

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS

of Wrocław University of Economics

Nr 426

Taksonomia 26

**Klasyfikacja i analiza danych –
teoria i zastosowania**



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2016

Redaktor Wydawnictwa: Agnieszka Flasińska

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

Informacje o naborze artykułów i zasadach recenzowania
znajdują się na stronach internetowych
www.pracnaukowe.ue.wroc.pl
www.wydawnictwo.ue.wroc.pl

Publikacja udostępniona na licencji Creative Commons
Uznanie autorstwa-Użycie niekomercyjne-Bez utworów zależnych 3.0 Polska
(CC BY-NC-ND 3.0 PL)



© Copyright by Uniwersytet Ekonomiczny we Wrocławiu
Wrocław 2016

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
e-ISSN 2392-0041
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Zamówienia na opublikowane prace należy składać na adres:
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
ul. Komandorska 118/120, 53-345 Wrocław
tel./fax 71 36 80 602; e-mail:econbook@ue.wroc.pl
www.ksiegarnia.ue.wroc.pl

Druk i oprawa: TOTEM

Spis treści

Wstęp	9
Jacek Batóg: Identyfikacja obserwacji odstających w analizie skupień / Influence of outliers on results of cluster analysis	13
Andrzej Bąk: Porządkowanie liniowe obiektów metodą Hellwiga i TOPSIS – analiza porównawcza / Linear ordering of objects using Hellwig and TOPSIS methods – a comparative analysis.....	22
Grażyna Dehnel: <i>MM</i> -estymacja w badaniu średnich przedsiębiorstw w Polsce / <i>MM</i> -estimation in the medium-sized enterprises survey in Poland.....	32
Andrzej Dudek: <i>Social network analysis</i> jako gałąź wielowymiarowej analizy statystycznej / Social network analysis as a branch of multidimensional statistical analysis.....	42
Iwona Foryś: Analiza dyskryminacyjna w wyborze obiektów podobnych w procesie szacowania nieruchomości / The discriminant analysis in selection of similar objects in the real estate valuation process	51
Gregory Kersten, Ewa Roszkowska, Tomasz Wachowicz: Ocena zgodności porządkowej systemu oceny ofert negocjatora z informacją preferencyjną / Analyzing the ordinal concordance of preferential information and resulting scoring system in negotiations.....	60
Iwona Konarzewska: Rankingi wielokryteriowe a współzależność liniowa kryteriów / Multi-criteria rankings and linear relationships among criteria	69
Anna Król, Marta Targaszewska: Zastosowanie klasyfikacji do wyodrębniania homogenicznych grup dóbr w modelowaniu hedonicznym / The application of classification in distinguishing homogeneous groups of goods for hedonic modelling.....	80
Marek Lubicz: Problemy doboru zmiennych objaśniających w klasyfikacji danych medycznych / Feature selection and its impact on classifier effectiveness – case study for medical data.....	89
Aleksandra Łuczak: Wpływ różnych sposobów agregacji opinii ekspertów w FAHP na oceny priorytetowych czynników rozwoju / Influence of different methods of the expert judgments aggregation on assessment of priorities for evaluation of development factors in FAHP.....	99
Iwona Markowicz: Tablice trwania firm w województwie zachodniopomorskim według rodzaju działalności / Companies duration tables in Zachodniopomorskie voivodship by the type of activity	108

