AUTOMATYKA • 2009 • Tom 13 • Zeszyt 2

Anna Zygmunt*, Jarosław Koźlak*, Łukasz Krupczak*, Bartosz Małocha*

Analiza blogów internetowych przy użyciu metod sieci społecznych

1. Wprowadzenie

Blog (*weblog*) można traktować jak swego rodzaju sieciowy dziennik, pamiętnik. Jest to strona internetowa, na której autor umieszcza datowane wpisy, wyświetlane zazwyczaj w odwrotnej kolejności chronologicznej (najpierw najświeższe). Blogosferę (*blogosphere*) można traktować jako wspólną przestrzeń wszystkich blogów.

Na blogach komentowane są opisywane wydarzenia polityczne, naukowe lub kulturalne. Blogi mogą również pełnić rolę osobistych pamiętników. Typowy blog składa się z kombinacji tekstu, zdjęć, odsyłaczy do innych blogów lub stron internetowych. Blogi można kategoryzować, opisując je za pomocą znaczników zwanych tagami.

Wielkie zainteresowanie blogami i ich ogromny rozwój związany jest niewątpliwie z właściwą im interaktywnością, a więc możliwością praktycznie nieograniczonego dopisywania komentarzy. Technorati [8] szacowało, że w 2008 roku codziennie tworzonych było 175 tys. komentarzy do blogów. Zazwyczaj dostęp do blogów jest w pełni otwarty, możliwy jest odczyt wpisów właściciela blogu, komentarzy wpisywanych przez użytkowników odwiedzających blog oraz zawartych na nim odsyłaczy do innych blogów.

Obecnie bardzo popularne są mikroblogi (*micro-blog*), które różnią się tym od klasycznych blogów, że wiadomości zapisywane przez użytkowników są bardzo krótkie, np. informacje o tym, co robi w danej chwili. Z tego względu mikroblogi porównuje się często do publikowanych statusów użytkowników na gadu-gadu czy skype. Najbardziej popularnym przykładem mikroblogu jest Twitter [9].

Początki blogowania sięgają roku 1994 i właściwie do 1999 roku rozwijały się stosunkowo wolny. Gwałtowny rozwój można było zaobserwować od 1999 roku, a obecnie jesteśmy świadkami lawinowej eksplozji liczby blogów w Internecie: „w styczniu 2004 było w Internecie około 1 miliona blogów, w połowie 2006 populacja blogosfery przekroczyła 50 milionów i stale rośnie” [10].

* Katedra Informatyki, Akademia Górniczo-Hutnicza w Krakowie

Obecnie blogowanie jest zjawiskiem powszechnym i globalnym [8]: w czerwcu 2008 roku blogi były pisane w 81 językach, przez blogerów pochodzących z 66 krajów na sześciu kontynentach.

Celem prowadzonych badań jest analiza wybranej kategorii blogów pod kontem identyfikacji wpływowych blogerów oraz analizy ewolucji ich popularności wraz z upływem czasu, ważności poruszanych tematów oraz wzajemnych wpływów.

2. Kierunki badań nad blogosferą

Blogosfera z racji swego ogromnego zasięgu i różnorodności prezentowanych problemów stanowi cenne źródło informacji i dlatego poddawana jest różnorodnym badaniom. Wśród najciekawszych można wymienić [1]:

