Symulacja w Badaniach i Rozwoju
Vol. 3, No. 3/2012

Zenon A. SOSNOWSKI, Katarzyna REMBOWICZ
Politechnika Białostocka, Wydział Informatyki
15-351 Białystok, ul. Wiejska 45a
E-mail: z.sosnowski@pb.edu.pl

Badanie sieci społecznych z wykorzystaniem algorytmu grupowania rozmytego

1 Wstęp

W dzisiejszym świecie znacząco wzrosło zainteresowanie sieciami społecznymi [1]. Ma to swoje podłoże głównie w szybkim wzroście liczby użytkowników Internetu, jak i w łatwym dostępie do globalnej komunikacji. Spowodowała je również konieczność posiadania informacji o zjawiskach, które szybko się rozprzestrzeniają w społeczeństwie i których wczesne wykrycie może zaoszczędzić wielu niepotrzebnych kosztów jak np. epidemie, globalny kryzys, terroryzm.

Najogólniej sieć społeczną można zdefiniować jako „dowolny wzór powiązań obecny w społecznych relacjach jednostek, grup i innych zbiorowości" [2]. Badane jednostki (zbiorowości) nazywane są aktorami lub węzłami, a powiązania, jakie tworzą, to relacje lub linki. Wspomnianym aktorem może być struktura tworzona przez ludzi, czyli pojedyncza osoba, grupa osób lub organizacja. Aktorami mogą też być inne układy, np. kraje, miasta, strony WWW, konkretna informacja (badanie jej przepływu), pliki znajdujące się w Internecie, wiadomość e – mail, lub nawet struktura fizycznych urządzeń tworząca sieć (np. urządzenia sieciowe, takie jak routery). Relacje to zależności między aktorami, które mają być zbadane. Mogą to być informacje dotyczące faktu znajomości między ludźmi, rodzaju relacji, jaką są związani, powiązań między organizacjami, poziomu stosunków międzynarodowych, hiperłącz wiążących strony internetowe, fizycznych połączeń pomiędzy urządzeniami.

Badania i analizę sieci społecznych określa się skrótem SNA (*ang. Social Network Analysis*). SNA jest nieodłączną częścią sieci społecznych. Oznacza to, że towarzyszyła już im od dziesiątek lat. Jednak dopiero w ostatniej dekadzie dzięki rozwojowi nauki i wzrostowi mocy obliczeniowej komputerów jest coraz bardziej powszechną i dynamicznie rozwijaną dziedziną nauki [3,4]. SNA to przede wszystkim specyficzna perspektywa analizy: skupienie się nie na aktorach, lecz związkach między nimi. Analiza sieci społecznych znajduje zastosowanie nie tylko w socjologii, ale również w takich branżach jak ekonomia, biologia, geografia, zdrowie publiczne, bezpieczeństwo narodowe (np. walka z terroryzmem, zorganizowanymi grupami przestępczymi) czy psychologia.

Głównym zadaniem SNA jest odnajdywanie wzorców, nieznanych połączeń oraz ciekawych zjawisk w sieciach społecznych. Odnalezione wzorce lub zjawiska służą do przewidywania zachowań aktorów, którzy wchodzą w skład danej sieci. Takie przewidywania czyni się na podstawie wcześniejszych badań innych sieci podobnego typu. Jednak za każdym razem trzeba podejść do analizowanej sieć w sposób indywidualny. Badane atrybuty, rodzaje relacji oraz różne oddziaływanie zjawisk

społecznych na te relacje mogą mieć inny wpływ na podobną sieć i jej analiza może przynieść odmienne wyniki.

Próba rozwiązania problemu wyszukiwania klastrów (grup) w danych sieci społecznej była głównym celem poniższej pracy. Przeprowadzone badania odbyły się z zastosowaniem własności zbiorów rozmytych, które nie były stosowane w żadnym oprogramowaniu do analizy sieci społecznych, pomimo że są bardziej intuicyjne i przyswajalne, gdy opisują rzeczywistość niż zbiory klasyczne. Zbiory rozmyte pozwalają na opisanie rzeczywistości w bardzo naturalny sposób, posługując się przedziałami wartości, a nie pojedynczymi wartościami. Zastosowanie tej własności do grupowania danych oznacza, że jeden węzeł może być zaliczony do dwóch różnych klastrów, co znajduje swoje odzwierciedlenie w rzeczywistych sieciach. Jeden aktor może tak samo „mocno" przynależeć do dwóch podsieci. Jedną z pierwszych prób zastosowania tej techniki podjęto w [5].

