Włodzimierz Gogołek
Instytut Dziennikarstwa, Uniwersytet Warszawski

Słowa kluczowe: informacja, Internet, Big Data, rafinacja informacji, wybory parlamentarne

Key words: information, Internet, Big Data, refining, parliamentary elections

INFORMACYJNY POTENCJAŁ RAFINACJI ZASOBÓW SIECIOWYCH

Wstęp

Na podobieństwo Wielkiego Wybuchu (*Big Bang*) – początku Wszechświata, Big Data stanowią nowy, eksplodujący wymiar informacyjnej przestrzeni świata. Już w 2011 r. zarejestrowano niemal dwa zeta bajty nowych, podatnych na wszechstronną analizę informacji[1]. Ich ekonomiczna i społeczna wartość wynika nie tylko z wielkości, lecz także z nieodkrytej jeszcze jakości. Nowe narzędzia do analizy Big Data, na wzór teleskopu Hawla, pozwalają dostrzec ogromny potencjał nowych zasobów informacyjnych, poznanie informacyjnego wymiaru świata XXI w. Nie jest on jednak neutralny, szczególnie wobec dysproporcji i celów możliwości jego wykorzystania[2]. W dalszej części opracowania pominąłem zagrożenia wynikające ze specyficznych zastosowań wspomnianego wymiaru świata informacji.

Bogactwem Big Data, wobec tradycji, poza zgromadzonymi informacjami o dowolnym przedmiocie jest ich wielowymiarowy kontekst. Jego umiejętna analiza – omówiona dalej rafinacja informacji – wydobywa nowe, trafne, precyzyjnie opisujące podmiot dane. Staje się nadzwyczaj wartościowym, nowym źródłem informacji dla wszystkich branż od biznesu, przez kulturę, edukację i naukę do świata mediów włącznie.

---

[1] Wolfgang Martin Team, *BI meets BPM and Big Data / Wolfgang Martin*, March 2013, http://www.wolfgang-martin-team.net/BI-BPM-SOA.php [dostęp: maj 2013].

[2] *Unlocking the Value of Personal Data: From Collection to Usage*, World Economic Forum, February 2013, http://www3.weforum.org/docs/WEF_IT_UnlockingValuePersonalData_CollectionUsage_Report_2013.pdf [dostęp: maj 2013].

Wyniki uzyskane z analizy Big Data tworzą wcześniej niedostępne źródła danych. Ich kreowanie może być postrzegane jako nowa faza rozwoju aplikacji IT (narzędzi i cyfrowych sieci wymiany informacji)[3]. Współcześnie obejmują one narzędzia mobilne, cloud computing, Big Data ze szczególnym wyróżnieniem zasobów sieci społecznościowych.

Umiejętna analiza Big Data pozwala na precyzyjniejsze i w stosownym czasie, także w czasie rzeczywistym, wyszukiwanie potrzebnych, krytycznych, a nawet wiarygodnie prognozujących informacji[4]. Umożliwią one doskonalenie i rozwijanie nowych generacji produktów i usług wykorzystywanych przez media.

Znaczącą część Big Data tworzą zasoby Internetu, a szczególnie sieci społecznościowe. Dane tego typu pochodzą od indywidualnych użytkowników sieci społecznościowych i o nich (blogi, fora, portale, maile lub strumień zapytań kierowanych do Internetu), przez profesjonalne publikacje i inne bogate zasoby informacyjne[5].

Przyjęto, że zasoby zgromadzone w Big Data tworzą informacje źródłowe, a wynik ich analizy to informacje wtórne. Proces owej analizy określany jest jako wspomniana rafinacja informacji sieciowych (rafinacja). Ważną jej cechą jest możliwość wyboru okresu rafinacji – od lat do ostatnich godzin, np. dzięki analizie bieżących wpisów na Twitterze, forach lub blogach.

Rafinacja

Idea rafinacji, która obejmuje specjalistyczną analizę nieustrukturyzowanych (głównie tekstowych) danych, ma swoje źródło w Data Mining. Jest to proces identyfikujący lub wyszukujący informacje z dużych ustrukturyzowanych, głównie ilościowych (nie tekstowych) zbiorów danych. Najczęściej wykorzystywany do odnajdywania prawidłowości w określonych procesach, do odkrywania nieznanych struktur i korelacji w celu formułowania hipotez. Często Data Mining stanowi integralną część powszechnie stosowanych CRMów[6]. Obecnie z powodzeniem używa-

[3] *Big data: The Next Frontier for Innovation, Competition, and Productivity*, http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation, [dostęp: maj 2011].

[4] http://readwrite.com/2012/02/08/harvard_researcher_uses_social_media_to_predict_st [dostęp: marzec 2013].

[5] S. Stephens-Davidowitz, *Google's Crystal Ball*, „The New York Times", http://campaignstops.blogs.nytimes.com/2012/10/20/googles-crystal-ball/?_php=true&_type=blogs&_r=0, October 20, 2012 [dostęp: kwiecień 2013].

[6] CRM (Customer Relationship Management) – specjalistyczne oprogramowanie do budowania profili klientów, zarządzania kampaniami, spersonalizowanej troski o klienta, oceny lojalności klienta oraz analizy sprzedaży i zaniedbań. CRM w reklamie społecznościowej – kombinacja znajomości oczekiwań konsumentów, rzeczywistości, a także ponoszonych wysiłków na istniejące media online.

ny w kontroli produktów, zarządzania ryzykiem, wykrywania oszustw itp. Podobnie jak rafinacja jest źródłem wcześniej niedostępnych informacji.

Jednym z ugruntowanych już filarów rafinacji jest Culturomics. Obejmuje on aktywności związane z eksploracją kulturowych trendów poprzez analizę bogatych zbiorów umożliwiających spojrzenie na funkcjonowanie społeczeństw. Korzystanie z narzędzi Culturomicsu sprawnie sygnalizują ważne kulturalne, naukowe i historyczne zmiany[7]. Mając na uwadze szersze spektrum informacji – Big Data – proces uzyskiwania nowych informacji, głównie z WWW, nazwano rafinacją informacji. Umożliwia ona dostrzeganie w obszarze informacji podstawowych – informacje wtórne, które są ukryte w zasobach WWW. Rafinacja jest jak mikroskop pozwalający zainteresowanym oglądać i mierzyć rzeczy – zarówno na poziomie poszczególnych komórek (rekordów), jak i grup społecznych. Jest to rodzaj rewolucji w pomiarach. Uzyskane w ten sposób dane tworzą nie tylko obraz potrzeb i zachowań indywidualnych użytkowników, lecz także społeczności jako całości.

Rafinacja pozwala uzyskać liczby opisujące wartości wybranych wymiarów przestrzeni badanego przedmiotu rafinacji. Na przykład liczby pozytywnych (jeden wymiar) i negatywnych (drugi wymiar) opinii na temat wybranej osoby, zdarzenia, firmy (przedmiot rafinacji). Wartości wymiarów umożliwiają monitorowanie zmiany przedmiotu rafinacji, określanie jego bieżącego stanu oraz szacowanie jego przyszłych zmian. Przykładem zastosowania uzyskanych w ten sposób informacji są alerty – automatyczne powiadamianie użytkownika o ważnych dla niego (przekroczenie progu zadanej wartości wymiaru/ów) zmianach stanu przedmiotu rafinacji. W odniesieniu do mediów chodzi tu także o automatyczne monitorowanie sieci w poszukiwaniu – ilościowo ocenianych – najpopularniejszych informacji.

Rafinacja obejmuje trzy zasady pozyskiwania informacji z Big Data:

– pierwsza – ilościowa ocena przedmiotu badań, np. liczby występowania określonych słów;
– druga – kontekstowa ocena przedmiotu badań, np. liczby występowania słowa „edukacja" w sąsiedztwie słów: „powszechna", „podstawowa", „krajowa", „zachodnia" itp.;
– trzecia – ocena sentymentów związanych z przedmiotem badań, np. liczby występowania słowa „edukacja" w sąsiedztwie słów: „dobra", „zła" itp.