- **Modelowanie blogosfery**, czyli opracowanie odpowiedniego modelu najlepiej oddającego charakter i strukturę blogosfery. Mając zbudowany model, można lepiej zrozumieć strukturę i właściwości blogosfery: zależności między blogerami, postami czy różnymi rodzajami blogów.
- **Klastrowanie blogów**, czyli automatyczne łączenie blogów w grupy podobnych w celu ułatwienia przeszukiwania blogosfery. Klasyfikacja taka może być przeprowadzona np. w oparciu o tagi opisujące blogi [2].
- **Eksploracja blogów** (*Blog Mining*), np. w celu śledzenia przekonań i opinii, reakcji, rozpoznawanie trendów, modnych sformułowań, badanie śladów przepływów informacji, dzielenia opinii oraz wpływów.
- **Badanie wpływów**, np. poprzez identyfikację wpływowych blogerów można wpływać na wartość sprzedaży czy głosy wyborcze. Zamiast analizować wszystkie blogi, lepiej jest wybrać te „najważniejsze", przy czym istotne jest nie tylko wykrycie wpływowych członków lub ekspertów uwzględniając sposób współdzielenia wiedzy w środowisku, ale także określić, do jakiego stopnia niektórzy są uważani za ekspertów przez członków społeczności.
- **Propagacja zaufania** (*propagation of trust*) poprzez szacowanie zaufania i reputacji ekspertów, np. przyznawanie punktów reputacyjnych.
- **Wyodrębnianie społeczności** (*community extraction*), traktowane jako rozszerzenie klastrowania: analiza zawartości postów i komentarzy oraz ich autorów.
- **Filtrowanie spamu** (*spam blog (splogs) filtering*), czyli odróżnianie normalnych stron od stron będących spamem, oparte np. na proporcjach statystycznych dotyczących występujących słów, średniej ilości słów oraz URL.
- **Analiza opinii i nastrojów** (*opinion and sentiment analysis*).

Istotną cechą blogów jest brak istnienia cenzury, co niesie ogromne ryzyko wykorzystywania ich w celach przestępczych. Wyniki otrzymane w przedstawionych powyżej obszarach wykorzystywane są w badaniach związanych z przeciwdziałaniem przestępczości w cyberprzestrzeni.

Blogi mogą być wykorzystywane do głoszenia nielegalnych poglądów (np. faszystowskich), promowania przestępczych zachowań (np. pedofilia), propagowania nienawiści, głoszenia treści manipulujących i nieprawdziwych, prezentowania zakazanych treści, jak również do budowy społeczności zwolenników lub użytkowników danych materiałów (wymiana informacji, propaganda, pozyskiwanie nowych zwolenników/użytkowników). Innym rodzajem przestępczych zachowań jest podszywanie się pod daną osobę (np. w celu jej skompromitowania poprzez głoszenie kontrowersyjnych poglądów w jej imieniu lub też udostępnianie zdjęć przedstawiających ją w niekorzystnym świetle) lub wykorzystanie skradzionej tożsamości osoby do dalszych działań o charakterze przestępczym.

3. Sieci społeczne i ich analiza

Interakcje międzyosobowe zachodzące we współczesnym świecie są coraz bardziej zróżnicowane (np. spotkania, rozmowy telefoniczne, wymieniane e-maile, połączenia za pomocą komunikatorów internetowych), a w miarę rozwoju technologii komputerowych coraz łatwiejsze do obserwacji i analizy. Zależności te mogą być opisane za pomocą sieci powiązań (grafów), w których role węzłów pełnią osoby (lub grupy osób), a powiązania reprezentują zachodzące między nimi określonego rodzaju interakcje. Dziedziną zajmującą się badaniem występujących w takich grafach zależności jest analiza sieci społecznych (*Social Network Analysis – SNA*). Metody SNA są szeroko stosowane w wielu dyscyplinach, przy analizie bardzo różnych zjawisk. Wyliczając i analizując wartości parametrów sieci, takich jak indeksy rozkładu, miary centralności, podobieństwa węzłów czy strukturę społeczności można opracować modele obrazujące rozprzestrzenianie się chorób czy wirusów komputerowych, formowanie grup. Traktowanie sieci społecznej jako grafu pozwala wykorzystać grafowe algorytmy eksploracji danych (*link mining*).

Ponieważ wszystkie blogi są umieszczone z definicji w Internecie i powiązane między sobą na wiele sposobów, można blogosferę traktować jak złożoną sieć społeczną i zamodelować w postaci grafu. Węzłami w takiej sieci są blogerzy, grupy blogerów, a krawędziami – odsyłacze (linki) między blogami lub wstawiane komentarze. W tak stworzonej sieci możemy badać stopień ważności węzłów i pełnione przez nich role w sieci, oraz wyodrębniać grupy blogerów mających wspólne zainteresowania.

