Paweł Szyman[1]

Wybrane aspekty związane z analizą sieci społecznościowej opartej na korespondencji e-mail instytucji publicznej

1. Wstęp

Współcześnie Internet stał się jednym z głównych kanałów komunikacji w społeczeństwie. Ludzie wymieniają się informacjami, korzystając z różnych ogólnodostępnych narzędzi. W zależności od potrzeb, wybór narzędzia do komunikacji może być inny. Jeżeli komuś zależy na szybkim kontakcie lub nawiązaniu relacji krótkotrwałej, skorzysta z jednego z komunikatorów, takich jak Messenger lub SnapChat. Jeżeli zamiarem jest nawiązanie kontaktu z większą liczbą osób, można przypuszczać, że do tego typu komunikacji zostanie wykorzystany portal społecznościowy. Istnieją jednak relacje, w których korzystanie z wyżej wymienionych sposobów komunikacji często nie jest pożądane, np. do kontaktów formalnych przeważnie wykorzystuje się pocztę elektroniczną.

Podstawowym powodem, dla którego firmy korzystają z poczty elektronicznej, jest wymiana korespondencji pomiędzy osobami – pracownikami firmy oraz kontrahentami. Dodatkowo, na serwerach poczty przechowywana jest historia korespondencji, co oznacza, że w każdej chwili można wrócić do danej wiadomości.

Z drugiej strony, stały rozwój technologii oraz narzędzi zapewnia naukowcom odpowiednie środowisko i zasoby do analizy społecznych interakcji osób, które komunikują się, wykorzystując pocztę elektroniczną.

W niniejszej pracy skupiono się na sieci społecznościowej opartej na komunikacji e-mail w instytucji publicznej. Celem pracy było dokonanie analizy takiej sieci, w szczególności określenie popularności i wpływu określonego węzła, będącego częścią takiej sieci.

W pierwszej części pracy przedstawiono podstawowe podejścia związane z analizą sieci społecznościowych, odnosząc je do sieci opartej na korespondencji

[1] Uniwersytet Morski w Gdyni, Wydział Przedsiębiorczości i Towaroznawstwa, Katedra Systemów Informacyjnych.

e-mail w instytucji publicznej, jak również dokonano przeglądu literatury związanej z analizą sieci społecznościowych opartych na takiej korespondencji. W drugiej części pracy zaprezentowano wyniki eksperymentu obliczeniowego, przeprowadzonego z wykorzystaniem metodologii analizy sieci społecznościowych, w którym określono podstawowe cechy takiej sieci, zbadano powiązania między jednostkami, jak również dokonano wizualizacji struktury organizacyjnej instytucji.

2. Analiza sieci społecznościowych

Analiza sieci społecznościowych[2] (ang. *Social Network Analysis*, SNA) opiera się na badaniu struktury, powiązań i zachowania określonych jednostek wewnątrz grup społecznych, reprezentowanych w postaci wierzchołków (odnoszących się przykładowo do osób lub też organizacji) oraz krawędzi (określających wzajemne powiązania lub przepływ informacji między tymi jednostkami). Metodologia analizy sieci społecznościowych wymaga określenia obszaru badanej sieci[3], który identyfikuje jednostki wchodzące w skład sieci i relacje między nimi. Są to głównie wszyscy pracownicy organizacji lub określona grupa. Dodatkowo udostępnia wiele miar, dzięki którym istnieje możliwość prowadzenia analiz właściwości danej sieci. Dzięki analizie sieci społecznościowych można określić m.in.: pozycje wybranych jednostek w danej sieci, ich role w organizacji czy też odkryć pewne wzorce w relacjach pomiędzy jednostkami reprezentowanymi w sieci.

Sieć ukazującą komunikację elektroniczną można rozumieć jako pewien rodzaj sieci społecznej, w której wierzchołki odpowiadają osobom, a krawędzie łączące wierzchołki reprezentują kontakty między ludźmi. Przykładowo połączenie wierzchołków może zostać utworzone w sytuacji, gdy przynajmniej dwie osoby wymienią się między sobą wiadomością *n* razy. Oczywiście nie jest to jedyny sposób określania krawędzi. Można przyjąć inne kryteria, które pozwolą na bardziej szczegółową analizę danej sieci.