[7] J.-B. Michel i in., *Quantitative Analysis of Culture Using Millions of Digitized Books*, „Science" 2011, Vol. 331, No. 6014, s. 176–182, za: http://www.sciencemag.org/content/331/6014/176 [dostęp: kwiecień 2013].

Narzędzia rafinacji

Zastosowanie odpowiednich narzędzi do rafinowania milionów postów, blogów i artykułów dostępnych online, pozwala na uzyskanie wcześniej niezauważanych informacji dotyczących społecznych fenomenów, państw, organizacji i osób indywidualnych. W wyniku rafinacji można także otrzymywać wartościowe informacje dotyczące oceny emocjonalnych relacji, takich jak: sympatia, uraza/rozgoryczenie, poczucie szczęścia, optymizm, pesymizm, obawa, niepokój. Proces ten określany jest mianem „analiza sentymentów".

Wyszukiwarki są względnie prostym (liczba funkcji), najczęściej wykorzystywanym narzędziem do rafinacji informacji. Paradoksalnie wyniki wyszukiwania są rezultatem kreacji specyficznego, zawężonego obrazu pierwotnych informacji dostępnych w sieci. Ta specyfika bazuje na szczególnym, zamierzonym wyborze odpowiedzi na zadawane pytania. Poszukujący informacji otrzymuje odpowiedzi, które są zgodne z zadanymi kryteriami/pytaniami i partykularnym celem konstruktorów wyszukiwarki. Innymi słowy wynik wyszukiwania jest rezultatem specyficznej funkcji oczekiwań użytkownika i jednocześnie spełnienia celów (np. komercyjnych, politycznych) właściciela wyszukiwarki[8].

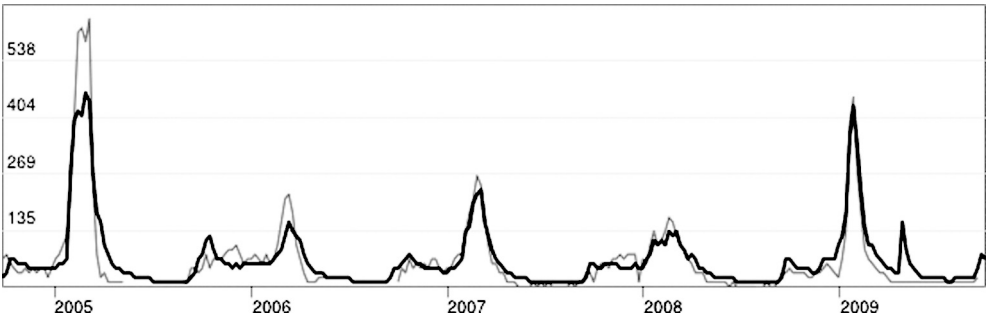
Dane behawioralne są względnie nową kategorią informacji pozyskiwanych z sieci; specyficzną formą rafinacji, której selektywność jest pochodną kryteriów opartych na zachowaniach użytkowników sieci. Uzyskane dzięki nim informacje są wynikiem potencjału znaczeniowego danych uzyskanych z analizy sposobów przeszukiwania i korzystania z sieci przez poszczególne osoby. Przykładem tego jest gromadzenie i analizowanie danych/słów/fraz używanych przez internautów podczas ich poszukiwań, np. realizowanych za pomocą wyszukiwarki Google. Dzięki temu przeprowadzane są analizy zachowań konsumentów w czasie rzeczywistym. Wynikiem tego było np. uzyskanie wartościowych i wiarygodnych informacji o ryzyku nadchodzącej epidemii grypy[9], co stanowiło efekt częstszych zapytań osób, które szukały w Internecie terminów powiązanych z ich poczuciem choroby (rys. 1 i 2). To nowe źródło informacji może mieć z powodzeniem różne zastosowania – także jako zasób informacji dziennikarskich.

Personalizacja jest innym – bezpośrednio powiązanym z rafinacją – narzędziem. Można wyróżnić dwie przestrzenie związane z personalizacją. Pierwsza – wielowymiarowa – opisuje różne typy użytkowników (wymiary/metadane: płeć, wiek, sprawność fizyczna, zainteresowania kulturalne, sportowe itp.). Druga przestrzeń – informacje pozyskiwane z sieci przez konkretną osobę są odpowiednio wybierane stosownie do konkretnych wartości wymiarów opisujących użytkownika. Dzięki

---

[8] E. Pfanner, *Google, in Settlement, Changes Ad Rules in France*, „The New York Times" 2010, October 28; D. Deacon, *Yesterday's Paper and Today's Technology, Digital Newspapers Archives and "Push Button" Content Analysis*, „European Journal of Communication" 2007, Vol. 22, No. 1, s. 17.

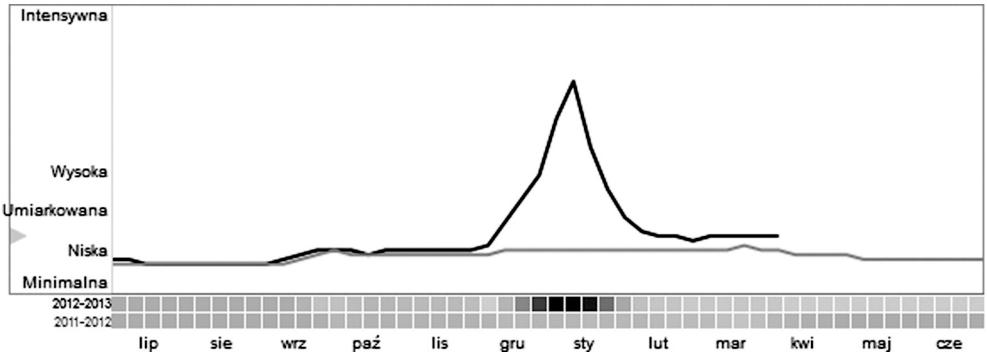
[9] D. Butler, *Web Data Predict Flu*, „Nature" 2008, No. 456, s. 287–288, http://www.nature.com/news/2008/081119/full/456287a.html [dostęp: luty 2012].



● Google Flu Trends – dane szacunkowe ● Polska – dane o zespole objawów grypopodobnych opublikowane przez Europejską Sieć Nadzoru nad Grypą Europejskiego Centrum ds. Zapobiegania i Kontroli Chorób.

Rys. 1. Aktywność grypy w Polsce. Szacowane zachorowania na grypę

Źródło: http://www.google.org/flutrends/about/how.html [dostęp: kwiecień 2013].



(● 2012–2013, ● 2011–2012)

Rys. 2. Tendencje aktywności wirusa grypy w Polsce

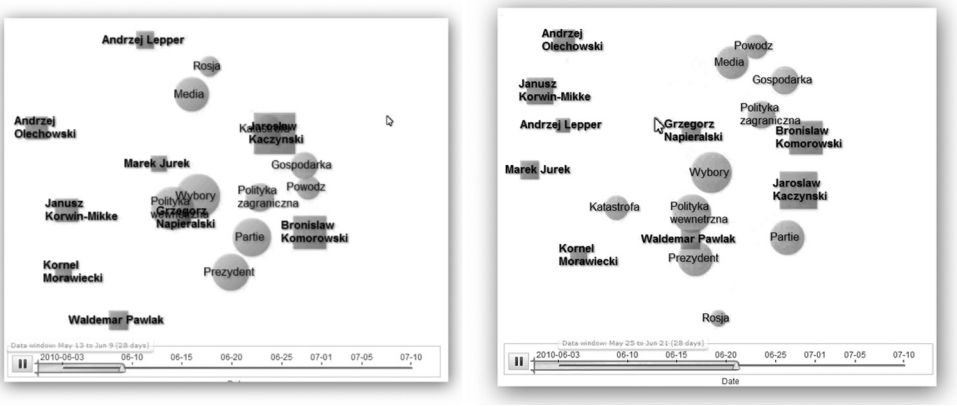
Źródło: http://www.google.org/flutrends/pl/#PL [dostęp: kwiecień 2013].

temu uzyskuje się dla użytkownika o określonych wartościach parametrów zawężony/spersonalizowany zbiór wyłuskany (rafinacja) z Big Data. W efekcie personalizacja spełnia nie tylko merytoryczne oczekiwania wirtualnego konsumenta, lecz także jego indywidualne cechy.