Podstawowym parametrem SNA jest centralność (*centrality*). Opisuje ona pozycję węzła w strukturze sieci stanowiąc miarę ważności, znaczenia i jego wpływu na inne obiekty w sieci [7]. Wykorzystuje się wiele sposobów pomiaru centralności [3]. Środek ciężkości (*bary center*) [5], to miara opisująca dany węzeł, obliczana na podstawie wszystkich najkrótszych ścieżek prowadzących do tego węzła z innych węzłów. Bloger o najwyższym środku ciężkości może w najszybszy sposób uzyskać informację zgromadzoną we wszystkich blogach w sieci. Centralność betweenness (*betweenness centrality*) jest obliczana na podstawie wszystkich najkrótszych ścieżek do wszystkich węzłów. Wartość centralności betweenness jest tym większa dla danego węzła, im większa jest ilość wszystkich najkrót-

szych ścieżek przechodzących przez ten węzeł. Blogi o wysokich wartościach tej miary mogą być traktowane jako punkty krytyczne sieci, których usunięcie może spowodować utrudnienie przepływu informacji w sieci.

Innymi parametrami wykorzystywanymi do analizy sieci społecznych jest stopień wierzchołków wchodzących, obliczany na podstawie ilości krawędzi wchodzących do danego wierzchołka oraz – podobnie – stopień wierzchołków wychodzących obliczany na podstawie krawędzi wychodzących. Blog o najwyższej wartości stopnia wierzchołków wychodzących ma najwięcej linków czy komentarzy do innych blogów, podczas gdy blog o najwyższej wartości stopnia wierzchołków wchodzących pojawia się najczęściej wśród linków na innych blogach, bądź też jest najczęściej komentowany.

W analizie sieci społecznych bada się również sposób, w jaki węzły wskazują na siebie. Służą do tego parametry [4]: stopień koncentracji (*hubness)* i autorytet (*authoritativeness*). Blogi o wysokim współczynniku autorytetu są wskazywane przez wiele innych blogów, a te o wysokim współczynniku stopnia koncentracji wskazują na wiele blogów o wysokim autorytecie.

Innym parametrem wykorzystywanym do pomiaru ważności wskazać jest ranking Page'a (*PageRank*), wykorzystywany przez przeglądarkę internetową Google. W przypadku sieci społecznej, parametr ten może być interpretowany jako wyznacznik posiadanej informacji.

4. Analiza wybranej kategorii blogów

W naszych pracach skupiliśmy się na analizie wybranych blogów dostępnych w portalu *www.salon24.pl*. Przygotowano środowisko do takich analiz obejmujące moduł do pobierania zawartości blogów, bazę danych do ich przechowywania oraz aplikacje do przeprowadzania różnych rodzajów analiz. Wykorzystywano metody sieci społecznych do analizy statycznej (w oparciu o wszystkie zebrane dane) oraz dynamicznej (analiza stanu sieci w kolejnych przedziałach czasu), statystyczną analizę otagowania postów a także analizę zawartości tekstowej poszczególnych postów oraz związków pomiędzy postami na różnych blogach biorąc pod uwagę reprezentację wektorową tekstu oraz czasy umieszczenia poszczególnych postów i komentarzy.

W analizach w oparciu o metody sieci społecznych wykorzystano klasyczne miary sieci społecznych takie jak: stopnie wierzchołków liczone dla powiązań przychodzących i wychodzących, centralność środka ciężkości, centralność betweeness, koncentracja i autorytet oraz ranking Page'a.

Analizowana sieć społeczna mogła być tworzona na trzy sposoby i w rezulacie można uzyskać trzy różne grafy rozpatrując dany zbiór blogów:

– w oparciu o dostępne na blogach stałe odsyłacze do innych blogów, odsyłacze te wskazują na blogi, które właściciel danego blogu uznaje za interesujące i godne polecenia, często dlatego, że podziela poglądy na nich prezentowane;

– w oparciu o informacje o autorach komentarzy do postów na danym blogu, jeśli autorem komentarza jest właściciel jakiego bloga, to można poprowadzić z węzła reprezentującego jego blog powiązanie do węzła blogu na którym wpisał on swój komentarz;
– uwzględniając oba omawiane powyżej rodzaje powiązań.

