ZESZYTY NAUKOWE UNIWERSYTETU SZCZECIŃSKIEGO
NR 863 STUDIA INFORMATICA NR 36 2015
DOI: 10.18276/si.2015.36-11

Barbara Probierz[*]
Jan Kozak[**]
Urszula Boryczka[***]
Uniwersytet Śląski w Katowicach

ANALIZA ZBIORU WIADOMOŚCI E-MAIL Z ZASTOSOWANIEM SIECI SPOŁECZNYCH

Streszczenie

W artykule zaproponowane zostało podejście związane z analizą sieci społecznych, a także praktyczne możliwości zastosowania tych sieci w badaniu organizacji pod kątem procesów przepływu wiadomości mailowych pracowników. Celem pracy jest analiza kontaktów pomiędzy poszczególnymi pracownikami korporacji zastosowana do wyznaczenia pracowników – liderów z punktu widzenia rozprzestrzeniania się informacji lub wpływania na osoby będące w bezpośrednim sąsiedztwie. Analiza ta w dalszych pracach powinna przyczynić się do stworzenia algorytmu, którego zastosowanie posłuży do poprawienia dokładności klasyfikacji wiadomości mailowych do poszczególnych folderów w skrzynkach pocztowych pracowników. Zaproponowana metoda została przetestowana na ogólnodostępnym zbiorze danych Enron E-mail.

Słowa kluczowe: Enron E-mail, sieci społeczne, SNA, analiza danych

[*] Adres e-mail: barbara.probierz@us.edu.pl.
[**] Adres e-mail: jan.kozak@us.edu.pl.
[***] Adres e-mail: urszula.boryczka@us.edu.pl.

Wprowadzenie

Historia powstania wiadomości e-mail zaczęła się prawie pół wieku temu, kiedy to w 1965 r. Louis Pouzin, Glenda Schroeder i Pat Crisman przesłali wiadomość tekstową pomiędzy dwoma użytkownikami. Niestety, usługa ta umożliwiała zostawienie wiadomości innym użytkownikom tego samego komputera, a adres poczty elektronicznej jeszcze wtedy nie istniał. Dopiero w 1971 r. amerykański inżynier i programista Raymond S. Tomlinson wpadł na pomysł, dzięki któremu udało się wysłać wiadomość tekstową pomiędzy dwoma komputerami. W celu oddzielenia nazwy użytkownika od nazwy komputera R. Tomlinson zastosował znak @, który w tamtych czasach używany był sporadycznie. Na tej podstawie w 1973 r. członkowie stowarzyszenia Internet Engineering Task Force uzgodnili standardową składnię dla komunikacji mailowej: *użytkownik@host*, która funkcjonuje do dziś.

E-mail jako narzędzie komunikacji może być wykorzystane do różnych celów. Najczęściej jest to chęć przekazania komunikatu, zareklamowania produktu, poinformowania o promocji, przesłania dokumentów czy też utrzymania kontaktu z klientami. To, co umieścimy w naszych wiadomościach nie tylko przekazuje konkretne treści, ale też jest swego rodzaju wizytówką i najlepiej świadczy o poziomie naszej edukacji i wychowania.

Większość ludzi nie wyobraża sobie możliwości funkcjonowania bez dostępu do elektronicznej skrzynki pocztowej. Jednak największym problemem użytkowników, zwłaszcza tych, dla których e-mail to podstawa komunikacji, jest odpowiednie uporządkowanie poczty elektronicznej i przypisanie wiadomości do poszczególnych folderów. Zwłaszcza, gdy kategoryzacja ta ma się odbywać w sposób automatyczny.

Określenie problemu

Typowy użytkownik dostaje ok. 40–50 wiadomości e-mail każdego dnia. Niektórzy otrzymują ich nawet setki dziennie, przez co użytkownicy znaczną część swojego czasu pracy poświęcają na czytanie i odpowiadanie na otrzymane wiadomości e-mail. W rezultacie ostatnio jest coraz większe zainteresowanie tworzeniem systemów, które w sposób automatyczny mogą pomóc użytkownikom w zarządzaniu pocztą elektroniczną.