---

[2] S. Wasserman, K. Faust, *Social Network Analysis: Methods and Applications*, Cambridge University Press, Cambridge, New York 1994, s. 2–3.

[3] M. Zdziarski, *Analiza sieci*, w: *Sieci międzyorganizacyjne. Współczesne wyzwanie dla teorii i praktyki zarządzania*, J. Niemczyk, E. Stańczyk-Hugiet, B. Jasiński (red.), Warszawa 2012, t. 1, s. 35–42.

W zależności od analizowanego zbioru danych poczty elektronicznej struktura sieci może przyjąć różne formy. Gdy analizie zostanie poddana sieć osoby, która komunikuje się z kilkoma osobami, a osoby te między sobą nie wymieniają wiadomości, taka sieć może przyjąć kształt gwiazdy. Natomiast jeżeli analiza dotyczyć będzie całej historii korespondencyjnej dużej firmy korporacyjnej lub dużej jednostki organizacji publicznej, sieć ta będzie zdecydowanie bardziej rozbudowana. Jednym z aspektów analizy sieci społecznej opartej na komunikacji za pomocą poczty elektronicznej jest badanie przepływu informacji w firmie. Analiza tych sieci pozwala również na zidentyfikowanie specyficznych grup docelowych, do których można zaplanować wysyłkę e-maili grupowych z konkretną informacją. Taki zabieg pozwoli uniknąć niepotrzebnych e-maili, zwanych spamem.

3. Metodyka badań nad sieciami społecznościowymi

Badanie sieci komunikacyjnej e-mail pozwala na pozyskanie kluczowych informacji. Służy głównie do filtrowania wiadomości na podstawie priorytetu przypisywania e-maili oraz do identyfikacji spamu. Można również dowiedzieć się, kto w danej sieci jest najwyżej w hierarchii, czy w firmie bądź organizacji nie dochodzi do łamania prawa poprzez udostępnianie ważnych informacji osobom trzecim. Warto zaznaczyć, że wysyłane wiadomości pomiędzy uczestnikami danej sieci mogą mieć charakter formalny, w przypadku np. komunikacji prezesa z pracownikiem, lub nieformalny, w przypadku zwykłej relacji koleżeńskiej. Badanie tych relacji wyodrębnionych z archiwum bądź logów serwerowych poczty e-mail stanowi duże wyzwanie dla naukowców.

Odnosząc się do powyższych zadań i możliwości eksploracji sieci opartej na komunikacji e-mail, poniżej przytoczono wybrane podejścia i prace osób, które podjęły się rozwiązania wybranych problemów i znalezienia odpowiedzi na kilka ważnych pytań.

P.A. Gloor[4] opisuje zastosowanie linku tymczasowego i analizy zawartości w danych firmy Enron. Pozwala mu to identyfikować głównych uczestników sieci oraz wygenerować mapy klastrowe treści e-mailowych. Dodatkowo, w łatwy

[4] P.A. Gloor, *Capturing Team Dynamics through Temporal Social Surfaces*, w: *Information Visualization*, E. Banissi, M. Sarfraz, J.C. Roberts, B. Loften, A. Ursyn, R.A. Burkhard, A. Lee, G. Andrienko (red.), 2005, s. 939–944.

sposób może zidentyfikować potencjalne wzorce podejrzanych aktorów, których działania szkodzą firmie.

A. McCallum, X. Wang oraz A. Corrada-Emmanuel[5] zaprezentowali model ART (Author-Recipient-Topic), czyli autor – odbiorca – temat. Model ten ma za zadanie uczenie się dystrybucji tematów na podstawie wysyłanych komunikatów kierunkowych pomiędzy jednostkami. Model opiera się na algorytmie LDA (Latent Dirichlet Allocation) oraz AT (Author-Topic). W dalszej części swojej pracy autorzy zaprezentowali rozszerzenie modelu RART, czyli rola – autor – odbiorca – temat.

