STUDIA INFORMATICA 2010
Volume 31 Number 2A (89)

Anna ZYGMUNT, Jarosław KOŹLAK, Łukasz KRUPCZAK
Akademia Górniczo-Hutnicza, Katedra Informatyki

IDENTYFIKACJA WPŁYWOWYCH JEDNOSTEK W BLOGOSFERZE[1]

Streszczenie: W artykule zaprezentowano identyfikację wpływowych jednostek w blogosferze stosując metodę analizy sieci społecznej. Identyfikacja jest dokonywana przez obliczenie podstawowych miar sieci społecznych dla węzłów, rozpatrywanie wpływu poszczególnych postów na blogosferę oraz wyodrębnienie w sieci grup jednostek mających silne wzajemne powiązania społeczne.

Słowa kluczowe: sieci społeczne, blogi

IDENTIFYING THE INFLUENTIAL INDIVIDUALS IN BLOGOSPHERE

Summary: In the paper, methods of identification of the influential persons in the blogosphere using the social network analysis approach are presented. The identification is performed thanks to an estimation of the values of social network measures for nodes, an analysis of the influence of given posts onto the posts written by other authors, or by the recognition of the groups in the network composed from entities having strong reciprocal social links.

Keyword: social networks, weblogs

1. Wprowadzenie

Blog (ang. *weblog*) można traktować jak swego rodzaju sieciowy dziennik. Jest to strona internetowa, na której autor umieszcza datowane wpisy. Blogosferę (ang. *blogosphere*) można traktować jako wspólną przestrzeń wszystkich blogów.

[1] Prace badawcze prowadzące do opisanych wyników były częściowo finansowane z umowy grantowej nr 218086 Siódmego Programu Ramowego Unii Europejskiej.

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. 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. Aby uzyskać możliwość komentowania postów innych, zazwyczaj należy być zarejestrowanym użytkownikiem. Blogosfera charakteryzuje się dużą dynamiką zmian, a wpisy na blogach mają zazwyczaj krótki cykl życia; większość z nich jest przestarzała i nigdy więcej niecytowana.

Gwałtowny rozwój blogosfery można zaobserwować od 1999 roku, a obecnie jesteśmy świadkami lawinowej eksplozji liczby blogów w Internecie. Blogowanie staje się zjawiskiem powszechnym i globalnym: w czerwcu 2008 roku blogi były pisane w 81 językach, przez blogerów pochodzących z 66 krajów na sześciu kontynentach[2].

Blogi pełnią istotną rolę w rozprzestrzenianiu się informacji, dlatego istotne jest zrozumienie tego fenomenu poprzez analizę wzorców zachowań oraz przyczyn takiego a nie innego sposobu rozpowszechniania się informacji.

Prowadzone przez autorów badania obejmują swym zakresem analizę wybranej kategorii blogów pod kątem wyszukiwania w nich wpływowych blogerów, analizę ewolucji ich popularności oraz sposobów oddziaływania. Zaprezentowane w artykule prace oparte są na metodyce analizy sieci społecznych (ang. *Social Network Analysis – SNA*). Na podstawie wybranych metryk można określić role pełnione przez poszczególnych blogerów i wyodrębnić tych, którzy mają dużo większy wpływ na innych. Opracowaną metodę znajdowania wpływowych blogerów rozbudowano o możliwość analizy tekstów (w oparciu o metodę wektorową) i znajdowania podobnych postów.

2. Przegląd kierunków badań nad blogosferą

Fenomen ogromnego i cały czas rosnącego zainteresowania pisaniem blogów i komentowania zawartości innych blogów jest bardzo interesującym zjawiskiem nie tylko socjologicznym. Przyczyną pojawienia się takiego zjawiska jak blogowanie był rozwój technologii komputerowych i w konsekwencji pojawienie się Web 2.0. I tylko za pomocąmetod komputerowych możliwa jest analiza blogosfery, tak by socjologowie mogli dołożyć swoją wiedzę do-

[2] http://www.technorati.com

tyczącą mechanizmów społecznych. Bez tego wspomagania blogosfera pozostałaby ogromnym niezbadanym obszarem. Tak więc podstawowym kierunkiem badań jest 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. Blogosfera nie jest jednorodna, można wyodrębnić grupy skupione wokół jednego lub kilku blogerów, które łączą wspólne tematy wynikające z podobnych zainteresowań, opcji politycznych czy przekonań religijnych. Interesującym kierunkiem badań jest automatyczne łączenie blogów w grupy podobnych w celu ułatwienia przeszukiwania blogosfery. Podobieństwo można badać poprzez analizę zawartości postów i komentarzy oraz ich autorów, jak również analizując opisujące blogi. Blogi można eksplorować np. w celu śledzenia przekonań i opinii, reakcji, rozpoznawanie trendów, modnych sformułowań, badanie śladów przepływów informacji, dzielenia opinii oraz wpływów. Z kolei poprzez identyfikację wpływowych blogerów można oddziaływać 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.