Wiadomość e-mail ma bardzo skomplikowany format w wielu różnych wymiarach, gdyż wiadomości mogą być przesyłane, przekazywane, kopiowane, a także może na nie odpowiadać wiele osób lub grup w różnym czasie. Dodatkowo e-maile mogą zawierać jako załączniki inne wiadomości e-mail lub dokumenty w postaci dołączonych plików. Ponadto, informacje uzyskane z tematu maila mogą mieć inne znaczenie niż informacje uzyskane z treści lub załączników.

Szczególnym przypadkiem w klasyfikacji jest przypisanie wiadomości mailowych do folderów (E-mail Foldering Problem – EFP), co polega na tym, że użytkownicy tworzą nowe katalogi, ale także przestają korzystać z niektórych folderów utworzonych wcześniej. Jednocześnie foldery nie zawsze odpowiadają tematom maili, czasami mogą dotyczyć zadań do wykonania, grup projektowych, niektórych odbiorców, a inne mają sens tylko w powiązaniu z poprzednimi wiadomościami (Bekkerman, 2004).

Proces przypisania wiadomości do folderów jest problemem złożonym, gdyż automatyczna metoda klasyfikacji może się sprawdzić u jednego użytkownika, a u innego może prowadzić do błędów. Ponadto informacje docierają w różnym czasie, co powoduje dodatkowe trudności. Do rozwiązania tego problemu może przyczynić się analiza kontaktów pomiędzy nadawcą a odbiorcami wiadomości mailowych, a także drzewiasta struktura folderów utworzona w poszczególnych skrzynkach odbiorczych.

Sieć społeczna

Sieć społeczna (Social Network – SN) to wielowymiarowa struktura złożona ze zbioru jednostek społecznych oraz połączeń między nimi. Jednostki społeczne to osoby funkcjonujące w danej sieci, natomiast połączenia odwzorowującą różnorodne relacje społeczne pomiędzy poszczególnymi osobami.

Najczęściej sieć społeczną przedstawia się w postaci grafu. Zgodnie z matematyczną definicją, graf to uporządkowana para:

$$G = (V, E),$$

gdzie V jest skończonym zbiorem wierzchołków grafu, natomiast E jest skończonym zbiorem wszystkich dwuelementowych podzbiorów zbioru V zwanych krawędziami, łączącymi poszczególne wierzchołki, takim, że:

$$E \subseteq \{\{u, v\} : u, v \in V, u \neq v\}.$$

W grafie wierzchołki reprezentują obiekty, natomiast krawędzie obrazują relacje między tymi obiektami. W zależności od tego, czy relacja ta ma charakter symetryczny czy też nie, graf wykorzystywany do opisu sieci może być grafem nieskierowanym lub grafem skierowanym.

Krawędzie w sieci społecznej reprezentują interakcję, przepływ informacji i dóbr, podobieństwo, afiliację lub związki społeczne. Dla związków społecznych miarą siły powiązania może być:

- częstotliwość interakcji lub przepływu informacji,
- wzajemność interakcji lub przepływu,
- rodzaj interakcji lub przepływu,
- atrybuty łączonych węzłów lub krawędzi (np. stopień pokrewieństwa),
- struktura sąsiedztwa łączonych węzłów (np. liczba wspólnych sąsiadów).

Związki społeczne, ze względu na ich siłę powiązania, można podzielić na dwie grupy – mocne bądź słabe. Cechą wyróżniającą wśród słabych związków jest tworzenie się mostów, czyli powiązań między poszczególnymi węzłami, które łączą różne grupy obiektów. Mosty zazwyczaj ułatwiają komunikację między grupami, zwiększają spójność, umożliwiają rozprzestrzenianie informacji i innowacji. Natomiast wśród silnych związków występuje element przechodniości. Jest to cecha powiązań, dla których istnienie związków między wierzchołkami A i B oraz B i C sugeruje istnienie związku między wierzchołkami A i C. Przechodniość prowadzi do powstawania klik i pseudo-klik.