X. Zhang, J. Zhu, Q. Wang oraz H. Zhao[6] zaproponowali nową metodę identyfikacji wpływowych węzłów w złożonych sieciach o strukturze społeczności. Ta metoda wykorzystuje prawdopodobieństwo transferu informacji między dowolną parą węzłów a algorytmem *k-medoid clustering*.

U. Boryczka, B. Probierz oraz J. Kozak[7] w swojej pracy zaproponowali nowe podejście do automatycznej kategoryzacji wiadomości e-mail na podstawie algorytmu mrówkowego. Dodatkowo zastosowali rozwiązania z eksploracji danych oraz SNA. Swój algorytm również testowali na danych e-mail Enron.

R. Bekkerman, A. McCallum oraz G. Huang[8] podjęli się badania porównawczego kategoryzowania wiadomości e-mail na podstawie zbioru e-mail Enron oraz zbioru uczestników projektu badawczego SRI. W tym badaniu wykorzystali kilka popularnych klasyfikatorów, między innymi maksymalnych entropi (MaxEnt), Naive Bayes, SVM (Support Vector Machine). Ostatni wariant okazał się bardzo efektywny pod kątem obliczeniowym oraz łatwy do wdrożenia.

---

5 A. McCallum, X. Wang, A. Corrada-Emmanuel, *Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email*, „Journal of Artificial Intelligence Research" 2007, vol. 30, s. 249–272.

6 X. Zhang, J. Zhu, Q. Wang, H. Zhao, *Identifying Influential Nodes in Complex Networks with Community Structure*, „Knowledge-Based Systems" 2013, vol. 42, s. 74–84.

7 U. Boryczka, B. Probierz, J. Kozak, *An Ant Colony Optimization Algorithm for an Automatic Categorization of Emails*, Springer, LNCS 8733 w: *Computational Collective Intelligence. Technologies and Applications*, D. Hwang, J.J. Janson, N.T. Nguyen (red.), 2014, s. 583–592.

8 R. Bekkerman, A. McCallum, G. Huang, *Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora*, Computer Science Department Faculty Publication Series 218, 2004.

4. Eksperyment obliczeniowy

W celu wykazania podstawowych cech sieci opartej na korespondencji e-mail pomiędzy pracownikami instytucji publicznej, a także dokonania analizy takiej sieci przeprowadzono eksperyment obliczeniowy. Celem eksperymentu było zbadanie, jak kształtują się relacje pomiędzy pracownikami wewnątrz organizacji, a także zbadanie sieci pod kątem ważności wierzchołków w sieci oraz określenie, który z nich odgrywa kluczową rolę.

Badania zaprezentowane w artykule zostały przeprowadzone na danych pozyskanych z logów serwerowych poczty elektronicznej środowiska akademickiego w instytucji publicznej. Skupiono się na dwóch wybranych działach organizacji, które posiadają zbliżoną liczbę osób. W celu przeprowadzenia eksperymentu pobrano logi serwerowe instytucji publicznej z całego roku, począwszy od kwietnia 2017 r. do kwietnia 2018 r. Ze względu na bardzo dużą ilość informacji zawartych w logach wybrano okres jednego miesiąca – marzec 2018. Przed przystąpieniem do badań istotne było oczyszczenie danych serwerowych z niepotrzebnych informacji i wydobycie tych najbardziej istotnych. W trakcie oczyszczania usunięto wszelkie duplikaty wiadomości. Następnie pobrano adresy e-mail wszystkich pracowników wybranych wcześniej działów i przefiltrowano dane tak, aby uzyskać informacje o wysłanych i odebranych wiadomościach e-mail. Z uwagi na wrażliwość danych, każdemu adresowi e-mail została przypisana kolejna liczba naturalna, zaczynając od 1.

W tabeli 1 zestawiono liczbę wiadomości e-mail przed i po oczyszczeniu.