W rosnącej liczbie przypadków wirtualny profil konkretnego konsumenta, np. przyjętej sylwetki konsumenta materiału medialnego, jest dostępny wszędzie i zawsze. Bez względu gdzie i za pomocą jakiego urządzenia (PC, laptop, tablet, komórka) odbiorca ma zawsze indywidualne okno dla transferu informacji dopasowanych do jego osobistego profilu. Ilustrują to bogate doświadczenia personalizowanej reklamy online, które łatwo mogą być adaptowane do potrzeb mediów, głównie dostępnych za pośrednictwem sieci, także przez urządzenia mobilne.

Zaawansowane narzędzia. Do rafinacji zasobów sieciowych mogą być bezpośrednio użyte takie narzędzia, jak np.: Attentio, Radian6, Sysomos, NetBase, Collective Intellect, Alterian, Google Alerts. Rafinacja sieciowa jest skutecznie realizowana poprzez wykorzystanie Attentio Brand Dashboard[10]. Dowodzą tego wyniki badań dynamiki zmian obrazu informacyjnego kandydatów w wyborach prezydenckich 2010 r. (wielkość i położenie figur związanych z nazwiskami kandydatów rys. 3)[11]. Innym profesjonalnym narzędziem rafinacji jest serwis monitorujący serwisy informacyjne – *Summary of World Broadcasts* (SWB). Umożliwia on monitorowanie pełnych tekstów i streszczeń artykułów prasowych, materiałów pokonferencyjnych, materiałów telewizyjnych i radiowych, periodyków i innych nieklasyfikowanych technicznych raportów w 130 językach[12].

Wizualizacja jest bardzo użyteczną składową rafinacji – ważnym ogniwem łańcucha łączącego źródła informacji ze skutecznym wykorzystaniem rezultatów rafinacji. Problemem wizualizacji jest uzyskiwanie większej ilości informacji, ale eliminowanie jej mniej ważnych części. Wagę wizualizacji potwierdza bogactwo multimediów w mediach – włączają wizualizację statycznych i ruchomych obrazów. Reprezentatywnym przykładem takich narzędzi do rafinacji i jednocześnie ich wizualizacji są te, które wykorzystuje się do badań środowiska naturalnego – The Carbon Capture Report[13].



Rys. 3. Prezydenckie wybory w Polsce 2010 – wyniki analizy blogów

Źródło: http://www.youtube.com/watch?v=v0k0DWbddX8 [dostęp: kwiecień 2013].

[10] Obszerny spis narzędzi do analizy zasobów informacyjnych dostępny jest w opracowaniu: Wolfgang Martin Team, *BI meets BPM and Big Data*.

[11] P. Kuczma, W. Gogołek, *Informacyjny potencjał sieci – na przykładzie wyborów prezydenckich 2010*, „Studia Medioznawcze" 2010, nr 4(43), s. 35–49.

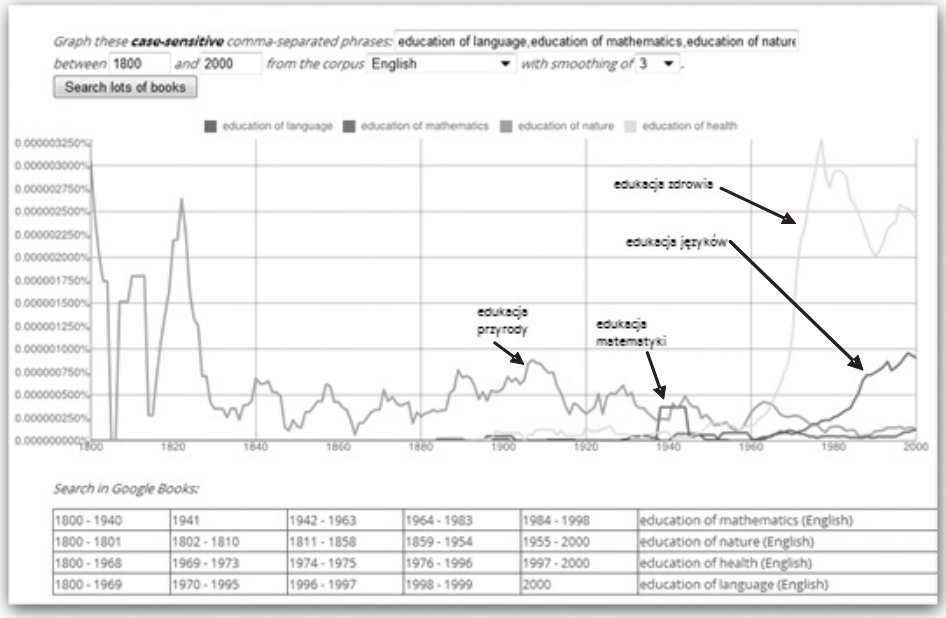
[12] „First Monday" 2011, Vol. 16, No. 9–5, September, http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3663/3040 [dostęp: styczeń 2012].

[13] *The Carbon Capture Report*, http://www.carboncapturereport.org/ [dostęp: wiosna 2012].

Wyniki rafinacji

Rafinacja stwarza możliwości wykrywania na zadanym poziomie ufności obrazu przeszłego i aktualnego statusu informacyjnego rzeczywistości, a nawet prognozowania przyszłość. Na przykład, odnośnie do przeszłości i współczesności, korpus ponad 5 mln książek w formie cyfrowej umożliwia na ilościową ocenę kulturalnych trendów, używając kolektywną pamięć opublikowanych książek, rozpoznawać adopcję technologii, cenzurę czy historię epidemiologii lub zmiany obecności słów związanych z edukacją (rys. 4)[14].

Innym przykładem nowych danych uzyskanych dzięki rafinacji jest łączenie tradycyjnych źródeł informacji z relatywnie nowymi – crowdsourcing. W sumie z zasobami tradycyjnymi pozwala to uzyskać rzetelniejszy obraz o świecie, np. w zakresie doskonalenia procedur doboru aktualnej tematyki publikacji medialnych (rys. 1 i 2)[15].



Rys. 4. Zmiany obecności zwrotów: edukacja matematyki, przyrody, zdrowia i języków, w latach 1800–2000

Źródło: Wynik usługi http://books.google.com/ngrams.

