwalających na dogłębne poznanie tego obszaru. Na szczęście to się zmieniło!

Trzymasz właśnie w rękach unikalną książkę, w której badacze z największych firm branży IT dzielą się skutecznymi technikami analizy danych. Z kolejnych rozdziałów dowiesz się, czym jest nauka o danych, model danych oraz test A/B. Ponadto zdobędziesz wiedzę na temat wnioskowania statystycznego, algorytmów, języka R oraz wizualizacji danych. Sięgnij po tę książkę, jeżeli chcesz się dowiedzieć, jak wykrywać oszustwa, korzystać z MapReduce oraz badać przyczynowość. To obowiązkowa pozycja na półce czytelników zainteresowanych badaniem danych.

Dzięki tej książce:

- poznasz zaawansowane sposoby analizy danych
- nauczysz się korzystać z MapReduce
- zwizualizujesz posiadane dane i wykryjesz oszustwa
- poznasz podstawy języka R
- wyciągniesz wartościowe wnioski z posiadanych danych

Wyciągnij wartościowe wnioski z posiadanych informacji!

helion.pl
księgarnia internetowa

Nr katalogowy: 25943

Księgarnia internetowa:
http://helion.pl

Zamówienia telefoniczne:
0 801 339900
0 601 339900

Helion

Sprawdź najnowsze promocje:
http://helion.pl/promocje
Książki najchętniej czytane:
http://helion.pl/bestsellery
Zamów informacje o nowościach:
http://helion.pl/nowosci

Helion SA
ul. Kościuszki 1c, 44-100 Gliwice
tel.: 32 230 98 63
e-mail: helion@helion.pl
http://helion.pl

Informatyka w najlepszym wydaniu

sięgnij po WIĘCEJ
KOD KORZYŚCI

ISBN 978-83-246-9626-0

Cena 54,90 zł

9 788324 696260