Tabela 1. Dane wykorzystane w eksperymencie

Kategoria wiadomości	Liczba
Wszystkie wiadomości e-mail (przed oczyszczeniem)	4 503 376
Wszystkie wiadomości e-mail wewnątrz instytucji	2 945 258
Wszystkie wiadomości wewnątrz instytucji w marcu 2018 r.	274 458
Wszystkie wiadomości na wybranej jednostce w marcu 2018 r.	62 596
Wiadomości wysłane i odebrane w obrębie działu nr 1 w marcu 2018 r.	241
Wiadomości wysłane i odebrane w obrębie działu nr 2 w marcu 2018 r.	127

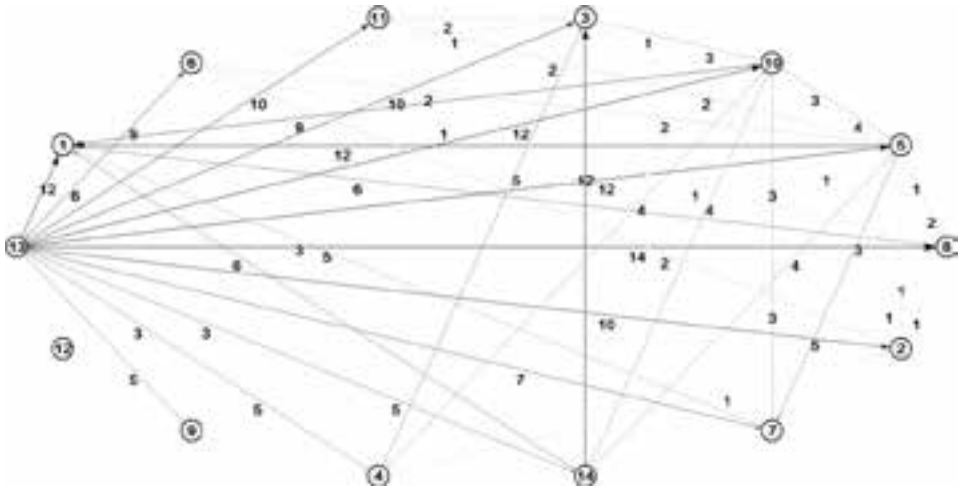
Źródło: opracowanie własne.

Kolejnym etapem badań było dokonanie konwersji danych do formatu akceptowanego przez program do analizy sieci społecznościowych. Na potrzeby

eksperymentu wykorzystano program Pajek[9]. Jest to program do graficznej reprezentacji i analizy dużych sieci.

Wykorzystując dane i narzędzie Pajek, utworzono dwie podsieci dotyczące komunikacji opartej na poczcie elektronicznej. Pierwsza z nich składa się z 14 wierzchołków, które reprezentują adresy e-mail pracowników. Pomiędzy wierzchołkami istnieją połączenia, informujące o zaistniałej relacji, czyli wymianie wiadomości. Druga podsieć składa się z 12 wierzchołków i podobnie zdefiniowanych relacji.

Na rysunku 1 przedstawiono wizualizację, której rezultatem jest graf skierowany, pokazujący połączenia pomiędzy pracownikami działu nr 1, a także liczbę wysłanych lub otrzymanych wiadomości e-mail w obserwowanym okresie. Liczba umiejscowiona bliżej grotu strzałki wskazuje na liczbę wysłanych wiadomości.



Rysunek 1. Wizualizacja pierwszej podsieci

Źródło: opracowanie własne.

W przypadku pierwszej podsieci widać, że najbardziej wpływowym wierzchołkiem jest „13". To właśnie z tego wierzchołka wychodzi najwięcej połączeń. Z drugiej strony nie trudno zauważyć, że osoba przypisana jako „12" nie wysłała ani nie odebrała żadnego e-maila w badanym okresie. Natomiast wierzchołek „9" odebrał tylko 5 wiadomości, nie wysyłając ani jednej. Można przypuszczać, że wierzchołek „13" to sekretariat podsieci.