Prace przez nas prowadzone dotyczą głównie znajdowania wpływowych blogerów i wykrywania oraz analizowania charakteru grup wpływów, które gromadzą się wokół takich blogerów. Interesująca jest dynamiczna analiza powiązań w blogosferze i zrozumienie mechanizmów

2.1. Wyszukiwanie wpływowych blogerów

Aby próbować znajdować wpływowych blogerów należy przede wszystkim zdefiniować, jakie powinien on mieć własności. Jak zauważono w [3], aktywny bloger niekoniecznie musi być wpływowy, jak również wpływowy niekoniecznie musi być aktywny. Zdefiniowanie aktywnego blogera jest stosunkowo proste: to ten, który pisze dużo postów. Niestety, trudno tę prostą statystykę przełożyć jako określenie wpływowego. Intuicyjnie wpływowy bloger, to taki, który ma dużo wpływowych postów, czyli takich, które mają dużo komentarzy i *inlinków* (linków na innych blogach), ale małą mała liczbę *outlinków* (linków w tekście do innych blogów), gdyż duża ich liczba sugerowałaby małą oryginalność poruszanego tematu. Zauważono również pozytywną korelację między długością posta a liczbą komentarzy; dłuższe posty zwracają uwagę.

W [1] zaproponowano algorytm *iRank*, który wykorzystuje przepływ informacji między postami, czyli linki. Linki w postach pojawiają się najczęściej w kontekście cytowania wypowiedzi innych, powoływania się na czyjąś opinię bądź komentowania pewnych zdarzeń w ce-

lu np. uwiarygodnienia swoich wypowiedzi. Ogólnie rzecz biorąc można przyjąć, że jeżeli dwa posty cytują te same linki, to ich treść dotyczy podobnych zagadnień.

W algorytmie tym wyszukujemy linków powtarzających się w co najmniej dwóch postach i – jeżeli blog A cytuje link zanim zacytuje go blog B – tworzymy krawędź od B do A. Im mniejsza różnica czasowa pomiędzy postami cytującymi ten sam link, tym większa waga krawędzi połączenia. W ten sposób powstanie graf skierowany, w którym wierzchołki odpowiadające najbardziej wpływowym osobom (czyli takim, które pierwsze zacytowały jakieś linki) mają największą liczbę krawędzi wchodzących. Według tego algorytmu wpływowe blogi to takie, które cytując pewne wiadomości powodują powstanie wielu dyskusji na ten temat lub zacytują pewien link, zanim stanie się popularny.

Można zauważyć, że w drugim przypadku blog może być uznany za wpływowy przez przypadek, dyskusja na dany temat nie musi być wywołana przez tego, który pierwszy umieścił dany link.

2.2. Wyszukiwanie grup

Jednym z podejść do znajdowania grup jest wykorzystanie zasady homofilii [6], która mówi, że ludzie częściej komunikują się między sobą, jeżeli mają podobne poglądy. Zgodnie z tą zasadą ci blogerzy, którzy mają więcej krawędzi między sobą, mogą być traktowani jako grupa. Oznacza to, że komunikują się oni dużo częściej między sobą niż z innymi, dzieląc wspólne poglądy, opinie, zainteresowania [2]. Do pomiaru siły i trwałości tych powiązań wprowadzono w [9] dwa parametry: gęstość (ang. *density*) i stabilność (ang. *stability*). Pierwszy z nich określa kompletność grupy w aspekcie istniejących powiązań między jej członkami i jest definiowany jako odsetek maksymalnej możliwej liczby powiązań, które występują między członkami [12] :

$$Density = \frac{2\sum_{i}^{n}\sum_{<j}^{n} a(i,j)}{n(n-1)}, \tag{1}$$

gdzie n jest licznością grupy, a $a(i,j)$ – zmienną binarną wskazującą czy istnieje link między blogiem i oraz j. Im grupa gęstsza, tym istnieje więcej powiązań między jej członkami.

Biorąc pod uwagę wysoce dynamiczny charakter blogosfery istotna jest wiedza dotycząca trwałości takich grup: czy wszyscy członkowie grupy są stali w swych poglądach, czy też można wyodrębnić pewien trzon złożony z kilku blogerów, a reszta migruje między różnymi grupami. Ocenę stabilności szacuje się na podstawie wyliczenia liczby osób, które pozostają członkami danej grupy w dwóch różnych punktach czasowych $t1$ i $t2$:

$$Stability = \frac{|G_{t1} \cap G_{t2}|}{|G_{t1} \cup G_{t2}|}, \tag{2}$$

gdzie G określa zbiór węzłów grupy.