Badanie blogów opierało się na porównaniu wartości uzyskanych dla poszczególnych miar oraz na pozycjach rankingowych zajętych dla poszczególne miary. W celu wyboru najbardziej znaczących blogów przyjęliśmy algorytm przydziału punktów dla 20 blogów uzyskujących najlepsze wartości dla każdej z miar (punktując od 1 za miejsce 20. do 20 za miejsce pierwsze), a następnie te punktowania zostały zsumowane. Analizując w ten sposób trzy zbudowane grafy uzyskaliśmy trzy rankingi blogów, różniące się od siebie. W tabeli 1 pokazane jest po 5 najwyżej punktowanych blogów dla każdego z grafów. Uwzględniając z kolei punkty zdobyte za miejsca w tych trzech sieciach przez 30 pierwszych blogów (i przyznając po 30 punktów za pierwsze miejsce i odpowiednio coraz mniej za kolejne, aż do 1 punktu za miejsce 30.), można uzyskać ostateczny sumaryczny ranking (tab. 2). Widać (tab. 1), że blog *http://kataryna.salon24.pl* znajdował się zawsze w pierwszej trójce, niezależnie od sposobu definiowania krawędzi w grafie reprezentującym sieć społeczną. Blogi takie jak:

– *http://maryla.salon24.pl*,
– *http://jankepost.salon24.pl*,
– *http://freeyourmind.salon24.pl*

zawdzięczają swoją wysoką pozycję dobrym wynikom w dwóch rankingach. Wśród blogów, które znajdują się wysoko we wszystkich rankingach znalazły się także: *http://boguslawchrabota.salon24.pl* oraz *http://dolinanicosci.salon24.pl*.

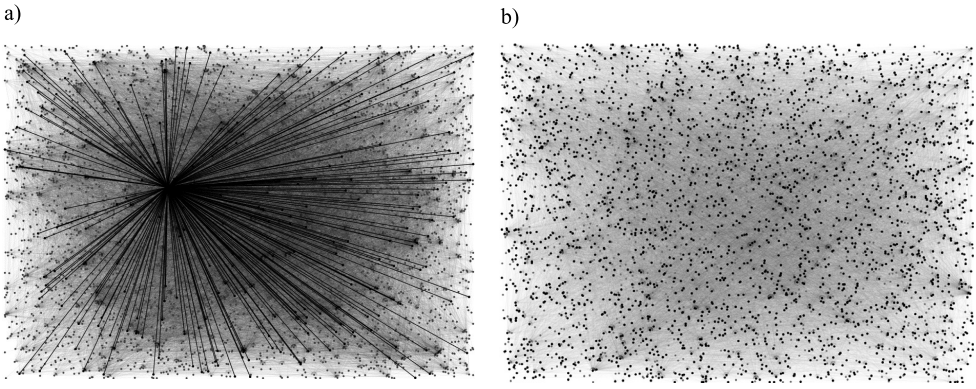
Tabela 1
Najwyższe pozycje rankingu blogów dla trzech tworzonych sieci społecznych

Lp.	Krawędzie według linków	Krawędzie według komentarzy	Krawędzie według linków i komentarzy
1	http://maryla.salon24.pl (166)	http://jankepost.salon24.pl (114)	http://kataryna.salon24.pl (142)
2	http://kataryna.salon24.pl (152)	http://tomaszsakiewicz.salon24.pl (103)	http://maryla.salon24.pl (137)
3	http://freeyourmind.salon24.pl (119)	http://kataryna.salon24.pl (99)	http://jankepost.salon24.pl (91)
4	http://krzysztofleski.salon24.pl (91)	http://jerzyjachowicz.salon24.pl (95)	http://freeyourmind.salon24.pl (88)
5	http://kluby-iv-rp.salon24.pl (82)	http://obserwator.salon24.pl (94)	http://zezem.salon24.pl (83)

Tabela 2
Sumaryczny ranking blogów

Blogger	razem	komentarze	linki	suma
Kataryna	30	28	29	87
Maryla	29	0	30	59
Igor Janke	28	30	0	58
Freeyourmind	27	0	28	55
Freeman	26	0	24	50
Bogusław Chrabota	16	25	2	43
Bronisław Wildstein	15	7	12	34