Dodatkowymi wskaźnikami charakteryzującymi daną sieć społeczną są stopnie wierzchołków oraz centralność wg tych stopni. Stopień wierzchołka (stopień wejściowy, stopień wyjściowy) to liczba krawędzi wchodzących lub wychodzących z danego węzła. Natomiast centralność wg stopni wierzchołków jest użyteczna do określania, które węzły są kluczowe z punktu widzenia rozprzestrzeniania informacji lub wpływania na węzły położone w bezpośrednim sąsiedztwie. Centralność często mierzy popularność lub wpływowość tych węzłów.

W badaniach nad zbiorami danych zawierających wiadomości e-mail niezwykle ważną rolę odgrywa analiza sieci społecznych (Social Network Analysis – SNA). Jest to przede wszystkim specyficzna perspektywa analizy, która

nie skupia się na indywidualnych jednostkach lub makrostrukturach, lecz bada powiązania między poszczególnymi jednostkami czy grupami.

Pierwsze badania sieci społecznych przeprowadził w 1923 r. Jacob L. Moreno, który uznawany jest za jednego z założycieli dyscypliny analizy sieci społecznych. Jest to gałąź socjologii, która zajmuje się ilościową oceną roli jednostki w grupie lub społeczności przez analizę sieci powiązań między jednostkami. Jego książka *Who Shall Survive?* z 1934 r. zawiera pierwsze graficzne przedstawienia sieci społecznych, a także definicje kluczowych terminów w analizie sieci społecznych i sieci socjometrycznych (Moreno, 1923).

Wiele badań SNA dotyczyło znajdowania korelacji między społeczną strukturą sieci a wydajnością (Gloor, 2006). Początkowo analiza sieci społecznych przeprowadzana była na podstawie ankiet wypełnianych ręcznie przez uczestników (Cummings, 2003), jednak z czasem popularne stały się badania przeprowadzane z zastosowaniem wiadomości e-mail (Aral, 2007). W niektórych badaniach stwierdzono, że zespoły badawcze są bardziej kreatywne, gdy posiadają większy kapitał społeczny (Gloor, 2012).

Analiza sieci społecznych ma szeroki zakres zastosowań. Przede wszystkim stosowana jest w dużych organizacjach i firmach jako narzędzie wspierające strategiczne zarządzanie zasobami ludzkimi czy też zarządzanie wiedzą w organizacji. SNA wspiera innowacyjność firmy, a także służy analizie procesów biznesowych oraz analizie potrzeb szkoleń. Dodatkowo wykorzystywana jest przy badaniach marketingowych w tworzeniu mapy sieci społecznej klientów. Analiza ta pozwala jednak przede wszystkim kadrze zarządzającej na zapoznanie się z nieformalną strukturą organizacji i przepływu informacji w firmie.

Baza Enron E-mail

Enron E-mail Dataset to zestaw danych zebrany i przygotowany przez Projekt CALO (ang. A Cognitive Assistant that Learns and Organizes) (CALO, 2014). Zawiera ponad 600 tys. wiadomości e-mail, które zostały wysłane lub odebrane przez 158 pracowników wyższego szczebla z Enron Corporation. Zbiór danych został przejęty przez Komisję Regulacji Energetyki Federalnej w trakcie dochodzenia po upadku firmy, a następnie został podany do publicznej wiadomości. Kopia bazy danych została wykupiona przez Leslie Kaelbling z Massachusetts Institute of Technology (MIT), po czym okazało się, że w zbiorze są duże

problemy związane z integralnością danych. Dzięki pracy zespołu z ośrodka SRI International, zwłaszcza Melinda Gervasio, dane zostały poprawione i udostępnione innym naukowcom do badań.

Baza danych uważana jest za jeden z cenniejszych zbiorów, gdyż zawiera rzeczywiste wiadomości e-mail dostępne publicznie, co często jest problematyczne z uwagi na prywatność danych z innych zestawów. Wiadomości te są przypisane do kont osobistych i podzielone na foldery. W zbiorze danych nie ma załączników do e-maili, a niektóre wiadomości zostały usunięte ze względu na występowanie duplikatów w innych folderach. Brakujące informacje zostały w miarę możliwości uzupełnione na podstawie innych treści, a gdy nie było możliwości określenia odbiorcy wprowadzono frazę *no_address@enron.com*.