3. Analiza wartości miar węzłów

Traktowanie blogosfery jako graf wydaje się być naturalnym podejściem: blogi umieszczone są w internecie i powiązane między sobą na wiele sposobów. W oparciu o te powiązania można zbudować sieć blogów (ang. *blog network*) oraz sieć postów (ang. *post network*). Można również stworzyć sieć złożoną z obydwu rodzajów powiązań. Sieć blogów powstaje w oparciu o stale dostępne na blogach 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 przedstawiane. Z kolei sieć postów budowana jest na podstawie informacji o autorach komentarzy do postów na danym blogu. Jeśli autorem komentarza jest właściciel jakiegoś bloga, to można poprowadzić z węzła reprezentującego jego blog powiązanie do węzła blogu, na którym wpisał on swój komentarz.

W tak stworzonych sieciach można badać ważność poszczególnych węzłów, zakres ich wpływu na innych blogerów oraz wyodrębniać blogerów mających wspólne zainteresowania i obserwować dynamikę przyłączania się do takiej grupy zewnętrznych blogerów. W naszych badaniach wykorzystaliśmy bardzo przydatną w takich sytuacjach analizę sieci społecznych (ang. *Social Network Analysis – SNA*). Podstawowym zadaniem analizy sieci społecznych jest znajdowanie i interpretowanie wzorców więzi społecznych zachodzących między aktorami, przy czym koncentrujemy się nie na indywidualnych cechach jednostek, ale na własnościach całego układu, na który składają się oddziałujące ze sobą elementy [7].

Koncepcja oceny centralności położenia jednostek w ich sieciach społecznych była jedną z pierwszych, która została zaproponowana i zrealizowana przez analityków sieci społecznych [14]. Miary centralności wskazują na ważność węzła w sieci poprzez badanie jego pozycji w strukturze sieci, która może wskazywać jego ważność, wpływowość lub popularność. Określają one również stopień, do jakiego sieć może rozbudowywać się wokół danego węzła.

Wykorzystuje się wiele sposobów pomiaru centralności [4]. Środek ciężkości (ang. *bary center*) [8], 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 (ang. *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 najkrótszych ścieżek przechodzących przez ten węzeł. Blogi o wysokich wartościach tej miary mogą być traktowa-

ne 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 [5]: stopień koncentracji (ang. *hubness)* i autorytet (ang. *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 (ang. *PageRank*), wykorzystywany przez przeglądarkę internetową Google. W przypadku sieci społecznej parametr ten może być interpretowany jako wyznacznik posiadanej informacji.

W naszych badaniach skupiliśmy się głównie na analizie wybranych blogów dostępnych w portalu www.salon24.pl. Duże zainteresowanie tym portalem wynika prawdopodobnie w dużej mierze z faktu, że właścicielami blogów są osoby ogólnie znane, a ich wpisy odzwierciedlają szybką reakcję na zachodzące w realnym świecie wydarzenia polityczne. Generalnie wpisy są rzetelne i często stają się początkiem ożywionych dyskusji. Dodatkową zaletą portalu jest fakt, że zgodnie z regulaminem „prowadzenie Bloga lub wpisywanie Komentarzy w Salon24.pl jest możliwe wyłącznie po założeniu konta Blogera lub Komentatora"[3], co powoduje że wpisy nie są anonimowe (można zidentyfikować autora).

Przed przystąpieniem do analiz przygotowano środowisko do zbierania danych ze stron blogów, zaprojektowano i zaimplementowano bazę danych pod kątem możliwych do przeprowadzenia analiz oraz aplikację usprawniającą przetwarzanie danych [33].