2 Przedstawienie algorytmu grupowania rozmytego danych sieci społecznych

Analiza sieci społecznych (SNA) przez wykorzystanie teorii grafów może być realizowana za pomocą algorytmów, które są dedykowane grafom i strukturom sieciowym. Do najpopularniejszych algorytmów tej kategorii, które poddają sieci społeczne grupowaniu, należą: algorytm klastrowania Marcov (MCL) użyty w programie UNICET oraz trzy algorytmy, Girvan–Newman, Wakita – Tsurumi, Clauset – Newman – Moore (CNM), użyte w programie NodeXL [6].

Na potrzeby tej pracy został wykorzystany algorytm grupowania rozmytego.

Algorytm ten jest metodą grupowania danych pozwalającą jednostce danych przynależeć do dwóch lub więcej klastrów. W swoim rozwiązaniu algorytm stosuje logikę zbiorów rozmytych, w której przynależność danej do zbioru zależy od wielkości funkcji przynależności. Ta funkcja osiąga wartości z przedziału [0,1], gdzie 0 oznacza całkowity brak przynależności, a 1 oznacza całkowitą przynależność do zbioru. Algorytm działa podobnie jak algorytm k – średnich. Też należy przeliczać centra klastrów oraz odległości między nimi a grupowanymi punktami, z tą różnicą, że przeliczoną odległość trzeba przemnożyć przez konkretną wartość z macierzy przynależności U. FCM również potrzebuje z góry określonej liczby klastrów [7].

Ogólna idea algorytmu polega na minimalizacji następującej funkcji celu J:

$$J = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2, \quad \text{gdzie } 1<m<\infty ;$$

m jest to tak zwany współczynnik rozmytości, *N* to liczba punktów w zbiorze, który jest grupowany, C to liczba centrów klastrów, u_{ij}^{m} – stopień przynależności x_i do *j*-tego klastra (jest to element (liczba) z macierzy przynależności U), x_i to *i*-ty element zbioru danych, c_j to centrum j-tego klastra, natomiast || || to wybrana metryka do obliczania odległości. Macierz przynależności U zawiera wartości ze zbioru [0,1] i określa stopień przynależności badanego elementu do danego centrum.

W pierwszym kroku algorytmu należy zainicjować wartości w macierzy U. Następnie na jej podstawie można obliczyć położenia centrów, stosując się do następującego wzoru:

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m}$$

Kolejny krok to przeliczenie wartości w macierzy U, stosując się do nowych centrów, według wzoru:

$$u_{ij} - \frac{1}{\sum_{k-1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

Po wyliczeniu nowej macierzy U sprawdzić trzeba warunek stopu algorytmu, który przedstawia się następująco:

$$max_{ij} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{k} \right| \right\} < \varepsilon$$

gdzie ε to wartość z przedziału [0,1], a k to numer iteracji. Im mniejsza jest wartość ε, tym więcej iteracji wykona algorytm. Jeżeli warunek stopu jest spełniony, to algorytm kończy swoje działanie, dostarczając jako wynik centra klastrów C oraz macierz przynależności U, która określa, w jakim stopniu dany obiekt należy do poszczególnych klastrów([7]).

3 Eksperymenty

Przedstawione zostaną przebieg i wyniki (Tab. 1) testów wykonane przy użyciu aplikacji wykonanej w ramach pracy dyplomowej [8]. Badania przeprowadzone zostały na dwóch zbiorach danych. Analizowane dane są rzeczywistymi danymi i opisują sieci stworzone za pośrednictwem Internetu. Oba zbiory danych zostały pozyskane za pośrednictwem oprogramowania NodeXL ([6]). Ich oryginalna forma to lista krawędzi, której wiersze opisują relacje dla dwóch wierzchołków oraz wagę tej relacji.