Zbiór danych E-mail Enron jest powszechnie stosowany do badań związanych z analizą sieci społecznych, przetwarzaniem języka naturalnego oraz uczenia maszynowego. Klasyfikacja e-mail może być używana do wielu różnych zastosowań, w szczególności do filtrowania wiadomości na podstawie priorytetu przypisywania e-maili do folderów utworzonych przez użytkownika, a także do identyfikacji spamu.

Doświadczenia związane z analizą zbioru danych

Sieci społeczne związane są również z budowaniem sieci komunikacji (Wilson, 2009), czyli procesem wymiany informacji, zasobów i możliwości, prowadzonym przy pomocy wzajemnie korzystnych kontaktów. Celem pracy jest analiza kontaktów pomiędzy poszczególnymi pracownikami korporacji zastosowana do wyznaczenia liderów z punktu widzenia rozprzestrzeniania się informacji lub wpływania na osoby będące w bezpośrednim sąsiedztwie. Analiza ta w dalszych pracach pozwoli na stworzenie algorytmu, którego zastosowanie posłuży do poprawienia dokładności klasyfikacji wiadomości mailowych do poszczególnych folderów w skrzynkach pocztowych pracowników firmy Enron (Boryczka, 2014).

Omawiana metoda nie tylko pozwoli na skrócenie czasu przeznaczanego na czytanie i odpowiadanie na otrzymane maile, ale przede wszystkim pozwoli na odtworzenie mapy kontaktów w postaci sieci powiazań społecznych, która ma kluczowe znaczenie dla procesów przepływu informacji w korporacji.

Skupiając się na stworzeniu i analizie mapy powiązań społecznych na podstawie zbioru Enron E-mail, badania zostały podzielone na trzy etapy:

- analiza całej sieci (analiza makro),
- analiza części sieci (analiza mezo),
- analiza poszczególnych pracowników (analiza mikro).

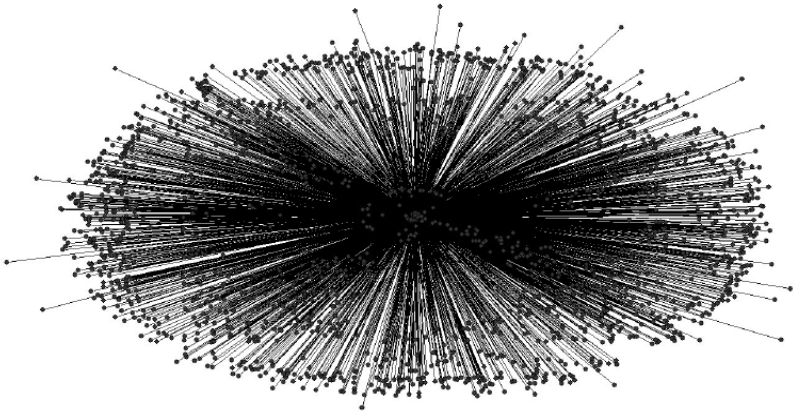
Dzięki takiemu podejściu do badań można otrzymać nie tylko ogólny obraz komunikacji w organizacji, ale przede wszystkim pozwala to na uzyskanie nieformalnej struktury firmy i mapy przepływu informacji w przedsiębiorstwie (Stępka, 2004).

Sieć jako całość (analiza makro)

Analiza makro to spojrzenie na organizację jako całość, dzięki czemu można określić charakter firmy pod względem komunikacji i współpracy wszystkich pracowników, a niekiedy także klientów. Poprzez mapowanie procesów komunikacji, czy analizy poziomu i struktury znajomości pracowników w danym przedsiębiorstwie, powstaje swoista, nieformalna struktura organizacyjna przedsiębiorstwa.