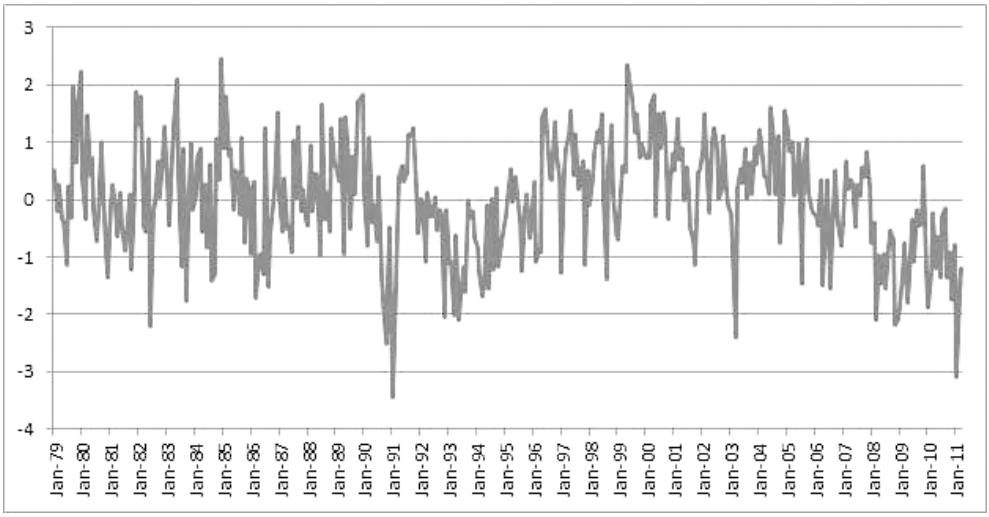
[14] M. Shen i in., *Quantitative Analysis of Culture Using Millions of Digitized Books*, „Science" 2011, Vol. 331, Issue 6014, s. 176–182.

[15] C. Anstey, *Empowering Citizen Cartographers*, „The New York Times" 2012, Jan. 13.

Przykłady

Inną wymowną ilustracją informacyjnej siły rafinacji sieci są rezultaty badań, które przeprowadził Kalev H. Leetaru. Dokonał on analizy sentymentów i geograficznych wymiarów 30-letnich archiwów światowych wiadomości do konstrukcji przewidywania w rzeczywistym czasie ludzkich zachowań, takich jak narodowe konflikty i precyzyjne daty specyficznych zdarzeń[16].

Podobna problematyka dotyczyła badań sentymentów w skali państwa – zakresu obejmowanych w analizie informacji (rafinacja) Egiptu, Tunezji i Libii w kontekście ostatnich politycznych zmian. Rysunek 5 ilustruje przeciętną miarę sentymentów treści w okresie od stycznia 1979 do marca 2011 r., 52 438 artykułów uzyskanych z SWB, które dotyczyły Egiptu. Rysunek ukazuje zmiany wartości sentymentów artykułów – pozytywne i negatywne. Odzwierciedlają one daty najbardziej istotnych zmian w tym kraju (styczeń 1991, marzec 2003, 1–24 stycznia 2011). Podobne rezultaty uzyskano odnośnie do dat istotnych zmian w Tunezji i Libii.



Rys. 5. Zmiany wartości miar sentymentów publikacji w Egipcie

Źródło: „First Monday" 2011, Vol. 16, No. 9–5, September.

Ocena barwy tonu informacji dotyczących części świata wyróżnia zapalne regiony kontynentów i innych informacji dotyczących indywidualnych postaci. Na przykład odnośnie do lokalizacji pobytu Bin Ladena. Rysunek 6 pokazuje wszystkie geograficzne wskazówki dotyczące Bin Ladena zawarte w treściach SWB od

[16] „First Monday" 2011, Vol. 16, No. 9–5, September.

Rys. 6. Wynik analizy (SWB) geolokalizacyjnych treści w światowych publikacjach od stycznia 1979 do kwietnia 2011 zawierających zwrot „Bin Laden"

Źródło: „First Monday" 2011, Vol. 16, No. 9–5, September, za: http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3663/3040 [dostęp: styczeń 2012].

stycznia 1979 do kwietnia 2011 r. Niemal połowa materiałów o Bin Ladenie była związana z Pakistanem[17].

Ilustracja rafinacji – wybory parlamentarne

Mając na uwadze zasygnalizowany potencjał Big Data, w Instytucie Dziennikarstwa Uniwersytetu Warszawskiego podjęto badania, których celem była ilustracja praktycznych zastosowań **potencjału rafinacji sieciowej, prezentacja i weryfikacja metodologii obróbki informacji na przykładzie tworzenia nowego źródła informacji zawierającego oceny bieżących preferencji wyborczych przed wyborami parlamentarnymi w Polsce w 2011 r.** Podstawą do osiągnięcia tego celu była ocena danych ilościowych i jakościowych oraz dynamika zmian treści ukazujących się w mediach społecznościowych oraz w sieciowych wydaniach niektórych drukowanych gazet.

Osiągnięcie założonego celu pozwoliło wskazać sposób kreowania miarodajnego źródła danych. W omawianych badaniach nad tego typu źródłem chodziło o wspomaganie diagnozowania stanu i dynamiki zmian obrazu informacyjnego o komitetach wyborczych (partiach politycznych) biorących udział w wyborach. Wiedza ta może stanowić wartościowe źródło informacji o przebiegu kampanii wyborczej dla mediów, zainteresowanych osób i grup społecznych.

[17] *Ibidem*.

Podobne badanie zostało przeprowadzone w 2010 r. przy okazji wyborów prezydenckich[18]. Jego wyniki w pełni potwierdziły zasadność kontynuowania ścieżki badawczej opartej na rafinacji informacji sieciowych.

Przyjęto następującą hipotezę: rafinacja sieci umożliwiała bieżący, wiarygodny monitoring zmiennych preferencji wyborczych Polaków w okresie poprzedzającym wybory parlamentarne w 2011 r. Innymi słowy treści w sieci, szczególnie w mediach społecznościowych, są odzwierciedleniem rzeczywistych postaw użytkowników i mogą zapowiadać ich realne działania, takie jak oddanie głosu na kandydata, partię, wybór określonej odpowiedzi w referendum. Osiągnięcie tak sformułowanego celu uzyskano dzięki wskaźnikom ilości treści związanych z partiami, oceny dynamiki zmian i oceny jakościowej.

Wskaźnik ilości treści został opracowany na podstawie liczby wszystkich wpisów/informacji w plikach zebranych przez Attentio Brands Dashboard, pochodzących ze źródeł online dotyczących danej partii oraz kontekstów. Wpisy zawierały treści uzyskane z forów, blogów, Facebooka, tweetów i artykułów gazetowych.

Ocena dynamiki zmian i trendów dotyczących treści/wyszukiwania została dokonana na podstawie liczb wpisów dotyczących partii zależnie od kontekstów i sentymentów w analizowanym czasie.

Ocenę jakościową przeprowadzono w dwóch kategoriach:

a) Wstępna analiza kontekstowa polegała na pogrupowaniu pozyskanych w wyniku monitoringu treści w konteksty na podstawie listy kontekstów merytorycznych i medialnych omówionych dalej.
b) Równolegle z analizą kontekstową przeprowadzono wstępną analizę sentymentów. Rozumiana jest ona jako wyróżnianie wpisów, które zawierają dowolną nazwę partii oraz słowo uznane jako „sentyment". Ze względu na wagę analizy sentymentów, stanowi ona odrębną – opublikowaną tutaj – część całego badania. Wyróżnienie słów uznanych jako „sentyment", w związku z brakiem autorytatywnej listy polskich słów uznanych jako „sentyment", wykonano na bazie listy wyrażeń nasyconych emocjonalnie ANEW 2012[19]. Spośród 1031 słów z tego zbioru wybrano słowa skrajnie pozytywne i skrajnie negatywne, a wśród nich te, które miały największą częstotliwość występowania w przywołanym zbiorze. Słowa te przetłumaczono następnie na język polski, rozszerzając, w razie potrzeby, ich znaczenia, np. przy tłumaczeniu słowa „love" użyto zarówno formy „miłość" – rzeczownik, jak i „kochać" – czasownik itd.

Do badania zostały zakwalifikowane komitety wyborcze powiązane z partiami/środowiskami politycznymi[20], których członkowie zasiadali w Sejmie Rzeczypospo-

[18] P. Kuczma, W. Gogołek, *Informacyjny potencjał sieci*, s. 35–48.