[9] V. Batagelj, A. Mrvar, *Pajek – Program for Large Network Analysis*, University of Ljubljana, Ljubljana 1997.

W celu sprawdzenia, który z wierzchołków jest najbardziej istotny, zbadano podstawowe miary centralności analizy sieci społecznościowych, którymi są stopień wierzchołka wejściowego i wyjściowego, bliskość oraz pośrednictwo. Na podstawie tych miar można określić popularność i wpływowość danego węzła w sieci, mowa tu o stopniu wierzchołka. Bliskość[10] w sieci społecznej opartej na komunikacji z wykorzystaniem poczty elektronicznej może być rozumiana jako czas, jak szybko dana osoba może skomunikować się z pozostałymi osobami w sieci. Pośrednictwo[11] natomiast określa, jakie jest prawdopodobieństwo, że dana osoba jest kluczowa dla przepływu informacji między dowolnymi dwoma innymi osobami. Wskazuje, jak wiele najkrótszych dróg stracimy, gdy usuniemy węzeł z sieci. Innymi słowy, aby skutecznie zakłócić działanie sieci, powinniśmy uszkodzić te węzły, których pośrednictwo jest największe.

Tabela 2. Miary centralności pierwszej podsieci

Wierzchołki	Stopień centralności – wejściowy	Stopień centralności – wyjściowy	Pośrednictwo	Bliskość
1	6	0	0,0000	0,6191
2	5	0	0,0000	0,5865
3	6	3	0,0310	0,6555
4	1	4	0,0000	0,5571
5	6	7	0,0737	**0,7429**
6	3	5	0,0214	0,5865
7	3	4	0,0085	0,5865
8	5	6	0,1165	0,6555
9	1	0	0,0000	0,4845
10	6	4	0,0085	0,6555
11	3	0	0,0000	0,5301
12	0	0	0,0000	0,0000
13	4	12	**0,1934**	**0,9286**
14	1	5	0,0000	0,5865

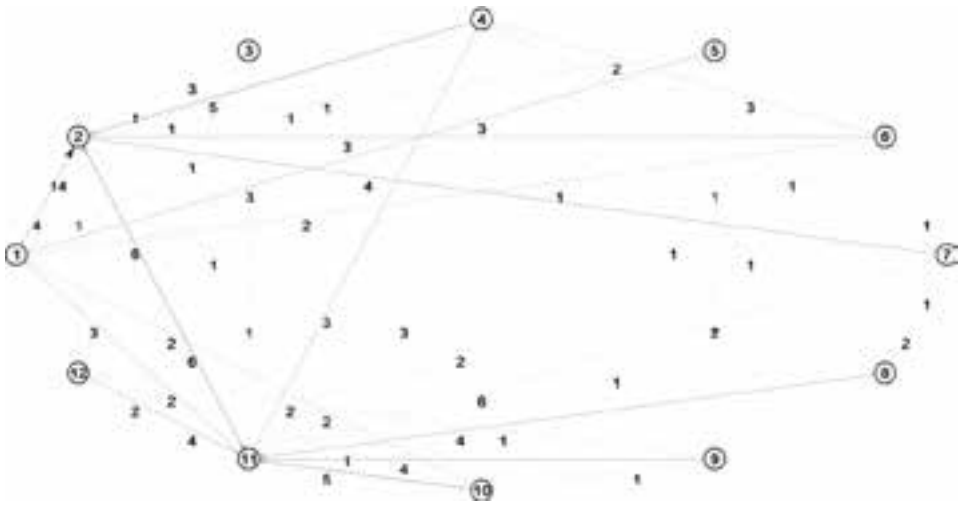
Źródło: opracowanie własne.

[10] R. Rousseau, E. Otte, *Social Network Analysis: A Powerful Strategy, also for the Information Sciences*, „Journal of Information Science" 2002, 28, s. 442–444.