W pierwszym etapie przeprowadzonych badań dotyczących tworzenia i analizy sieci powiązań na podstawie zbioru Enron Email pod uwagę wzięte zostały wszystkie wysłane i odebrane maile. Obiektami w sieci zostały wszystkie adresy mailowe występujące choć raz w całym zbiorze. Natomiast powiązania między tymi obiektami to informacja o wysłaniu bądź odebraniu przynajmniej jednej wiadomości mailowej.

Na rys. 1 przedstawiona została wizualizacja sieci dla całego zbioru wiadomości mailowych firmy Enron. Sieć ta jest zbudowana z 1 914 obiektów, w skład których wchodzą nie tylko pracownicy korporacji Enron, ale także klienci i inne osoby zewnętrze, którzy kontaktowali się ze sobą za pomocą poczty elektronicznej. Wszystkie obiekty sieci połączone są ze sobą 4 378 krawędziami. Liczba ta wskazuje powiązania pomiędzy poszczególnymi osobami. Dla takiego zbioru obiektów i krawędzi częstotliwość przepływu informacji, czyli liczba przesłanych wiadomości mailowych wynosi 462 976.



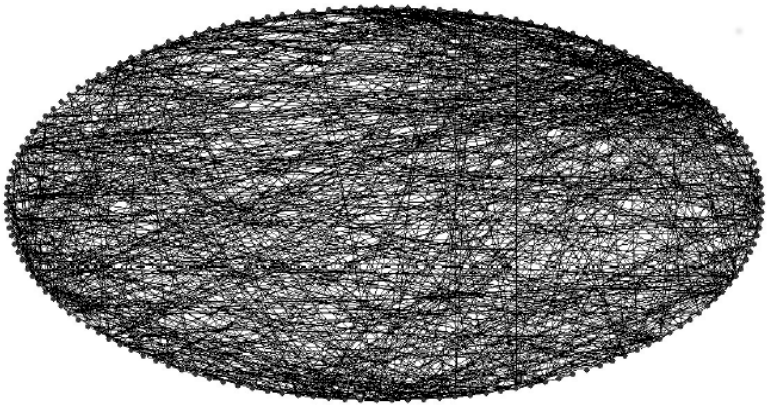
Rys. 1. Wizualizacja sieci społecznej dla zbioru Enron E-mail
Źródło: opracowanie własne.

Analiza podsieci (analiza mezo)

W kolejnym etapie przeprowadzonych badań skupiono się na analizie grupy pracowników będących na stanowiskach menadżerskich firmy Enron. Analiza sieci ograniczona do pewnej grupy społecznej w ramach danego przedsiębiorstwa nosi nazwę analizy mezo. Analiza ta skupia się na wewnętrznych relacjach danej grupy obiektów, wyodrębnionych ze względu na formalne kryteria podziału, tj. przynależność do odpowiednich działów, staż pracy lub stanowisko. Stosując tę metodę można określić nieformalne grupy pracowników, którzy w szczególny sposób ze sobą współpracują bądź komunikują się ze sobą, dzięki swojej wiedzy, bądź uczestniczą w tym samym procesie dotyczącym np. danego projektu.

Skrzynki pocztowe tych osób wraz z utworzonymi folderami i przypisanymi do nich wiadomościami mailowymi dostępne są w postaci zbioru Enron E-mail. Skrzynek tych jest 158, znajduje się w nich ok. 600 tys. wiadomości mailowych. Na potrzeby stworzenia sieci społecznej z badanego zbioru danych zostały usunięte wszystkie wiadomości mailowe wysłane przez pracowników firmy na własny adres mailowy. Ze zbioru wyeliminowane zostały także foldery utworzone przez programy pocztowe w sposób automatyczny, które w nazwie zawierają hasła sent lub inbox. Dodatkowo usunięto maile, których nadawca lub odbiorca występował tylko jeden raz w całym zbiorze maili.