Pierwszy zbiór danych opisuje sieć stworzoną przez plik wideo z serwisu społecznościowego YouTube. NodeXL wyszukał wszystkie pliki dla wyrażenia „ladygaga", które jest jednym z najczęściej wyszukiwanych wyrażeń w wyszukiwarkach internetowych. Relacje między tymi plikami są zaznaczone krawędziami tylko wtedy, gdy dwa pliki mają taką samą nazwę lub taką samą nazwę autora, lub były kiedykolwiek skomentowane przez tego samego użytkownika serwisu. Waga krawędzi między dwoma plikami to suma wartości ilości wymienionych rodzajów relacji. Otrzymana sieć posiada 3526 krawędzi.

Algorytm klastrujący dostępny w NodeXL - Wakita – Tsurumi podzielił dane na cztery klastry. W danych wejściowych dla algorytmów fuzzy c- means podawana jest wartość cztery, jako liczba klastrów w tym zbiorze danych.

Drugi zbiór danych jest to sieć stworzona przez użytkowników poczty elektronicznej, którzy wymieniają się wiadomościami e - mail. Dane są uzyskane z lokalnego komputera, na którym był zainstalowany program NodeXL. NodeXL pozwala za pośrednictwem programów do obsługi skrzynki pocztowej, tj. Outlook lub Windows Mail, stworzyć sieć wszystkich odbiorców i nadawców wiadomości e – mail. Węzłami w tak stworzonej sieci są wszystkie kontakty (adresy pocztowe e - mail). W badanej sieci jest 427 węzłów i 669 krawędzi. Krawędzie łączą tylko tych aktorów, którzy kiedykolwiek mieli ze sobą kontakt za pośrednictwem wiadomości e - mail. Wagą

odnalezionych krawędzi jest suma ilości wysłanych i odebranych wiadomości między aktorami. Liczba klastrów wykryta przez algorytmy w NodeXL to dwanaście. Dlatego też podczas podziału tego zbioru na klastry algorytm fuzzy c–means wyszukuje dwunaście zbiorów.

Analiza wyników algorytmu grupowania fuzzy c–means miała sprawdzić, czy algorytm rozmyty fuzzy c – means nadaje się do klastrowania węzłów sieci społecznych. Badania były przeprowadzone dla czterech różnych metryk: Euklidesa, miejskiej, Minkowskiego oraz współczynnika korelacji. Badania były prowadzone z podziałem ze względu na dostarczone zbiory danych. Porównanie otrzymanych wyników znajduje się w tabeli 1.

Tab. 1. Zestawienie procentowej zgodności grupowania

Tab. 1. Comparison of the results of the classification

Metryka	I zbiór danych (sieć plików wideo)	II zbiór danych (sieć e - mail)
	% zgodności	% zgodności
Euklidesowa	100 %	9%
Miejska	96%	57%
Minkowskiego	99%	1%
Współ. korelacji	51%	18%

Analizując otrzymane wyniki dla pierwszego zbioru danych, można stwierdzić, że są one zadowalające. W przypadku metryki euklidesowej, miejskiej i Minkowskiego grupowanie węzłów było prawie w stu procentach zgodne. Rozbieżności pojawiły się w ostatniej metryce współczynnika korelacji. Patrząc na trzy pierwsze metryki, dla pierwszego zbioru można śmiało stwierdzić, że rozmyty algorytm fuzzy c–means nadaje się bardzo dobrze do grupowania węzłów sieci społecznych. Algorytm fuzzy c–means z użyciem współczynnika korelacji nie dał zadowalających efektów na tle pozostałych metryk.

Wyniki dla drugiego zbioru danych w żadnej metryce nie dały stuprocentowej zgodności. Najlepszy wynik pod względem zgodności osiągnięto dla metryki miejskiej, gdzie grupowanie zgadzało się w 57%. Kolejne miejsce zajęła metryka współczynnika korelacji z wynikiem 18%, później metryka euklidesowa (9%) i na końcu odległość Minkowskiego – z wynikiem tylko 1%. Rozbieżność wyników jest zapewne spowodowana specyfiką zbioru danych, który był analizowany. Jak już wcześniej wspomniano, był to zbiór, który charakteryzował się niską spójnością i był dzielony aż na dwanaście klastrów. Taka sytuacja sprawiła, że algorytm rozmyty inaczej klasyfikował wybrane węzły. Grupowanie algorytmu rozmytego rozpoznawało wiele węzłów granicznych, np. ponad połowę aktorów zaliczało do dwóch lub jednego klastra jednocześnie tworząc grupy mocno na siebie zachodzące.