Małgorzata Markowska, Danuta Strahl: Filary inteligentnego rozwoju a wrażliwość unijnych regionów szczebla NUTS 2 na kryzys ekonomiczny – analiza wielowymiarowa / Smart development pillars and NUTS 2 European regions vulnerability to economic crisis – a multidimensional analysis.....	118
Kamila Migdał-Najman, Krzysztof Najman: Hierarchiczne deglomeracyjne sieci SOM w analizie skupień / The hierarchical divisive SOM in the cluster analysis	130
Kamila Migdał-Najman, Krzysztof Najman: Hierarchiczne aglomeracyjne sieci SOM w analizie skupień / The hierarchical agglomerative SOM in the cluster analysis	139
Barbara Pawelek, Józef Pocięcha, Jadwiga Kostrzewska, Mateusz Baryła, Artur Lipieta: Problem wartości odstających w prognozowaniu zagrożenia upadłością przedsiębiorstw (na przykładzie przetwórstwa przemysłowego w Polsce) / Problem of outliers in corporate bankruptcy prediction (case of manufacturing companies in Poland)	148
Wojciech Roszka: Syntetyczne źródła danych w analizie przestrzennego zróżnicowania ubóstwa / Synthetic data sources in spatial poverty analysis.....	157
Małgorzata Rószkiewicz: Czynniki różnicujące efektywność pracy ankietera w wywiadach <i>face-to-face</i> w środowisku polskich gospodarstw domowych / Factors affecting the efficiency of face-to-face interviews with Polish households.....	166
Adam Sagan, Marcin Pelka: Analiza wielopoziomowa z wykorzystaniem danych symbolicznych / Multilevel analysis with application of symbolic data	174
Marcin Salamaga: Zastosowanie drzew dyskryminacyjnych w identyfikacji czynników wspomagających wybór kraju alokacji bezpośrednich inwestycji zagranicznych na przykładzie polskich firm / The use of classification trees in the identification of factors supporting the choice of FDI destination on the example of Polish companies.....	185
Agnieszka Stanimir: Pomiar wykluczenia cyfrowego – zagrożenia dla Pokolenia Y / Measurement of the digital divide – risks for Generation Y ...	194
Mirosława Sztemberg-Lewandowska: Grupowanie danych funkcjonalnych w analizie poziomu wiedzy maturzystów / Functional data clustering methods in the analysis of high school graduates' knowledge	206
Tadeusz Trzaskalik: Modelowanie preferencji w wielokryterialnych dyskretnych problemach decyzyjnych – przegląd bibliografii / Preference modeling in multi-criteria discrete decision making problems – review of literature	214

Joanna Trzęsiok: Metody nieparametryczne w badaniu zaufania do instytucji finansowych / Nonparametric methods in the study of confidence in financial institutions	226
Hanna Wdowicka: Analiza sytuacji na lokalnych rynkach pracy w Polsce / Local labour market analysis in Poland.....	235
Artur Zaborski: Zastosowanie skalowania dynamicznego oraz metody wektorów dryfu do badania zmian w preferencjach / The use of dynamic scaling and the drift vector method for studying changes in the preferences.....	245

Wstęp

W dniach 14–16 września 2015 r. w Hotelu Novotel Gdańsk Marina w Gdańsku odbyła się XXIV Konferencja Naukowa Sekcji Klasyfikacji i Analizy Danych PTS (XXIX Konferencja Taksonomiczna) „Klasyfikacja i analiza danych – teoria i zastosowania”, zorganizowana przez Sekcję Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego oraz Katedrę Statystyki Wydziału Zarządzania Uniwersytetu Gdańskiego. Przewodniczącymi Komitetu Organizacyjnego konferencji byli prof. dr hab. Mirosław Szreder oraz dr hab. Krzysztof Najman, prof. nadzw. UG, sekretarzami naukowymi dr hab. Kamila Migdał-Najman, prof. nadzw. UG oraz dr hab. Anna Zamojska, prof. nadzw. UG, a sekretarzem organizacyjnym Anna Nowicka z Fundacji Rozwoju Uniwersytetu Gdańskiego.

Konferencja Naukowa została dofinansowana ze środków Narodowego Banku Polskiego.

Zakres tematyczny konferencji obejmował takie zagadnienia, jak:

a) teoria (taksonomia, analiza dyskryminacyjna, metody porządkowania liniowego, metody statystycznej analizy wielowymiarowej, metody analizy zmiennych ciągłych, metody analizy zmiennych dyskretnych, metody analizy danych symbolicznych, metody graficzne),

b) zastosowania (analiza danych finansowych, analiza danych marketingowych, analiza danych przestrzennych, inne zastosowania analizy danych – medycyna, psychologia, archeologia, itd., aplikacje komputerowe metod statystycznych).