Po takiej modyfikacji w zbiorze zostało 150 skrzynek pocztowych zawierających wiadomości mailowe przypisane do różnych folderów. W tej części przeprowadzonych badań stworzono sieć społeczną zawierającą 150 obiektów, które powiązane były ze sobą 1361 krawędziami. Pomiędzy tymi osobami zostało przesłanych 15 024 wiadomości mailowych co jest określane jako częstotliwość przepływu informacji. Wizualizacja opisanej sieci społecznej została przedstawiona na rys. 2, gdzie najwyższy stopień wierzchołka, czyli liczba krawędzi wchodzących i wychodzących z danego obiektu wynosi 926.



Rys. 2. Wizualizacja sieci społecznej dla 150 skrzynek pocztowych
Źródło: opracowanie własne.

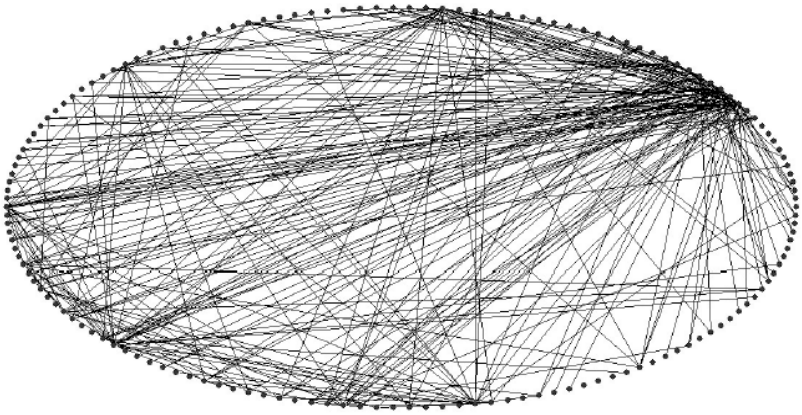
Analiza sieci dla najważniejszego obiektu (analiza mikro)

W kolejnym kroku przeprowadzonych badań skupiono się na stworzeniu oraz analizie sieci społecznej dotyczącej przepływu wiadomości mailowych pomiędzy jednym pracownikiem a pozostałymi osobami. Pracownik ten został wybrany na podstawie kryterium przepływu wiedzy i informacji, w związku z czym został określony jako najważniejszy obiekt w sieci, ponieważ z jego skrzynki pocztowej zostało przesłanych najwięcej wiadomości mailowych.

Metody badań SNA pozwalają na analizowanie małego wycinka sieci jakim jest sieć relacji poszczególnego pracownika. Taka analiza nosi nazwę analizy mikro. Dzięki tej metodzie istnieje możliwość zidentyfikowania pracowników

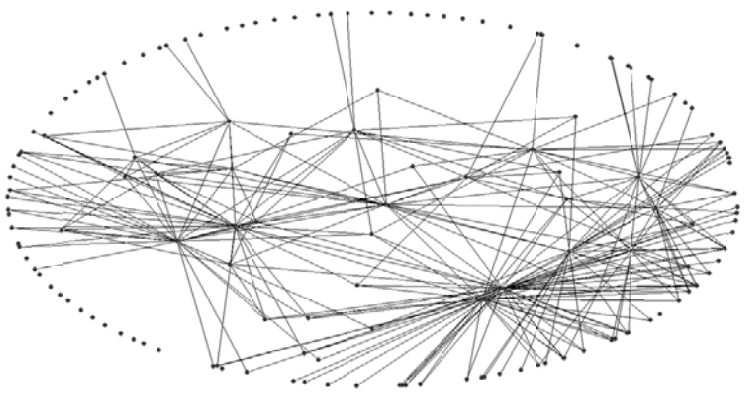
tworzących tzw. wąskie gardła w ramach procesu przepływu informacji, ale także pracowników będących liderami w swojej dziedzinie.

Na rys. 3 przedstawiona jest wizualizacja sieci społecznej dla najważniejszego obiektu, wybranego ze względu na największy stopień wierzchołka. Przedstawiona sieć składa się ze 150 obiektów, połączonych ze sobą 301 krawędziami. Częstotliwość przepływu informacji wynosi 5844 przesłanych maili.