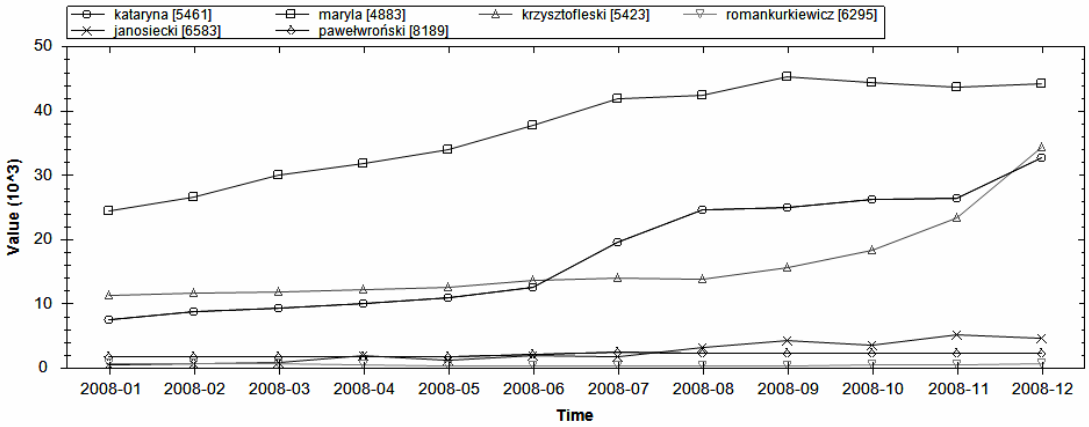
Obecnie w bazie przechowywane są dane z lat 2007÷2009 z 2 325 blogów, postów jest 14 756, komentarzy – 348 084 i 7 086 autorów. Zebrane dane wykorzystano w pierwszym podejściu do analizy statycznej oraz analizy stanu sieci w kolejnych przedziałach czasu, statystyczną analizę otagowania postów, a w ramach dalszych prac analizowano wartość tekstową poszczególnych postów oraz związki pomiędzy postami na różnych blogach, biorąc pod uwagę reprezentację wektorową tekstu oraz czasy umieszczania poszczególnych postów i komentarzy [33].

[3] Regulamin portalu salon24.pl http://www.salon24.pl/information/terms

3.1. Analiza statyczna i dynamiczna

W ramach analizy statycznej badano rankingi blogów w oparciu o wartości poszczególnych miar oraz pozycje rankingowe przez nie zajęte [33]. Analizując w ten sposób trzy zbudowane grafy uzyskano trzy różniące się między sobą rankingi blogów. Uwzględniając z kolei punkty zdobyte za miejsca w tych trzech sieciach przez 30 pierwszych blogów można uzyskać sumaryczny ranking. Zauważyliśmy, że są blogi, które znajdują się zawsze w pierwszej trójce, niezależnie od sposobu definiowania krawędzi w grafie reprezentującym sieć społeczną. Szczegółowy opis zastosowanej metody oraz otrzymane wyniki omówiono w [33]. W ramach tych badań podjęto próbę wyszukiwania znaczących tematów dyskutowanych na blogach poprzez analizę otagowania postów.

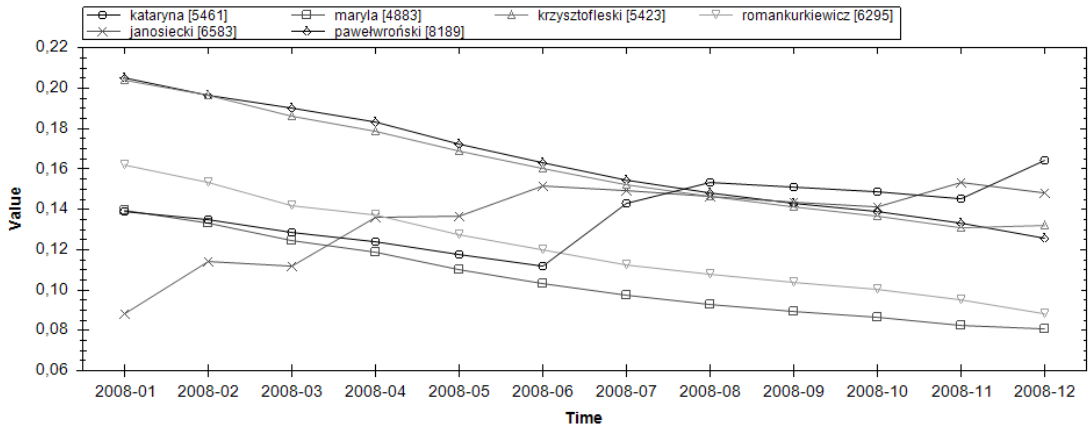
Znacznie istotniejsze informacje ułatwiające zrozumienie zachowania sieci można uzyskać obserwując jak interakcje w sieci zmieniały się w czasie. Następnymi wykonanymi analizami były analizy dynamiczne, pokazujące ewolucje wartości poszczególnych miar oraz pozycji w rankingach poszczególnych blogów w ciągu 2008 roku. Na rysunku 1 pokazano, jak zmieniała się wartość parametru centralności betweeness dla kilku blogów.



Rys. 1. Ewolucja wartości centralności betweeness
Fig. 1. Evolution of betweeness centrality

Widać, że trzy blogi (*kataryny, maryli i krzysztofaleskiego*) mają wysokie wartości tego parametru, świadczące o tym, że znajdują się na wielu najkrótszych ścieżkach wiodących pomiędzy blogami i można je traktować jako brokerów między różnymi podsieciami. Dodatkowo widać, że nastąpił znaczący wzrost miary dla tych blogów: dla *maryli* wzrost ten się zaczął w pierwszym półroczu, a dla dwóch pozostałych – w drugim.