Zasadniczymi celami konferencji SKAD były prezentacja osiągnięć i wymiana doświadczeń z zakresu teoretycznych i aplikacyjnych zagadnień klasyfikacji i analizy danych. Konferencja stanowi coroczne forum służące podsumowaniu obecnego stanu wiedzy, przedstawieniu i promocji dokonań nowatorskich oraz wskazaniu kierunków dalszych prac i badań.

W konferencji wzięło udział 81 osób. Byli to pracownicy oraz doktoranci następujących uczelni i instytucji: AGH w Krakowie, Politechniki Łódzkiej, Politechniki Gdańskiej, Politechniki Opolskiej, Politechniki Wrocławskiej, Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie, Szkoły Głównej Handlowej w Warszawie, Uniwersytetu im. Adama Mickiewicza w Poznaniu, Uniwersytetu Ekonomicznego w Katowicach, Uniwersytetu Ekonomicznego w Krakowie, Uniwersytetu Ekonomicznego w Poznaniu, Uniwersytetu Ekonomicznego we Wrocławiu, Uniwersytetu Gdańskiego, Uniwersytetu Jana Kochanowskiego w Kielcach, Uniwersytetu Łódzkiego, Uniwersytetu Mikołaja Kopernika w Toruniu, Uniwersytetu Przyrodniczego w Poznaniu, Uniwersytetu Szczecińskiego, Uniwer-

sytetu w Białymstoku, Wyższej Szkoły Bankowej w Toruniu, a także przedstawiciele NBP i PBS Sp. z o.o.

W trakcie dwóch sesji plenarnych oraz trzynastu sesji równoległych wygłoszono 58 referatów poświęconych aspektom teoretycznym i aplikacyjnym zagadnienia klasyfikacji i analizy danych. Odbyła się również sesja plakatowa, na której zaprezentowano 14 plakatów. Obradom w poszczególnych sesjach konferencji przewodniczyli profesorowie: Józef Pocięcha, Eugeniusz Gatnar, Tadeusz Trzaskalik, Krzysztof Jajuga, Marek Walesiak, Barbara Pawełek, Feliks Wysocki, Ewa Roszkowska, Andrzej Sokołowski, Andrzej Bąk, Tadeusz Kufel, Mirosław Krzyśko, Krzysztof Najman, Małgorzata Rószkiewicz, Mirosław Szreder.

Teksty 25 recenzowanych artykułów naukowych stanowią zawartość prezentowanej publikacji z serii „Taksonomia” nr 26. Pozostałe recenzowane artykuły znajdują się w „Taksonomii” nr 27.

W pierwszym dniu konferencji odbyło się posiedzenie członków Sekcji Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego, któremu przewodniczył prof. dr hab. Józef Pocięcha. Ustalono plan przebiegu zebrania obejmujący następujące punkty:

- A. Sprawozdanie z działalności Sekcji Klasyfikacji i Analizy Danych PTS.
- B. Informacje dotyczące planowanych konferencji krajowych i zagranicznych.
- C. Organizacja konferencji SKAD PTS w latach 2016 i 2017.
- D. Wybór przedstawiciela Rady Sekcji SKAD PTS do IFCS.
- E. Dyskusja nad kierunkami rozwoju działalności Sekcji.