Podsumowując, rozmyty algorytm grupowania fuzzy c–means dla spójnych sieci społecznych (takich jak pierwszy zbiór danych) daje zadowalające wyniki w metrykach: euklidesowej, Minkowskiego i miejskiej, natomiast dla sieci mało spójnych, z podziałem na wiele podgrup, rozmyty algorytm nie sprawdza się dobrze w żadnej

metryce. Wynika to z faktu, że przy podziale sieci na wiele klastrów fuzzy c–means zbyt wiele węzłów zalicza do granicznych, tworząc mocno zachodzące na siebie społeczności. Wykorzystanie rozmytego algorytmu fuzzy c–means w analizie spójnych sieci społecznych może mieć swoje szczególne przeznaczenie, to znaczy wykrywać nieostre granice między klastrami, co wiąże się ze znajdowaniem aktorów, którzy są kluczowymi węzłami (mostami) w kontaktach między podgrupami sieci.

4 Wnioski końcowe

Wyniki otrzymanych badań pokazują, że wartości obliczone przez fuzzy c – means są bardzo zbliżone do wartości algorytmu k – średnich, jak i do wartości wyznaczonych przez program NodeXL. Śmiało można więc stwierdzić, że w kwestii klastrowania danych sieci społecznych z pewnymi zależnościami własności zbiorów rozmytych spełniają się bardzo dobrze. Tymi zależnościami są: spójność badanej sieci, liczba klastrów oraz metryka użyta do implementacji algorytmu. Klastrowanie sieci spójnych z niedużą liczbą grup daje zadowalające wyniki dla metryk: Euklidesa, Minkowskiego i miejskiej. Natomiast grupowanie niespójnych sieci, bez względu na użytą metrykę, nie daje zadowalających efektów. Nie zmienia to jednak faktu, że pozostałe miary sieci społecznych ([9]) nie znalazły do tej pory zastosowania zbiorów rozmytych. Ta kwestia pozostaje do dalszych badań.

Literatura

1. Easley D, Kleinberg J.: *Networks, Crowds, and Markets: Reasoning about a Highly Connected World,* Cambridge University Press, Cambridge, 2010
2. Kuper A., Kuper J.: *The Social Science Encyclopedia*, Routledge, New York City, 2004
3. Scott J.: *Social network analysis a handbook*, Sage Publications, London, 1987
4. Marin A., Wellman B.: *Social Network Analysis: An Introduction*, University of Toronto, Toronto, 2009
5. Zhang S., Wang R.S., Zhang X.: Identification of overlapping community structure in complex networks using fuzzy c-means clustering, *Physica A: Statistical Mechanics and its Applications*, vol. 374, Issue 1, 15 January 2007, pp. 483–490
6. Microsoft Code Plex Open source community, *NodeXL*, Witryna internetowa, http://www.codeplex.com, stan z 03.06.2011
7. Bezdek J., *Pattern Recognition with Fuzzy Objective Function Algoritms*, Plenum Press, New York, 1981
8. Rembowicz K.: *Zbiory rozmyte w analizie sieci społecznych*, praca magisterska, Politechnika Białostocka, Wydz. Informatyki, 2011
9. Hanneman R., Riddle M.: *Introduction to social network methods*, University of California, Riverside, CA, 2005

Streszczenie

W pracy przedstawiono propozycje, uzyskane wyniki oraz wypływające z nich wnioski dotyczące zastosowania teorii zbiorów rozmytych do analizy sieci społecznych. Wyniki symulacji pokazują, że proponowane podejście wykorzystujące własności zbiorów

rozmytych sprawdza się bardzo dobrze w analizie spójnych sieci społecznych z niedużą liczbą klastrów.

Słowa kluczowe: sieci społeczne, grupowanie rozmyte

Social Networks Analysis using Fuzzy Clustering Algorithm

Summary

The paper presents proposals, the obtained results and the resulting conclusions concerning the use of fuzzy set theory to the analysis of social networks. The simulation results show that the proposed approach using fuzzy property works very well in the analysis of social networks consistent with a small number of clusters.

Keywords: Social Networks, Fuzzy C-Means

Praca finansowana w ramach badań statutowych Wydziału Informatyki Politechniki Białostockiej nr S/WI/2/08.