Obserwując z kolei przebieg wartości miary HITS na rysunku 2, widać duży wzrost wartości miar dla blogu *kataryny*, po regularnym spadku trwającym od początku roku do czerwca. Na przestrzeni roku wzrosła także znacząca wartość miary dla blogu *janaosieckiego*. Ponieważ miarę HITS interpretuje się jako siłę powiązań z ważnymi blogami, należy więc uznać, że zarówno *kataryna*, jak i – w mniejszym stopniu – *janosiecki* uzyskały dużą ilość powiązań z innymi wysoko punktowanymi blogami.



Rys. 2. Ewolucja wartości parametru HITS
Fig. 2. Evolution of HITS

Można stwierdzić, że uzyskane rankingi w znacznej mierze potwierdzają wiedzę wynikającą z obserwacji, a dotyczącą popularnych blogerów i ich oddziaływań.

4. Analiza propagacji wpływów

Kolejną dziedziną prowadzonych analiz było badanie reakcji społeczności blogerów na poszczególne posty, a w szczególności posty wcześniej zidentyfikowanych znaczących blogerów. Odbywa się to poprzez analizę podobieństw zawartości poszczególnych postów, sąsiedztwa czasowego oraz porządku wysłania. Analiza zawartości poszczególnych postów oparta była na modelu statystycznym dokumentu – porównywano liczności wystąpień poszczególnych wyrazów w dokumentach w połączeniu z wagą słów wyliczaną przy użyciu metody TF-IDF oraz odległości cosinusowej.

Wyliczanie miary cosinusowej umożliwiło nam badanie podobieństwa poszczególnych postów, a w konsekwencji – wnioskowanie na temat tego, jakie posty inspirowały kolejne posty. W analizach przyjęto, że jeżeli post wzorowany powstał w przeciągu x dni od publikacji postu wzoru oraz autor postu potencjalnie powstałego pod wpływem komentował post wzorcowy, to jest duża szansa, że post wzorowany został zainspirowany komentowanym. Inną niż komentarz przesłanką może być fakt stworzenia hiperpołączenia do blogu z postem wzorcowym przez autora postu wzorowanego.

Tak znalezione posty wzorowane oceniamy i sortujemy na podstawie miary podobieństwa w modelu wektorowym tekstów.

Przeprowadzono szereg testów i udało nam się znaleźć sytuacje, gdy inspiracja danym postem podczas pisania innego była niewątpliwa (sam autor drugiego postu o tym pisał). Jeden z takich wykrytych przypadków przedstawimy poniżej:

Krzysztofleski jest jedną z osób, które miały wysokie wartości parametrów sieci społecznościowych (rys.1, rys. 2). W swoim poście z dnia 2007-10-16 g.15:05 pt. „Mario przegiął na potęgę" komentuje działania Centralnego Biura Antykorupcyjnego. Próba znalezienia podobnych postów autorów powiązanych z *krzysztofleski* dała wyniki, jak widać w tabeli 1, gdzie podanych jest kilka postów wskazanych jako posty, który mogły powstać w odpowiedzi na post rozpatrywany.

Tabela 1

Posty podobne do rozpatrywanego postu *K. Leskiego*

Post	Długość	Ilość powtórzonych słów	Odległość wektorowa	Data
http://pwilkin.salon24.pl/39776,index.html	607	102	0,156368391	2007-10-16 20:31
http://rejent.salon24.pl/42229,index.html	123	17	0,12877953	2007-10-24 11:17
http://eumenes.salon24.pl/39719,index.html	205	36	0,120150302	2007-10-16 17:06
http://krzych.salon24.pl/40273,index.html	209	21	0,088326578	2007-10-18 11:19
http://chupacabras.salon24.pl/39795,index.html	581	78	0,084519913	2007-10-16 21:45

Analizując zawartość tych postów można zauważyć duże podobieństwo wyszukanych postów, opisywane przez duże wartości odległości wektorowej (powyżej 0.1). Trzy z przedstawionych postów zostały napisane w przeciągu kilku godzin od publikacji Leskiego. Najciekawszą sytuację można zauważyć czytając post *eumenes*, który zaczyna się słowami: „Krzysztof Leski się martwi…". W tym przypadku wyraźnie widać zjawisko, którego poszukiwaliśmy: jeden z bloggerów został zainspirowany postem kogoś innego i sam wypowiedział się na ten sam temat. Post ten powstał w przeciągu zaledwie dwóch godzin od powstania notki Leskiego, można więc stwierdzić, iż niektórzy użytkownicy portalu salon24.pl uważnie śledzą blogi swoich „znajomych".