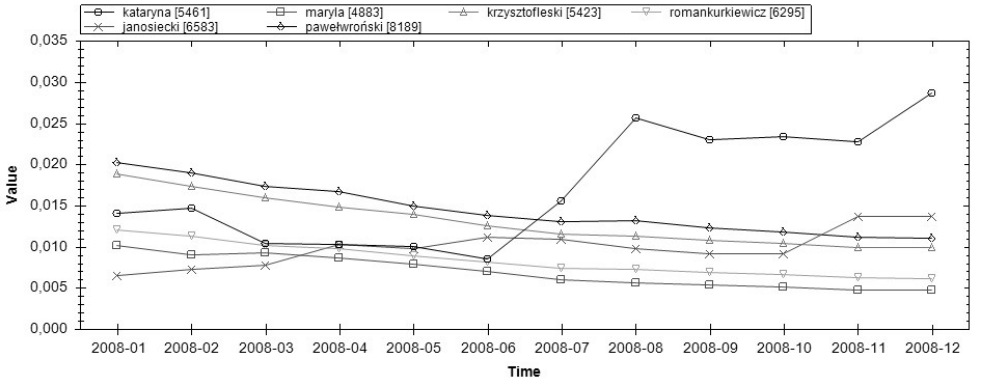
Na rysunku 1 pokazane są powiązania wybranych wysoko punktowanych blogów z innymi blogami.



Rys. 1. Przykłady znaczących blogów: a) wszystkie powiązania blogu http://jankepost.salon24.pl; b) wierzchołki do/z których z blogu http://kataryna.salon24.pl prowadzi ścieżka o długości maksymalnej 2

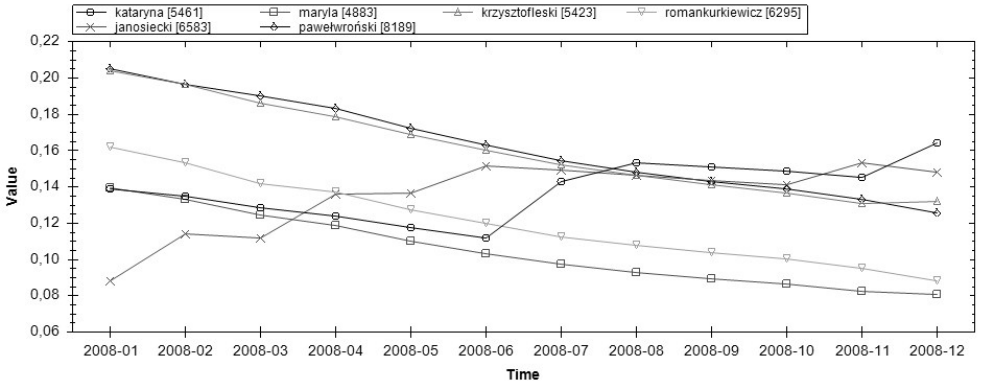
Kolejnym rodzajem przeprowadzonych analiz było wyszukiwanie znaczących tematów dyskutowanych na blogach poprzez analizę otagowania postów. Po odrzuceniu słów kluczowych w niewielkim stopniu charakteryzujących temat wypowiedzi (gdyż posiadające ogólne znaczenie jak np. *Polska*, *historia* czy *demokracja*) pozostały następujące słowa: *lustracja* (95), *PiS* (94), *dziennikarze* (78), *Kaczyńsk*i (73), *Kościół* (69), *Jarosław Kaczyński* (64), *Tusk* (64), *Rosja* (62), *Kwaśniewski* (54), *Michnik* (42), *Giertych* (41), *Gazeta Wyborcza* (40), *USA* (39), *agenci* (39), *etyka* (39).

Aby opisać zachowanie sieci społecznej, nie wystarczają jednak same zbiorcze informacje na temat interakcji w długim przedziale czasu, istotna jest także wiedza, jak interakcje te zmieniały się w poszczególnych okresach oraz jaki jest ich aktualny trend. Dlatego następnymi analizami, które zostały przeprowadzone, były analizy dynamiczne, pokazujące ewolucje wartości poszczególnych miar oraz pozycji w rankingach poszczególnych bloków w trakcie 2008 roku. Na rysunkach 2, 3 oraz 4 pokazane są wybrane wykresy, pokazujące ewolucję rankingu Page'a, HITS (opartego na autorytecie i koncentracji) oraz centralności betweeness dla wybranych, uznanych za znaczące, blogów.