Prof. dr hab. Józef Pocięcha otworzył posiedzenie Sekcji SKAD PTS. Sprawozdanie z działalności Sekcji Klasyfikacji i Analizy Danych PTS przedstawiła sekretarz naukowy Sekcji dr hab. Barbara Pawełek, prof. nadzw. UEK. Poinformowała, że obecnie Sekcja liczy 231 członków. Przypomniała, że na stronie internetowej Sekcji znajdują się regulamin, a także deklaracja członkowska. Poinformowała, że zostały opublikowane zeszyty z serii „Taksonomia” nr 24 i 25 (PN UE we Wrocławiu nr 384 i 385). W „Przeglądzie Statystycznym” (zeszyt 4/2014) ukazało się sprawozdanie z ubiegłorocznej konferencji SKAD, która odbyła się w Międzyzdrojach, w dniach 8–10 września 2014 r. Prof. Barbara Pawełek przedstawiła także informacje dotyczące działalności międzynarodowej oraz udziału w ważnych konferencjach członków i sympatyków SKAD.

W konferencji Międzynarodowego Stowarzyszenia Towarzystw Klasyfikacyjnych (IFCS – International Federation of Classification Societies) w dniach 6–8 lipca 2015 r. w Bolonii, zorganizowanej przez Università di Bologna, udział wzięło 19 osób z Polski (w tym 17 członków Sekcji), które wygłosiły 15 referatów (wkład członków SKAD – 79,0%). Ponadto prof. Józef Pocięcha był członkiem Komitetu Naukowego Konferencji z ramienia SKAD, członkiem Międzynarodowego Komitetu Nagród IFCS oraz organizatorem i przewodniczącym sesji nt. „Classification models for forecasting of economic processes”.

W konferencji „European Conference on Data Analysis” (Colchester, 2–4 września 2015 r.) zorganizowanej przez The German Classification Society (GfKI) we współpracy z The British Classification Society (BCS) i Sekcją Klasyfikacji i Analizy Danych PTS (SKAD) udział wzięło 18 osób z Polski (w tym 14 członków Sekcji), które wygłosiły 15 referatów (wkład członków SKAD – 66,0%). Ponadto profesorowie Krzysztof Jajuga oraz Józef Pociecha byli członkami Komitetu Naukowego konferencji, prof. Andrzej Dudek został poproszony przez organizatorów o przygotowanie referatu i wygłoszenie na Sesji Plenarnej „Cluster analysis in XXI century, new methods and tendencies”, prof. Krzysztof Jajuga był przewodniczącym sesji plenarnej, przewodniczącym sesji nt. „Finance and economics II” oraz organizatorem i przewodniczącym sesji nt. „Data analysis in finance”, prof. Józef Pociecha był organizatorem i przewodniczącym sesji nt. „Outliers in classification procedures – theory and practice”, prof. Andrzej Dudek był przewodniczącym sesji nt. „Machine learning and knowledge discovery II”.

Kolejny punkt posiedzenia Sekcji obejmował zapowiedzi najbliższych konferencji krajowych i zagranicznych, których tematyka jest zgodna z profilem Sekcji. Prof. dr hab. Józef Pociecha poinformował o dwóch wybranych konferencjach krajowych (były to XXXIV Konferencja Naukowa „Multivariate Statistical Analysis MSA 2015”, Łódź, 16–18 listopada 2015 r. i X Międzynarodowa Konferencja Naukowa im. Profesora Aleksandra Zeliasia nt. „Modelowanie i prognozowanie zjawisk społeczno-gospodarczych”, Zakopane, 10–13 maja 2016 r.) oraz o trzech wybranych konferencjach zagranicznych. Konferencja „European Conference on Data Analysis” odbędzie się na Uniwersytecie Ekonomicznym we Wrocławiu w dniach 26–28 września 2017 r. W przeddzień tej konferencji, tj. 25.09.2017 r., odbędzie się Niemiecko-Polskie Sympozjum nt. „Analizy danych i jej zastosowań GPSDAA 2017”. Następna konferencja Międzynarodowego Stowarzyszenia Towarzystw Klasyfikacyjnych (IFCS) odbędzie się w 2017 r. w Tokio. W 2019 r. Niemiecko-Polskie Sympozjum nt. „Analizy danych i jej zastosowań GPSDAA 2019” organizuje prof. Andreas Geyer-Schultz w Karlsruhe.