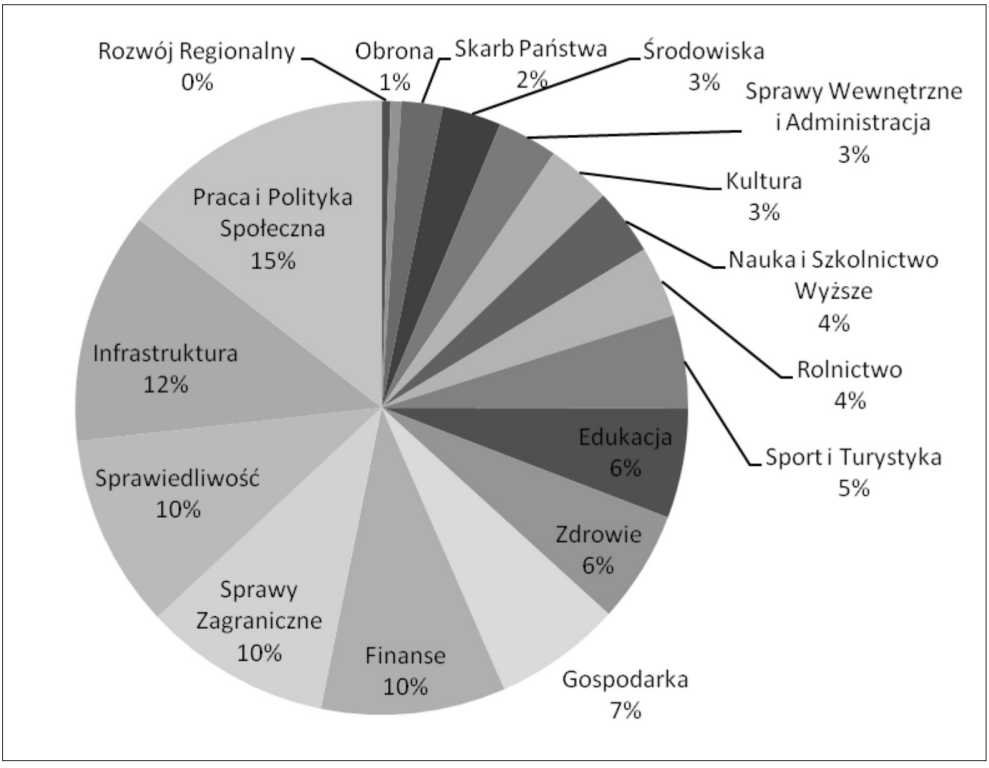
[19] M. M. Bradley, P. J. Lang, ANEW – Affective Norms for English Words. Lista wszystkich słów nie jest publicznie dostępna. Została uzyskana na specjalną prośbę od twórców strony *The Center for the Study of Emotion and Attention*, http://csea.phhp.ufl.edu/media/anewmessage.html i jest w posiadaniu autorów artykułu.

[20] Ruch Poparcia – partia założona przez Janusza Palikota – oraz Polska Jest Najważniejsza, która 1 stycznia 2011 była jeszcze stowarzyszeniem, a została zarejestrowana jako partia polityczna w 2011 r.,

litej 1 stycznia 2011 (w tym nowo powstałe twory polityczne obecne w Sejmie związane z posłem Januszem Palikotem i Joanną Kluzik-Rostkowską).

W celu dokonania ilościowej oceny krotności występowania nazw partii w tekstach zamieszczanych w sieci wyróżniono odpowiednie konteksty. Były nimi klucze (zwroty/słowa) wiązane z rządem, jego funkcjami i kompetencjami poszczególnych ministerstw[21] (**konteksty merytoryczne**, rys. 7). Słowa opisujące kompetencje każdego z ministerstw zostały oparte na kompetencjach ministerstw zapisanych w ich statutach[22]. Liczba pozyskanych wpisów z niektórych drukowanych gazet przekroczyła 550 000.

Drugą grupę kontekstów – **konteksty medialne** – stanowią te, które są związane z bieżącymi wydarzeniami relacjonowanymi w mediach (rys. 8). Zostały one



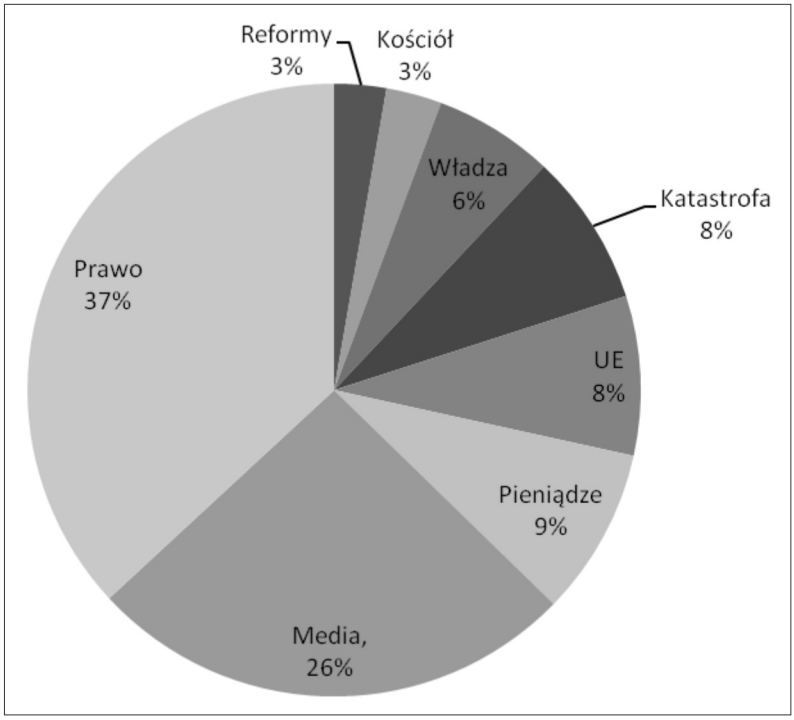
Rys. 7. Konteksty merytoryczne

Źródło: Obliczenia własne.

http://www.rp.pl/artykul/561429,603438.html [dostęp: listopad 2012].

[21] Na podstawie struktury Rady Ministrów za: Postanowienie Prezydenta Rzeczypospolitej Polskiej z dnia 16 listopada 2007 r. nr 1131–50-07 o powołaniu w skład Rady Ministrów.

[22] Spis statutów znajduje się w Aneksie opublikowanym na stronie: http://www.id.uw.edu.pl/zasoby/profile/59/Aneks_nr_2-Wykaz_statutow_ministerstw.pdf [dostęp: kwiecień 2012].



Rys. 8. Konteksty medialne

Źródło: Obliczenia własne.

wyłonione na podstawie formalnej analizy treści prasowych (przy użyciu programu QDA Miner v3.2 wraz z WordStat 6.0.1)[23] z największych polskich dzienników opiniotwórczych (w wersji elektronicznej)[24] o odmiennym profilu politycznym, jakimi są „Gazeta Wyborcza" oraz „Rzeczpospolita". Do tej analizy wykorzystano elektroniczną wersję dzienników drukowanych dostępnych za pomocą wyszukiwarki Factiva[25]. Artykuły pochodziły z okresu 1–28 lutego 2011 r., czyli z miesiąca poprzedzającego rozpoczęcie właściwego badania. Wszystkie artykuły wraz tytułami przeanalizowano pod względem ilościowym. Otrzymano w ten sposób listę 39 153 słów. Spośród nich wybrano listę 1000 słów, które powtarzały się statystycznie istotnie najczęściej – przynajmniej 32 razy we wszystkich analizowanych artykułach[26].

[23] Programy dostępne na stronie www.provalisresearch.com/Download/download.html [maj 2010]. Używana była wersja testowa.

[24] www.wirtualnemedia.pl/artykul/gazeta-wyborcza-i-fakt-to-najchetniej-czytane-dzienniki# [dostęp: maj 2010].

[25] https://han.buw.uw.edu.pl/han/ISIEM/site.securities.com/search/pub_search.html?pc=PL&sv=EMIS, [dostęp: maj 2010].