5. Klastry wokół wpływowych blogerów

Przeprowadzono analizy grup/społeczności powstających wśród blogerów na portalu salon24.pl oraz obserwowano ewolucję tych grup. Grupy zawierają autorów wzajemnie czytających i komentujących swoje blogi.

Do zrealizowania przedstawionego celu użyto programu CFinder [4]. Program ten wykorzystuje implementację algorytmu Clique Percolation Method (CPM) [9,10]. Algorytm polega na znalezieniu w grafie *k*-klik. *K*-klika to graf pełny o *k* wierzchołkach, czyli taki, w którym z każdego wierzchołka można bezpośrednio dojść do dowolnego innego. Grupy to

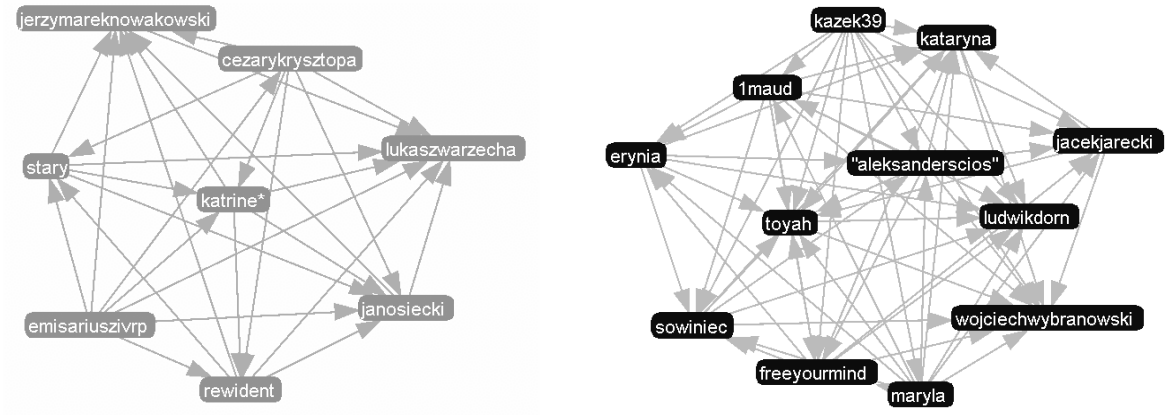
[4] http://cfinder.org/wiki/?n=Main.HomePage

zbiory adjacentnych *k*-klik. Warto zaznaczyć, że z powyższej definicji wynika, że jeden wierzchołek może należeć do wielu społeczności. Program wyszukuje grupy dla wartości *k* od 3 do 7-9 (w zależności od liczby wierzchołków w grafie). Grafy analizowane w poszukiwaniu grup to grafy skierowane. Krawędzie w wykresach należy zatem interpretować tak, że jeżeli od bloggera A prowadzi krawędź do bloggera B, to A częściej komentował B niż na odwrót.

W analizach badaliśmy, jak zmieniały się grupy w czasie, w okresie od 2007 do 2009, podzielonym na półroczne przedziały czasowe. Staraliśmy się wyodrębnić najsilniej powiązane grupy oraz przyjrzeć się ich ewolucji wraz z upływem czasu. Zauważono dużą dynamikę tworzonych grup w kolejnych okresach.

5.1. Analiza grup powstałych wokół wpływowych blogerów

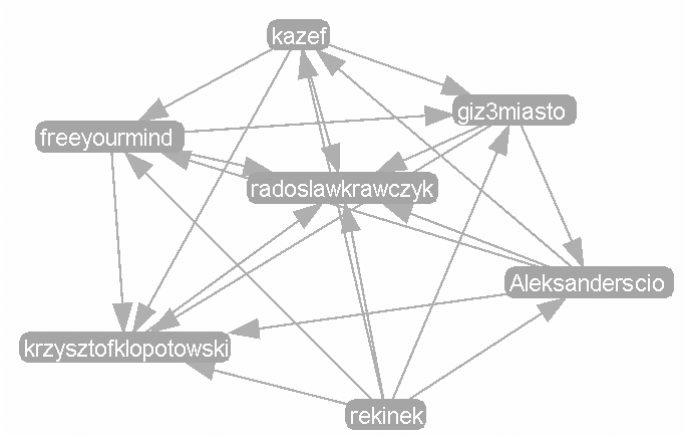
W rozdziale 3 przedstawiono wartości wybranych metryk dla najbardziej wpływowych węzłów. Dominujące węzły w tych rankingach (czyli blogi *kataryny* i *janaosieckiego*) zostały wybrane jako punkty odniesienia do analizy grup, w których skład wchodziły. Do prezentacji wybrano grupy o najmocniejszej strukturze, reprezentujące odpowiednio sytuację w pierwszej i drugiej połowie roku 2008. Pierwsza z tych grup to grupa 8-osobowa, o bardzo silnej strukturze (*k*=8, gęstość=1), w skład której wchodzili m.in. *lukaszwarzecha* oraz *janosiecki.* W grupie tej wyraźnie są zaznaczone węzły głównie komentujące (*emisariuszivrp, rewident*) oraz głównie komentowane (*janosiecki, lukaszwarzecha*). Silną grupą uzyskaną w drugim półroczu jest grupa 12-osobowa (*k*=9, gęstość=0,909). Jest to grupa skupiona wokół takich postaci, jak: *kataryna, ludwikdorn* czy *wojciechwybranowski*.