W następnym punkcie posiedzenia podjęto kwestię organizacji kolejnych konferencji SKAD. SKAD 2016 zorganizuje Katedra Metod Statystycznych Wydziału Ekonomiczno-Socjologicznego Uniwersytetu Łódzkiego.

W kolejnej części zebrania dokonano wyboru przedstawiciela Rady Sekcji SKAD PTS do IFCS na kadencję 2016–2019. Powołano Komisję Skrutacyjną, której przewodniczącym został prof. Tadeusz Kufel, a członkami dr hab. Iwona Konarzewska i dr Dominik Rozkrut. Profesor Józef Pociecha poprosił zebranych o proponowanie kandydatur zgłaszając jednocześnie prof. Andrzeja Sokołowskiego. Wobec braku następnych kandydatur listę zamknięto. Komisja Skrutacyjna przeprowadziła głosowanie tajne. W głosowaniu uczestniczyło 41 członków Sekcji. Profesor Andrzej Sokołowski został przedstawicielem Rady Sekcji SKAD PTS do

IFCS na kadencję 2016–2019, uzyskując następujący wynik: 39 głosów na „tak”, 1 głos na „nie”, 1 głos był nieważny.

W ostatnim punkcie zebrania dyskutowano nad kierunkami rozwoju działalności Sekcji obejmującymi następujące problemy: udział w międzynarodowym ruchu naukowym (wspólne granty, publikacje), umiędzynarodowienie konferencji SKAD (uczestnicy zagraniczni, dwujęzyczność konferencji), wydawanie własnego czasopisma.

Profesor Józef Pociecha zamknął posiedzenie Sekcji SKAD.

Krzysztof Jajuga, Marek Walesiak

Andrzej Dudek

Uniwersytet Ekonomiczny we Wrocławiu
e-mail: andrzej.dudek@ue.wroc.pl

**SOCIAL NETWORK ANALYSIS JAKO GAŁĄŻ
WIELOWYMIAROWEJ ANALIZY STATYSTYCZNEJ**

**SOCIAL NETWORK ANALYSIS AS A BRANCH
OF MULTIDIMENSIONAL STATISTICAL ANALYSIS**

DOI: 10.15611/pn.2016.426.04

Streszczenie: *Social Network Analysis* (SNA, analiza sieci społecznościowych, analiza danych społecznościowych) to zyskująca popularność wraz z rozwojem portali społecznościowych, takich jak Facebook czy Nasza Klasa, grupa metod zajmujących się analizą wzajemnych powiązań i wyszukiwaniem struktur we wspólnotach ludzkich. Zazwyczaj danymi wejściowymi do metod SNA są rzeczywiste dane dotyczące wzajemnych powiązań użytkowników portali społecznościowych. W artykule zostaną przedstawione podstawowe techniki *Social Network Analysis*, modele i miary w niej używane oraz przykłady dla rzeczywistych danych z wykorzystaniem pakietu *sna* środowiska obliczeniowego R, w tym badanie wykorzystujące autorską propozycję połączenia metod SNA z analizą skupień.

Słowa kluczowe: *Social Network Analysis*, sieci społecznościowe, pakiet *sna*.

Summary: Social Network Analysis is a group of methods for the analysis of interrelationships and search structures in the human community, gaining in popularity with the growth of social-networking portals such as Facebook or Twitter. Typically, input data to the SNA methods are the actual data of users of social-networking portals interconnectedness. The paper will present the basic techniques of Social Network Analysis, models and metrics used in it and examples for real data sets using *sna* package of R statistical environment including example of combining *sna* measures with cluster analysis.

Keywords: Social Network Analysis, social networks, *sna* package.