[26] W związku z tym, że słowo na miejscu 1000 miało częstotliwość występowania 32, do analizy włączono wszystkie słowa z częstotliwością przynajmniej 32. Było ich w sumie 1016.

Intensywność występowania kontekstów merytorycznych (150 000) jest niemal dwukrotnie większa od intensywności występowania kontekstów medialnych (90 000). Proporcje częstotliwości występowania poszczególnych słów-kluczy tworzących konteksty medialne i merytoryczne wskazują rangę, jaką przypisywano owym kluczom.

W mediach, od sierpnia do dnia wyborów, niektórym sprawom zasadniczym (merytoryczne konteksty, np. zdrowie i edukacja po 6%) poświęcano znacznie mniej uwagi niż atrakcyjnym medialnie kontekstom. Na przykład problemy mediów (w proporcji bezwzględnej do kontekstów merytorycznych – 26% × 9/15) poruszane były niemal trzykrotnie częściej (15,6%) od problemów merytorycznych – edukacja i zdrowie.

Wydaje się, że tego rodzaju informacje mogą mieć istotne znaczenie dla twórców materiałów dziennikarskich, np. w trosce o ograniczenie tabloidyzacji publikowanych treści.

Dane wejściowe obejmowały wszystkie dostępne, związane z partiami politycznymi w okresie marzec–wrzesień 2011 r. zasoby informacyjne forów, blogów, Facebooka, Twittera i dostępne w sieci portale informacyjne[27]. Zważywszy na względnie małe liczby wpisów na Facebooku i Twitterze (poniżej 1%), dane pozyskane z tych źródeł nie podlegały dalszej analizie.

Istotnym etapem procedury wykorzystania danych źródłowych do dalszych badań było wskazanie zmiennych niezależnych. Stanowiły one punkt odniesienia do oceny wiarygodności uzyskanych wyników rafinacji. Przyjęto, że owe zmienne tworzą: koszty poniesione przez partie na kampanię wyborczą, finalnie otrzymane liczby głosów, i wyniki sondaży CBOS.

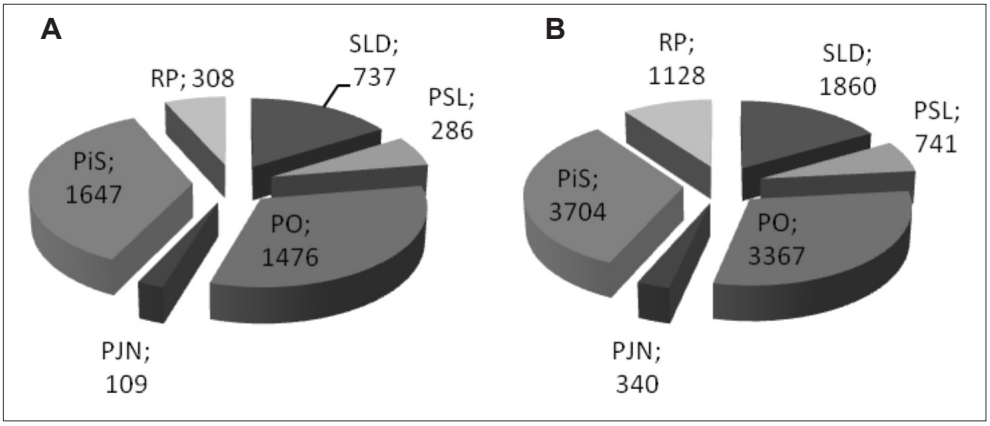
Analiza jakościowa

Przystępując do jakościowej analizy uzyskanych wyników, obliczono korelację liczb głosów uzyskanych przez wszystkie partie z wynikami sondaży (CBOS), wyniosła ona $r = 0{,}96$ ($p > 0{,}001$). Stanowi to przyjętą formę oceny wiarygodności sondaży przeprowadzonych przez CBOS i uzasadnienie założenia, że wyniki CBOS (w poszczególnych miesiącach) są wiarygodnym odniesieniem do dalszych badań.

Zważywszy na dominującą liczbę wpisów pozytywnych i negatywnych dla PO i PIS (fora – 62%, blogi – 67%), dalsza analiza informacyjnej siły rafinacji przeprowadzana była na przykładzie tylko tych dwóch partii (por. rys. 9).

Korelacja liczb głosów uzyskanych przez partie z liczbami pozytywnych wpisów na blogach wyniosła $r = 0{,}93$ ($p > 0{,}001$) (por. rys. 12). Dowodzi to niemal

[27] W. Gogołek, P. Kuczma, *Rafinacja informacji sieciowych na przykładzie wyborów parlamentarnych. Część 2. Portale internetowe, konteksty medialne i merytoryczne*, „Studia Medioznawcze" 2013, nr 3(54).



Rys. 9. Rozkład sumy wpisów pozytywnych i negatywnych (marzec–październik) na forach (A) i blogach (B) dla wszystkich partii w procentach

Źródło: Obliczenia własne.

pewnej zależności uzyskanych z rafinacji wyników z rzeczywistymi wynikami głosowania. Wskazuje jednocześnie na zasadność pogłębionej analizy zasygnalizowanej prawidłowości – znaczące zbieżności wyników rafinacji z oficjalnymi wynikami głosowań.

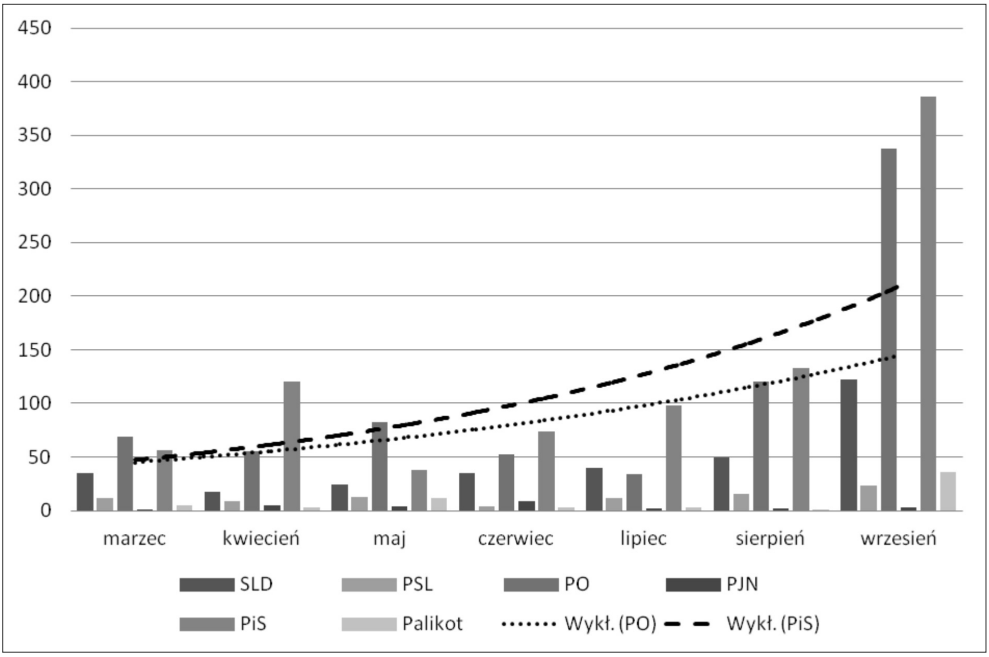
Fora

W trzecim kroku rafinacji dokonano wstępnej analizy ilościowej uzyskanych wyników – głównie na podstawie ich wizualizacji. Poza bezpośrednimi wynikami uzyskanymi z rafinacji, dodatkową formą wizualizacji są krzywe ilustrujące kierunek zmian liczb wpisów pozytywnych i negatywnych dla wiodących w wyborach dwóch partii: PIS i PO (rys. 10).