Rys. 3. Grupy z 2008. Pierwsze (*k*=9, gęstość=1) i drugie półrocze (*k*=9, gęstość=0,909)
Fig. 3. Groups from 2008. First (*k*=9, density=1) and second halfyear (*k*=9, density=0,909)

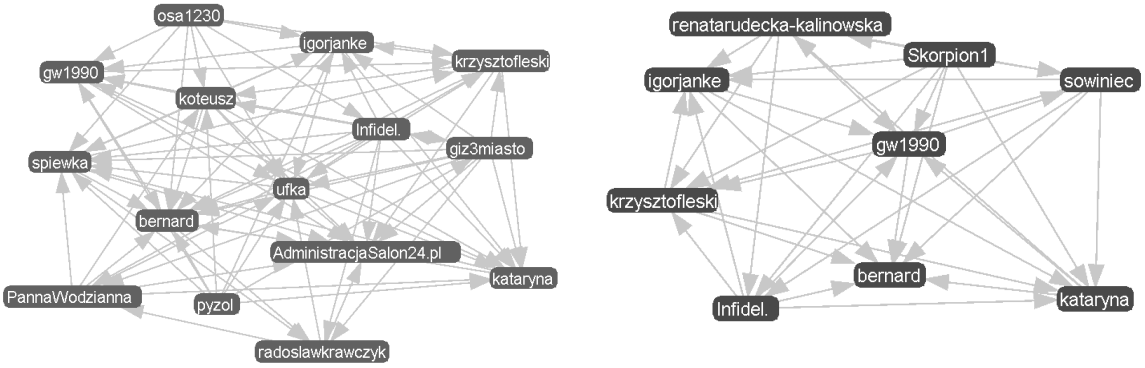
5.2. Analiza ewolucji grupy

Kolejna przykładowa analiza dynamiki grup stworzonych wokół węzłów o znacznych pozycjach w rankingach dotyczy końcowych miesięcy 2009 roku oraz grup stworzonych wokół bloggerów *pannawodzianna* i *sowiniec*. Z przeprowadzonych analiz zostaną zaprezentowane najsilniej związane grupy w październiku, listopadzie i grudniu, w skład których wchodzili rozpatrywani bloggerzy. W październiku oba analizowane węzły znajdowały się w tej samej grupie o *k* równym 7 i gęstości równej 1 (rys. 4).



Rys. 4. Październik 2009. Grupa blogerów *pannawodzianna* i *sowiniec*
Fig. 4. October 2009. Group of bloggers *pannawodzianna* and *sowiniec*

Z kolei w listopadzie wokół analizowanych użytkowników wykształciły się 2 grupy (rys. 5). Gęstość grupy *pannawodzianna* wyniosła 0,79 a grupy *sowiniec* – 0,97.



Rys. 5. Listopad 2009. Grupa *pannawodzianna* oraz grupa *sowiniec*
Fig. 5. November 2009. Group of *pannawodzianna* and group of *sowiniec*

Warto zauważyć, że tacy użytkownicy, jak: *infidel, kataryna, krzysztofleski, gw1990* znajdują się w obydwu grupach. Można by przypuszczać, że mogą oni połączyć obydwu bloggerów. Rzeczywiście, w grudniu obydwaj blogerzy są znów w jednej grupie (7 osób, gęstość: 1). W tej samej grupie znalazł się także użytkownik *gw1990*, który występował z obydwoma użytkownikami, gdy znajdowali się oni w osobnych grupach.

Jak widać, najsilniej związane grupy charakteryzują się dużą zmiennością uczestników. Niemniej jednak dla każdej z rozpatrywanych grup około połowa członków wystąpiła także

w innych grupach. Niektórzy autorzy, jak np. *1maud,do-kasacji* czy *krzysztofleski* byli obecni 3 miesiące z rzędu w grupie z dwoma analizowanymi bloggerami.