Rys. 3. Wizualizacja sieci społecznej dla najważniejszego obiektu
Źródło: opracowanie własne.

Ze względu na brak kontaktów niektórych pracowników z najważniejszym obiektem w sieci wyodrębnionych zostało 45 obiektów niepowiązanych ze sobą, co zostało przedstawione na rys. 4. Jednak należy pamiętać, że przedstawiona sieć dotyczy tylko relacji najważniejszego obiektu z pozostałymi. W przypadku wybrania innego obiektu jako najważniejszego, relacje w nowej sieci społecznej są zupełnie inne niż dotychczas przedstawione, a obiekty niepowiązane w tej sieci posiadają połączenia z innymi obiektami nowej sieci.



Rys. 4. Analiza sieci społecznej dla najważniejszego aktora
Źródło: opracowanie własne.

Podsumowanie

Przedstawiona praca dotyczy analizy zbioru wiadomości mailowych opartej na budowie sieci społecznych. Celem wykonanych badań było stworzenie i analiza mapy kontaktów pomiędzy poszczególnymi pracownikami korporacji Enron zastosowanej do wyznaczenia najważniejszych osób z punktu widzenia rozprzestrzeniania się informacji lub wpływania na osoby będące w bezpośrednim sąsiedztwie tych osób.

Na podstawie stworzonej sieci powiązań oraz przeprowadzonej analizy można przypuszczać, że wprowadzenie mechanizmu sieci społecznych do algorytmu zaproponowanego w innych pracach autorów, może w jeszcze większy sposób przyczynić się do poprawy przypisania wiadomości e-mail do folderów, co zostanie zbadane w najbliższej przyszłości.

Bibliografia

Aral S., Van Alstyne M. (2007), *Network structure & information advantage*.

Bekkerman R., McCallum A., Huang G. (2004), *Automatic categorization of email into folders: Benchmark experiments on Enron and SRI corpora*, Center for Intelligent Information Retrieval, Technical Report IR.

Boryczka U., Probierz B., Kozak J. (2014), *An Ant Colony Optimization Algorithm for an Automatic Categorization of Emails*, Computational Collective Intelligence. Technologies and Applications, LNCS, Springer, Berlin, s. 583–592.

CALO (2014), *A Cognitive Assistant that Learns and Organizes*, www.ai.sri.com/project/CALO (2.11.2014).

Cummings J.N., Cross R. (2003), *Structural properties of work groups and their consequences for performance*, „Social Networks", no. 25, s. 197–210.

Gloor P.A. (2006), *Swarm Creativity: Competitive Advantage through Collaborative Innovation Networks*, Oxford Univrsity Press.

Gloor P., Grippa F., Putzke J., Lassenius C., Fuehres H., Fischbach K., Schoder D. (2012), *Measuring social capital in creative teams through sociometric sensors*, „International Journal of Organisational Design and Engineering".

Moreno J.L. (1953), *Who Shall Survive? Foundations of Sociometry, Group Psychotherapy and Sociodrama*, Beacon House, Beacon, Nowy Jork.

Stępka P., Subda K. (2004), *Wykorzystanie analizy sieci społeznych (SNA) do budowy organizacji opartej na wiedzy*, „E-mentor", nr 1 (28).

Wilson G.C., Banzhaf W. (2009), *Discovery of email communication networks from the enron corpus with a genetic algorithm using social network analysis*.

AN ANALYSIS OF THE SET OF E-MAILS WITH SOCIAL NETWORKS

Summary

In this article is proposed an approach based on the Social Network Analysis and its practical applicability in the study of the organization in terms of flow processes e-mails employees. The aim of this paper is to analyze the interaction between individual employees corporation used to designate staff-leaders from the point of view of spreading information or to influence those in the immediate vicinity. This analysis further work should contribute to the creation of the algorithm, the application of which will be used to improve the accuracy of the classification of e-mail messages to specific folders in the mailboxes of employees. The proposed method has been tested on a public dataset Enron Email.

Translated by Barbara Probierz

Keywords: Enron E-mail, Social Network Analysis, data analysis