Dzięki temu wyróżniono widoczne prawidłowości/zależności zmiennych uzyskanych z rafinacji wpisów. Liczby wpisów pozytywnych i negatywnych na forach wskazały celowość obliczania różnic pomiędzy liczbami tych wpisów. Ilustracja różnic (rys. 11), wskazuje widoczną siłę predykcyjną ostatecznych wyników wyborów na podstawie wyników rafinacji.

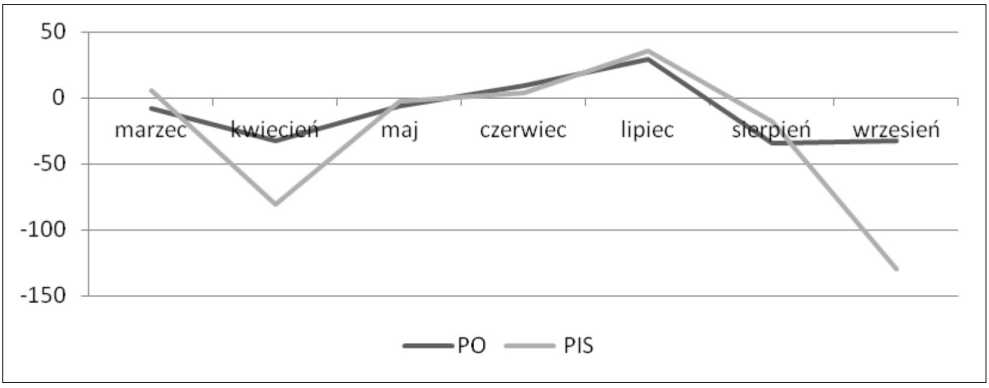
Informacyjnej wagi uzyskanych dzięki rafinacji informacji dowodzą miary ilościowe zależności zmiennych badań: zmiennych niezależnych (danych źródłowych) z wynikami rafinacji (zmiennych zależnych).

Przyjętym kryterium wiarygodności uzyskanych wyników rafinacji była wartość korelacji ilościowych wyników sondaży z wynikami rafinacji. Przykładem tego jest korelacja pomiędzy liczbami pozytywnych i negatywnych wpisów na forach



Rys. 10. Rozkład liczb wpisów negatywnych na forach oraz wykładnicze krzywe trendów (wykł.) zmian tych liczb dla PO i PIS

Źródło: Obliczenia własne.



Rys. 11. Rozkład różnic pomiędzy liczbami wpisów pozytywnych a liczbami wpisów negatywnych na forach dla PO i PIS (bilans między wpisami negatywnymi i pozytywnymi, wartości dodatnie odzwierciedlają przewagę wpisów pozytywnych)

Źródło: Obliczenia własne.

a wynikami sondaży (CBOS) dla wszystkich komitetów wyborczych w kolejnych miesiącach od marca do października 2011 r. Wartości tych współczynników są statystycznie znaczące ($p > 0{,}001$) z wyjątkiem października, który (ze względu na datę wyborów) nie obejmował danych z całego miesiąca. Dowodzi to statystycznie istotnej zbieżności wyników sondaży z wynikami rafinacji (tabela 1).

Tabela 1. Korelacja (r) związku wartości wyników sondaży CBOS z liczbami pozytywnych i negatywnych wpisów na forach dla wszystkich partii w kolejnych miesiącach

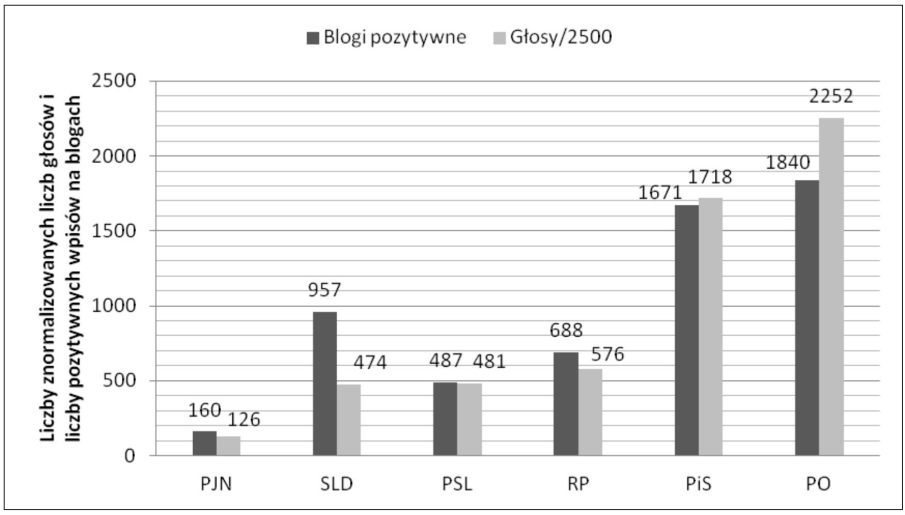
Miesiące	r dla wpisów pozytywnych	r dla wpisów negatywnych
Marzec	0,70	0,95
Kwiecień	0,81	0,76
Maj	0,83	0,98
Czerwiec	0,89	0,84
Lipiec	0,55	0,46
Sierpień	0,88	0,93
Wrzesień	0,90	0,87
Październik	0,23	0,12

Źródło: Obliczenia własne.

Wykazano także istotny związek pomiędzy liczbami pozytywnych wpisów na forach z liczbami uzyskanych przez partie głosów – współczynnik korelacji liczb pozytywnych wpisów na forach z liczbami głosów wyniósł 0,93 ($p < 0{,}001$).

Blogi

Nieco większą od forów wartość informacyjną o uzyskanej przez partie liczbie głosów mają blogi. O wspomnianej wartości informacyjnej świadczy ilustracja proporcji pozytywnych wpisów na blogach z wynikami wyborów (rys. 12). Współczynnik korelacji pomiędzy tymi zmiennym wyniósł 0,95 ($p < 0{,}001$). **Dowodzi to o niemal pewnej wiarygodności informacji pozyskiwanych z rafinacji wpisów na blogach. Można przyjąć, że blogi są istotnym źródłem informacji o liczbach głosów uzyskanych w wyborach.**



Rys. 12. Ilustracja podobieństwa/różnic proporcji (nie bezwzględnej wielkości) pozytywnych wpisów na blogach z liczbą uzyskanych głosów. W celu poprawienia efektu wizualizacji danych dokonano normalizacji danych wejściowych: liczbę głosów pomniejszono 2500 razy.

Źródło: Obliczenia własne.

Wiarygodność wyliczonych statystyk

Wartość informacyjną wyliczonych współczynników zależności potwierdzają badania ich statystycznej istotności. Służyły temu badania związku wartości wyników sondaży CBOS z liczbami negatywnych i pozytywnych wpisów na blogach dla wszystkich partii (tabela 2). Na podstawie uzyskanych wyników (Spearman i Pearson

Tabela 2. Korelacja (r) związku wartości wyników sondaży CBOS z liczbami pozytywnych i negatywnych wpisów na blogach dla wszystkich partii w kolejnych miesiącach

Miesiące	r dla wpisów pozytywnych	r dla wpisów negatywnych
Marzec	0,95	0,78
Kwiecień	0,97	0,84
Maj	0,98	0,84
Czerwiec	0,95	0,93
Lipiec	0,72	0,42
Sierpień	0,84	0,79
Wrzesień	0,67	0,24
Październik	0,38	0,81

Źródło: Obliczenia własne.

> 0,9 z istotnością > 0,95) można z pomijalnym błędem statystycznym mówić o niemal pełnej statystycznej zależności treści pozytywnych i negatywnych wpisów na blogach z sondażami. Przyjęto, że wyniki te można uogólnić i obejmują one także rezultaty uzyskane z rafinacji wpisów uzyskanych z forów.