[11] D. Vargas, A. Bridgeman, D. Schmidt, P. Kohl, B. Wilcox, L. Carr, *Correlation Between Student Collaboration Network Centrality and Academic Performance*, Carr Department of Physics, Colorado School of Mines, Golden, CO 80401, USA, August 2, 2018, s. 6.

Po analizie danych zebranych w tabeli 2 można powiedzieć, że są podstawy do stwierdzenia, że wierzchołek „13" jest najbardziej wpływowy w tej sieci. Wartość miary bliskości na poziomie 0,9286 świadczy o tym, że osoba identyfikowana jako wierzchołek „13" kontaktuje się z prawie wszystkimi osobami w podsieci. Dużą wartość bliskości posiada również wierzchołek „5". Pośrednictwo na poziomie 0,1934 dla wierzchołka „13" mówi, że jeżeli zostanie usunięty, to spowoduje zakłócenia sieci w postaci zerwania komunikacji pomiędzy innymi wierzchołkami. Stopień centralności wyjściowy wskazuje na to, że osoba identyfikowana jako wierzchołek „13" może rozpowszechniać informację masowo, poprzez wysyłanie jednej wiadomości e-mail do wielu osób.

Przyglądając się drugiej podsieci, odnoszącej się do działu nr 2 (rysunek 2), można stwierdzić, że jej struktura jest podobna do pierwszej. W porównaniu z pierwszą podsiecią jest ona mniejsza o 2 wierzchołki, a liczba wysłanych i odebranych wiadomości różni się o połowę.



Rysunek 2. Wizualizacja drugiej podsieci

Źródło: opracowanie własne.

W podsieci drugiej można zauważyć, że wszystkie wierzchołki wykazują aktywność. Najbardziej wpływowym wierzchołkiem jest „11".

Osoba identyfikowana jako wierzchołek „11" komunikuje się z każdą osobą w tej sieci. Mówi o tym miara bliskości wierzchołków, która jest równa 1, czyli wartości maksymalnej. W tej podsieci bliskość wierzchołków jest wysoka, co świadczy o tym, że osoby komunikują się z większością pracowników w podsieci. W celu

skutecznego zakłócenia działania podsieci drugiej należałby usunąć wierzchołek „11", którego miara pośrednictwa jest równa 0,4174. Zbliżoną aktywność do „11" wykazuje wierzchołek „2", który może pełnić podobną funkcję w tej podsieci.

Tabela 3. Miary centralności drugiej podsieci

Wierzchołki	Stopień centralności – wejściowy	Stopień centralności – wyjściowy	Pośrednictwo	Bliskość
1	5	6	0,1379	0,7333
2	9	3	0,0674	**0,8461**
3	3	2	0,0000	0,6111
4	2	5	0,0303	0,6875
5	3	4	0,0409	0,6111
6	2	6	0,0462	0,6875
7	4	3	0,0280	0,6471
8	3	2	0,0045	0,6111
9	3	3	0,0212	0,6111
10	3	4	0,0242	0,6875
11	11	6	**0,4174**	**1,0000**
12	2	6	0,0909	0,6875

Źródło: opracowanie własne.

5. Podsumowanie i dalsze badania

Poczta elektroniczna jest obecnie jedną z najpopularniejszych form komunikacji, głównie z powodu jej wydajności, niskich kosztów operacyjnych i kompatybilności z różnymi rodzajami informacji. Komunikacja pomiędzy pracownikami firm odbywa się głównie z wykorzystaniem tego narzędzia. Bogaty zasób informacji zbierany podczas komunikacji daje szerokie możliwości badania relacji pomiędzy pracownikami.

Przedstawione w artykule wybrane aspekty analizy sieci społecznościowych opartych na korespondencji e-mail pokazują, jak kształtuje się struktura organizacyjna wyodrębnionych podsieci, które wierzchołki w danej sieci są ważne ze względu na dystrybucję wiadomości w podsieci. Dodatkowo wykazują, które wierzchołki są nieaktywne w sieci. Wierzchołki te mogą źle wpływać na przepływ informacji w sieci.