Rys. 2. Ewolucja wartości rankingu Page'a

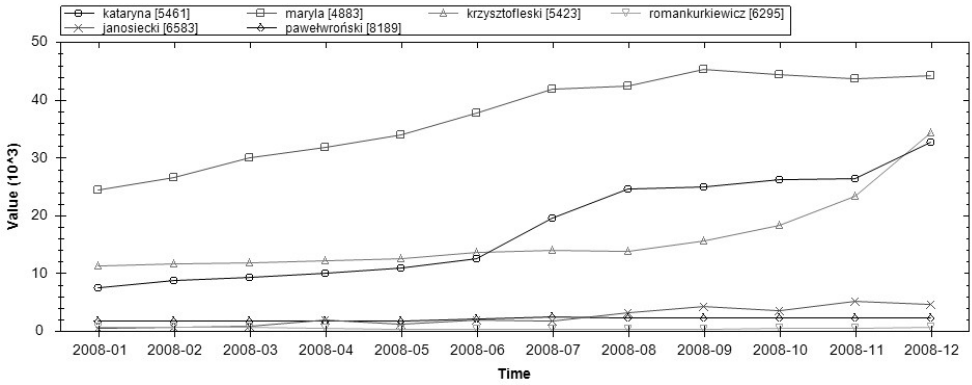
Analizując rysunek 2 można zauważyć znaczny wzrost wartości miar dla blogu *Kataryny*, oraz umiarkowany dla blogu *Jana Osieckiego*, przy pozostałych rozpatrywanych blogach tracących na wartości.



Rys. 3. Ewolucja wartości rankingu HITS

Na rysunku 3 pokazującym wartości HITS widać duży wzost wartości miar dla blogu *Kataryny*, po regularnym spadku trwającym od początku roku do czerwca, na przestrzeni

roku wzrosła także znacząca wartość miary dla blogu *Jana Osieckiego*. Miary pokazane na rysunkach 3 i 4 należy interpretować jako powiązania z ważnymi blogami, więc należy uznać, że *Kataryna* i w mniejszym stopniu *Jan Osiecki* uzyskały dużą ilość powiazań z innymi wysoko punktowanymi blogami.



Rys. 4. Ewolucja wartości rankingu centralności betweeness

Na rysunku 4 widać, że blogi *Kataryny*, *Maryli* oraz *Krzysztofa Leskiego* znajdują się na wielu najkrótszych ścieżkach wiodących pomiędzy blogami w blogosferze, oraz że nastąpił znaczący wzrost miary dla tych blogów, z tym, że dla *Maryli* raczej w pierwszym półroczu, a dla *Kataryny* i *Krysztofa Leskiego* – w drugim.

5. Podsumowanie

Uzyskane rankingi w znacznej mierze potwierdzają wynikającą z obserwacji wiedzę na temat popularnych blogerów i ich oddziaływania. W dalszych pracach zamierzamy skupić się na analizie zawartości tekstów w blogach, wyszukiwaniu blogów i postów przystających do zadanych wzorców oraz porównywaniu blogów przy użyciu statystycznych metod wektorowych. Planujemy także przeprowadzić analizy blogów dostępnych na innych serwerach.

Literatura

[1] Agarwal N., Liu H., *Blogosphere: Research Issues, Tools, and Applications.* SIGKDD Exploration, 10(1), 18–31, July 2008.

[2] Brooks C.H., Montanez N., *Improved annotation og the blogosphere via autotagging and hierarchical clustering.* Proceedings of the 15th International Conference on World Wide Web, ACM Press, NY, USA, 2006.

[3] Hanneman R.A., Riddle M., *Introduction to Social Network Methods.* University of California Press, 2005.

[4] Kleinberg J.M., *Authoritative Sources in a Hyperlinked Environments.* Journal of the ACM, 46(5), 1999.
[5] O'Madadhain J. Fisher D., Smyth P., White S., Boey Y.-B., *Analysis and Visualization of Network Data Using Jung.* Journal of Statistical Software, vol. VV, no. II.
[6] Wellman B., *From Little Boxes to Loosely Bounded Networks: The Privatization and Domestication of Community.* Sociology for the 21th Century: Continuities and Cutting Edges, University of Chicago Press, 1999.
[7] Wolfe A.P., Smyth P., *Algorithms for Estimating relative importance in network*. Proc. of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, 2003.
[8] http://www.technorati.com.
[9] http://www.twitter.com.
[10] Paul Gillin, The New Influencers, 2007.