O detekcyjnej sile rafinacji świadczy m.in. analiza ilościowa kosztów poniesionych przez poszczególne partie w okresie wyborczym. Wskazuje na całkowity brak związku (r = 0,06) pomiędzy kosztami utworzenia i utrzymania strony internetowej komitetu a liczbą uzyskanych głosów (tabela 3). Dowodzi to nadzwyczaj małej (wymownej) skuteczności deklarowanych przez komitety nakładów na strony internetowe. Znacznie większe zależności głosów stwierdzono z innymi nakładami.

Tabela 3. Korelacje (r) kosztów kampanii z liczbami uzyskanych głosów

Koszty kampanii	Współczynnik korelacji Pearsona
Koszty utworzenia i utrzymania strony internetowej komitetu	0,06
Wykonanie materiałów wyborczych, w tym prace koncepcyjne, prace projektowe i wytworzenie	0,61
Suma wszystkich nakładów	0,72
Korzystanie ze środków masowego przekazu i nośników plakatów	0,75
Internet vs całość wydatków na media	0,76
Reklama w Internecie (koszt usługi emisji)	0,86
Suma wszystkich nakładów na Internet	0,89
Reklama w Internecie	0,98

Źródło: Obliczenia własne.

Zakończenie

Ogrom cyfrowych zasobów informacyjnych i technologie szybko rewolucjonizują narzędzia i metody konstrukcji, gromadzenia i dystrybucji materiałów dziennikarskich. Rafinacja owych zasobów kreuje nową przestrzeń wartościowych źródeł informacji i otwiera nowe drogi do badań nad ich poszukiwaniem.

Uzyskane wyniki badań dowodzą podobieństwa, niemal identyczność, uzyskiwanych dzięki rafinacji danych o poparciu dla poszczególnych partii politycznych uczestniczących w wyborach parlamentarnych 2011 r. z wynikami sondaży opinii

publicznej oraz oficjalnymi wynikami ogłoszonymi przez Państwową Komisję Wyborczą. Dowodzi to oczekiwanej wartości informacyjnej zastosowań rafinacji.

Analiza rozkładu sentymentów stwarza nawet szanse na predykcje co do przyszłych zmian szacowanych wartości danych. Zmierzanie ku temu celowi – stworzenie funkcji predykcji – wymaga zaangażowania narzędzi statystycznych uwzględniających jednocześnie wiele parametrów, np. w postaci funkcji regresji wielokrotnej. Owe parametry to m.in. wiarygodnie wskazane w artykule wartości sentymentów pozyskiwanych z wpisów na forach i na blogach. Niemniej ważne od wartości merytorycznych uzyskanych wyników jest wielokrotnie mniejszy koszt uzyskanych dzięki rafinacji informacji, wobec kosztów zdobywania tych samych informacji w tradycyjny sposób – drogą sondaży.

Wyniki przedstawionych badań dowodzą, że Big Data to już nie *terra incognita* dla nauk społecznych. Ważnym wyzwaniem jest doskonalenie metodologii rafinacji oraz opracowanie stosownych narzędzi do rafinacji informacji sieciowej, spolegliwego dostarczania wyników użytkownikom, a co najważniejsze – przekonania o użyteczności tego nowego źródła informacji.

Bibliografia

Anstey C., *Empowering Citizen Cartographers*, „The New York Times" 2012, Jan. 13.

Beck A., *Big Data Is Never Too Big When You Can Act On It*, 2012, May 2, http://www.clickz.com/print_article/clickz/column/2171482/act?wt.mc_ev=click&WT.tsrc=Email&utm_term=&utm_content=Print%20version&utm_campaign=05%2F02%2F12%20-%20Behavioral%20Marketing&utm_source=ClickZ%20Media&utm_medium=Email [dostęp: maj 2012].

Big data: The Next Frontier for Innovation, Competition, and Productivity, http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation [dostęp: maj 2011].

Boswell W., *Five Search Engines You Can Use to Search the Deep Web*, http://websearch.about.com/od/invisibleweb/tp/deep-web-search-engines.htm [dostęp: marzec 2012].

Butler D., *Web Data Predict Flu*, „Nature" 2008, No. 456, http://www.nature.com/news/2008/081119/full/456287a.html [dostęp: luty 2012].

Deacon D., *Yesterday's Paper and Today's Technology, Digital Newspapers Archives and "Push Button" Content Analysis*, „European Journal of Communication" 2007, Vol. 22, No. 1.

„First Monday" 2011, Vol. 16, No. 9–5, September, http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3663/3040 [dostęp: styczeń 2012].

Gogołek W., Kuczma P*., Rafinacja informacji sieciowych na przykładzie wyborów parlamentarnych. Część 2. Portale internetowe, konteksty medialne i merytoryczne*, „Studia Medioznawcze" 2013, nr 3(54).

Jean-Baptiste M. i in. *Quantitative Analysis of Culture Using Millions of Digitized Books* „Science", Vol. 331, No. 6014, http://www.sciencemag.org/content/331/6014/176, accessed 1 June 2011. http://www.culturomics.org/cultural-observatory-at-harvard [dostęp: kwiecień 2012].

Kuczma P., Gogołek G., *Informacyjny potencjał sieci – na przykładzie wyborów prezydenckich 2010*, „Studia Medioznawcze" 2010, nr 4(43).
Pfanner E., *Google, in Settlement, Changes Ad Rules in France*, „The New York Times" 2010, October 28.
Shen M. in., *Quantitative Analysis of Culture Using Millions of Digitized Books*, „Science" 2011, Vol. 331, Issue 6014.
Stephens-Davidowitz S., *Google's Crystal Ball*, „The New York Times", http://campaignstops.blogs.nytimes.com/2012/10/20/googles-crystal-ball/?_php=true&_type=blogs&_r=0, October 20, 2012 [dostęp: kwiecień 2013].
The Carbon Capture Report, http://www.carboncapturereport.org/ [dostęp: wiosna 2012].
Unlocking the Value of Personal Data: From Collection to Usage, World Economic Forum, February 2013, http://www3.weforum.org/docs/WEF_IT_UnlockingValuePersonalData_CollectionUsage_Report_2013.pdf [dostęp: maj 2013].
Wolfgang Martin Team, *BI meets BPM and Big Data / Wolfgang Martin,* March 2013. http://www.wolfgang-martin-team.net/BI-BPM-SOA.php [dostęp: maj 2013].
Wollan M., *For Start-Ups That Aim at Giants, Sorting the Data Cloud Is the Next Big Thing,* The New York Times, December 25, 2011.

Streszczenie

Ocenia się, że ilość danych wytworzonych przez człowieka w ciągu ostatnich dwóch lat (2011–2013) przewyższyła ilość informacji wyprodukowanych do tego momentu w całej historii ludzkiej cywilizacji. Specjalistyczna analiza tych informacji stwarza nowe źródło informacji dziennikarskich. Proces uzyskiwania nowych informacji z Big Data określono rafinacją informacji sieciowej. Umożliwia on opisywanie przeszłego i bieżącego informacyjnego obrazu rzeczywistości, a nawet służy do predykcji i wskazywania ważnych problemów cywilizacji. Opracowanie zawiera kilka przykładów tych możliwości, m.in. związanych z informacjami opisującymi wybory parlamentarne w Polsce.

INFORMATION POTENTIAL OF REFINED NETWORK RESOURCES

Summary

It is estimated that the amount of data processed by man in the past two years (2011–2013) exceeds the sum of information produced to the present time in the anals of human civilization. Specialist analysis of this information establishes a new source of journalist information. The process of new information acquisition from the Big Data is defined as Network information refining. It enables the delineation of the reality of the past and present information picture, and even assists in the prediction and indication of important civilization problems. The study incorporates instances of these possibilities, among others, those related to information concerning parliamentary elections in Poland.