1. Wstęp

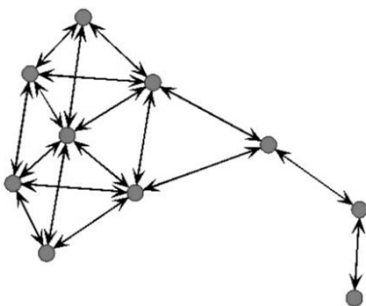
Social Network Analysis (SNA)¹, to grupa metod zajmujących się analizą wzajemnych powiązań i wyszukiwaniem struktur we wspólnotach z wykorzystaniem sieci

¹ Nazwa *Social Network Analysis* nie ma jeszcze ugruntowanego polskiego odpowiednika w krajowej literaturze. Należałoby rozważyć jej tłumaczenia jako: analiza sieci społecznościowych, analiza sieci społecznych, analiza sieci wzajemnych powiązań lub analiza danych społecznościowych.

i elementów teorii grafów. Podwaliny pod tę gałąź nauki położył już w latach 30. XX w. A. Radcliffe-Brown, badając problemy struktury społecznej (por. [Scott 2013, s. 1]). D. Cartwright i F. Harary [1956] do reprezentowania struktur społecznych zaproponowali sieci międzyludzkich relacji pomiędzy członkami grup społecznych. Od lat 70. poprzedniego wieku ta grupa metod zaczęła zyskiwać popularność, co zostało jeszcze wzmocnione przez pojawienie się portali społecznościowych, takich jak Facebook i Twitter. W artykule omówione zostaną podstawowe miary wykorzystywane w SNA, oprogramowanie komputerowe realizujące metody SNA oraz przedstawione dwa przykłady zastosowania SNA dla rzeczywistych danych pochodzących z portali społecznościowych Facebook i Twitter. W drugim miary SNA zostaną potraktowane jako dane wejściowe do analizy skupień.

2. Miary bliskości, pośrednictwa i centralności

Do reprezentacji danych w metodach *Social Network Analysis* służą grafy wzajemnych powiązań (por. rys. 1) lub odpowiadające im tablice powiązań, w których 1 na pozycji i, j oznacza istnienie powiązania (relacji) między obiektem i -tym i j -tym, a 0 oznacza brak powiązania.



Rys. 1. Typowa reprezentacja danych w metodach *Social Network Analysis* – graf powiązań

Źródło: opracowanie własne.

Tabela 1. Typowa reprezentacji danych w metodach *Social Network Analysis* – tablica powiązań

	O ₁	O ₂	O ₃	O ₄	O ₅	O ₆	O ₇	O ₈	O ₉	O ₁₀
O ₁	0	1	1	1	1	1	1	0	0	0
O ₂	1	0	1	1	0	1	0	1	0	0
O ₃	1	1	0	0	1	0	1	1	0	0
O ₄	1	1	0	0	1	1	0	0	0	0
O ₅	1	0	1	1	0	0	1	0	0	0
O ₆	1	1	0	1	0	0	0	0	0	0
O ₇	1	0	1	0	1	0	0	0	0	0
O ₈	0	1	1	0	0	0	0	0	1	0
O ₉	0	0	0	0	0	0	0	1	0	1
O ₁₀	0	0	0	0	0	0	0	0	1	0

Źródło: opracowanie własne.

Miara stopni (wejściowych, wyjściowych) to liczba połączeń (wejściowych/wyjściowych) z/do węzła.

Pośrednictwo (*betweenness centrality*) oznacza stosunek liczby najkrótszych ścieżek przechodzących przez węzeł do liczby innych (nie najkrótszych) ścieżek przechodzących przez węzeł.

Bliskość (*closeness centrality*) stosunek liczby węzłów, do których istnieje połączenie z danego węzła do liczby wszystkich krawędzi w grafie.