Wśród kierunków dalszych badań można wskazać zbadanie większej podsieci bądź całej sieci, co mogłoby przynieść więcej interesujących informacji o strukturze organizacyjnej jednostki publicznej. Ewentualne zwiększenie obszaru czasowego z jednego miesiąca, np. na kwartał bądź pół roku, mogłoby wykazać całkowicie nowe obserwacje i wnioski.

Bibliografia

Batagelj V., Mrvar A., *Pajek – Program for Large Network Analysis,* University of Ljubljana, 1997.

Bekkerman R., McCallum A., Huang G., *Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora*, Computer Science Department Faculty Publication Series 218, 2004.

Boryczka U., Probierz B., Kozak J., *An Ant Colony Optimization Algorithm for an Automatic Categorization of Emails,* Springer, LNCS 8733, w: *Computational Collective Intelligence. Technologies and Applications*, D. Hwang, J.J. Janson, N.T. Nguyen (red.), 2014, s. 583–592.

Gloor P.A., *Capturing team dynamics through temporal social surfaces*, w: *Information Visualization*, E. Banissi, M. Sarfraz, J.C. Roberts, B. Loften, A. Ursyn, R.A. Burkhard, A. Lee, G. Andrienko (red.), 2005, s. 939–944.

McCallum A., Wang X., Corrada-Emmanuel A., *Topicand Role Discovery in Social Networks with Experiments on Enron and Academic Email*, „Journal of Artificial Intelligence Research" 2007, vol. 30, s. 249–272.

Rousseau R., Otte E., *Social Network Analysis: A Powerful Strategy, also for the Information Sciences*, „Journal of Information Science" 2002, 28, s. 442–444.

Vargas D., Bridgeman A., Schmidt D., Kohl P., Wilcox B., Carr L., *Correlation Between Student Collaboration Network Centrality and Academic Performance*, Carr Department of Physics, Colorado School of Mines, Golden, CO 80401, USA, August 2, 2018, s. 6.

Wasserman S., Faust K., *Social Network Analysis: Methods and Applications*, Cambridge University Press, Cambridge, New York 1994, s. 2–3.

Zdziarski M., *Analiza sieci*, w: *Sieci międzyorganizacyjne. Współczesne wyzwanie dla teorii i praktyki zarządzania*, J. Niemczyk, E. Stańczyk-Hugiet, B. Jasiński (red.), t. 1, Warszawa 2012, s. 35–42.

Zhang X., Zhu J., Wang Q., Zhao H., *Identifying Influential Nodes in Complex Networks with Community Structure*, „Knowledge-Based Systems" 2013, vol. 42, s. 74–84.

Źródła sieciowe

https://www.cs.cmu.edu/~enron/ (dostęp: 22.04.2018).

* * *

Selected aspects of analysis of social networks based on communication by electronic mail in a public institution

Abstract

Social Network Analysis (SNA) is based on the study of the structure, links and behaviour of specific units within social groups, represented in the form of vertices (referring to, for example, persons or organizations) and edges (defining interrelations or flow of information between these units). Among the network properties usually analysed one can indicate centrality, the number and strength of connections between vertices, or their transitivity. SNA can specify positions of selected units in a given network, their roles in the organization, or discover certain patterns in the relations between the units represented in the network.

The observed constant development of information technologies, the widespread use of social networking sites, or the use of electronic communication tools in contacts between people, including between employees and/or groups of employees in an organization, suggests that a huge amount of data related to these activities is stored in various data repositories can provide interesting information about the people themselves as well as about the relationships between them.

The work focuses on the analysis of the social network, created on the basis of communication of individuals by means of electronic mail in a public institution. Selected aspects related to the analysis of such a network were presented, in particular the basic features of such a network were identified, the relationships between individuals were examined, the hierarchy of users of such a social network was created, as well as exploration of data contained in such a network. Using the basic SNA measures, the most important vertices in the network are indicated. For the purpose of the experiment, the Pajek tool was used.

Keywords: e-mail communication, Social Network Analysis