Rys. 6. Grudzień 2009. Grupa bloggerów *pannawodzianna* i *sowiniec*
Fig. 6. December 2009. Group of bloggers *pannawodzianna* and *sowiniec*

6. Podsumowanie

W artykule przedstawione zostały różne podejścia umożliwiające identyfikowanie wpływowych jednostek w blogosferze oraz ich wpływu na pozostałych blogerów. Pokazano wyniki oferowane przez te podejścia na przykładzie aktywnie działającej społeczności blogerów, jaką są blogi polityczne na portalu *salon24.pl*. Przedstawione wybrane metody mogą dać dość dokładny naszym zdaniem obraz grup wyodrębniających się w społeczności blogerów oraz ich dynamiki.

BIBLIOGRAFIA

1. Adar E., Zhang L., Adamic L. A., Lukose R. M.: Implicit Structure and the Dynamics of Blogspace. Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference, 2004.
2. Agarwal N., Liu H.: Modeling and Data Mining in Blogosphere, Morgan & Claypool, 2009.
3. Agarwal N., Liu H., Tang L., Yu P. S.: Identifying the Influential Bloggers in a Community, WSDM'08, Palo Alto, USA, 2008.
4. Hanneman R. A., Riddle M.: Introduction to Social Network Methods: University of California Press, 2005.
5. Kleinberg J. M.: Authoritative Sources in a Hyperlinked Environments. Journal of the ACM, 46(5), 1999.

6. McPherson M. L., Smith-Lovin L., Cook J.: Birds of a Feather: Homophily in Social Networks. Annual Review of Sociology, (27), 2001, s. 415÷444.
7. Noy W., Mrvar A., Batagelj V.: Exploratory Social Network Analysis with Pajek. Cambridge University Press, 2005.
8. O'Madadhain J., Fisher D., Smyth P., White S., Boey Y. –B.: Analysis and Visualtization of Network Data Using Jung, Journal of Statistical Software, Vol. VV, No. II.
9. Palla G., Abel D., Derényi I., Farkas I., Pollner P., Vicsek1 T.: K-clique percolation and clustering in directed and weighted networks, (http://www.inma.ucl.ac.be/~blondel/workshops/2008/files/palla.pdf).
10. Palla G., Derényi I., Farkas I., Vicsek T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 2005, s. 814÷818.
11. Wasserman S., Faust K.: Social Network Analysis: Methods and applications. Cambridge University Press, 1994.
12. Xu J., Marshall B., Kaza S., Chen H.: Analyzing and Visualizing Criminal Network Dynamics: A State Study. ISI, 2004, s. 359÷377.
13. Zygmunt A., Koźlak J., Krupczak Ł., Małocha B.: Analiza blogów internetowych przy użyciu metod sieci społecznych. Automatyka: półrocznik Akademii Górniczo-Hutniczej w Krakowie, Vol.13, No. 2, 2009, s. 673÷681.
14. Zygmunt A., Koźlak J.: Sieci społeczne w badaniach kryminalistycznych. Nawarecki E., Dobrowolski G, Kisiel-Dorohinicki M.: Metody sztucznej inteligencji w działaniach na rzecz bezpieczeństwa publicznego, Kraków, Wydawnictwa AGH, 2009, s. 103÷148.

Recenzenci: Dr inż. Paweł Kasprowski
Dr hab. inż. Robert Wrembel

Wpłynęło do Redakcji 31 stycznia 2010 r.

Abstract

In the paper is presented an identification of the influential persons in the blogosphere using different approaches offered by the Social Network Analysis. The first one is the estimation of basic measures of the social network (such as betweenness centrality, hits, Page rank etc.) for the given nodes, observation of their changes in the analysed tome period and an indication of these nodes which have highest values of these measures. The second method is an analysis of the influence of posts, especially posts written by the users with high values of measures, on the blogosphere. It is performed using the analysis of the similarities

of the given posts concerning numbers of words, time neighbouring and sending order. The last method is based on the identification of the sub-societies using algorithms of cliques and strongly connected groups of nodes recognition. The analysis of the groups evolution in the time period for which the evolutions measures had been presented was carried out.

Adresy

Anna ZYGMUNT: Akademia Górniczo-Hutnicza w Krakowie, Katedra Informatyki, Al. Mickiewicza 30, 30-059 Kraków, Polska, azygmunt@agh.edu.pl .

Jarosław KOŹLAK: Akademia Górniczo-Hutnicza w Krakowie, Katedra Informatyki, Al. Mickiewicza 30, 30-059 Kraków, Polska, kozlak@agh.edu.pl .

Łukasz KRUPCZAK: Akademia Górniczo-Hutnicza w Krakowie, Al. Mickiewicza 30, 30-059 Kraków, Polska, krupczak@student.agh.edu.pl .