Wśród pozostałych miar *Social Network Analysis* wymienić można:

- bliskość i pośrednictwo estymowane aproksymacją logistyczną;
- miarę centralności estymowaną wartościami własnymi (*eigenvector centrality*), czyli miarę wpływu (znaczenia) węzła w sieci definiowaną na podstawie macierzy podobieństw zawierającej długości najkrótszych ścieżek pomiędzy węzłami. W celu obliczenia tej miary stosuje się dekompozycję macierzy podobieństw, wybierając największe wektory własne w sensie twierdzenie Perrona-Frobeniusa;
- przesiąkanie (*percolation centrality*) – miarę, w której oprócz długości najkrótszych ścieżek między węzłami brane są pod uwagę wagi definiujące wpływ jednego węzła na pozostałe;
- miarę centralności Katza (*Katz centrality*) – miarę znaczenia węzła w sieci, biorącą pod uwagę nie tylko najkrótsze ścieżki między węzłem a pozostałymi uczestnikami sieci, lecz także pozostałe ścieżki;
- miarę centralności krzyżowej względem grup (*cross-clique centrality*) – miarę znaczenia węzła w sieci, w której to mierze zamiast najkrótszych odległości od innych uczestników sieci brane są pod uwagę odległości od grup (klik) występujących w grafie powiązań.

3. Oprogramowanie komputerowe

J. Scott [2013] łączy rozwój metod SNA z pojawieniem się oprogramowania komputerowego implementującego te metody. Historycznie wśród najważniejszych aplikacji tego typu wymienia on programy UCINET i PAJEK. Najważniejsze pozostałe programy tego typu wymienione są w tab. 2.

Tabela 2. Oprogramowanie realizujące metody *Social Network Analysis*

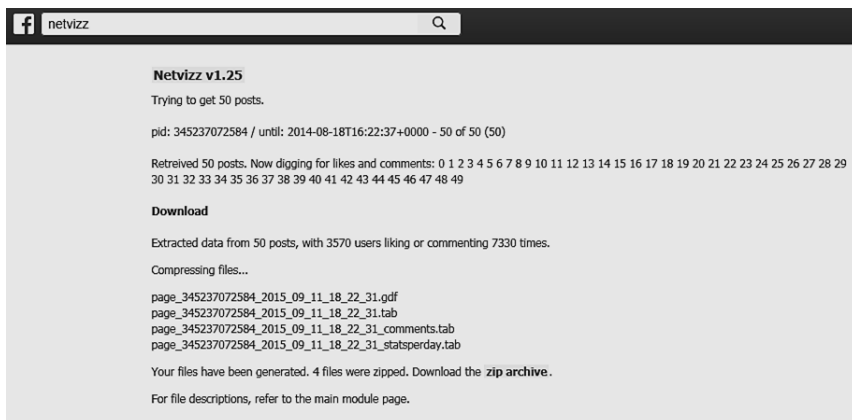
Oprogramowanie	Licencja	System	Format wejściowy	Format wyjściowy
1	2	3	4	5
#k@ http://hashkat.org/	GPL	Mac OS X, GNU/Linux	Plik tekstowy	Gephi, NetworkX
EgoNet http://sourceforge.net/projects/egonet/	Open source	Kompatybilny z językiem Java	Xml	Csv
Commetrix http://www.commetrix.net/	Komercyjny	Windows	Csv, Xml	Csv

1	2	3	4	5
Java Universal Network/Graph (JUNG) http://jung.sourceforge.net/	Open source /BSD	Kompatybilny z językiem Java	GraphML, Pajek, pliki tekstowe	GraphML, Pajek, pliki tekstowe
Network Overview Discovery Exploration for Excel (NodeXL) http://nodexl.codeplex.com/	Free MS-PL	Windows	GraphML, Pajek, Csv, Xls	GraphML, Csv, Xls
NetMiner http://www.netminer.com/	Komercyjny	Windows	Xml, Pajek, Csv, Gml	Xml, Pajek, Csv, Gml
R pakiety: sna Igraph PAFit network tnet ergm Bergm hergm latentnet degreenet networksis	GPL	Linux/Unix MacOS Windows	większość istniejących formatów	większość istniejących formatów
socilyzer https://socilyzer.com/	Komercyjny	Linux/Unix MacOS Windows / przez www	Własny moduł wprowadzania danych	Csv, Vna

Źródło: opracowanie własne.

4. Integracja z serwisem Facebook – przykład

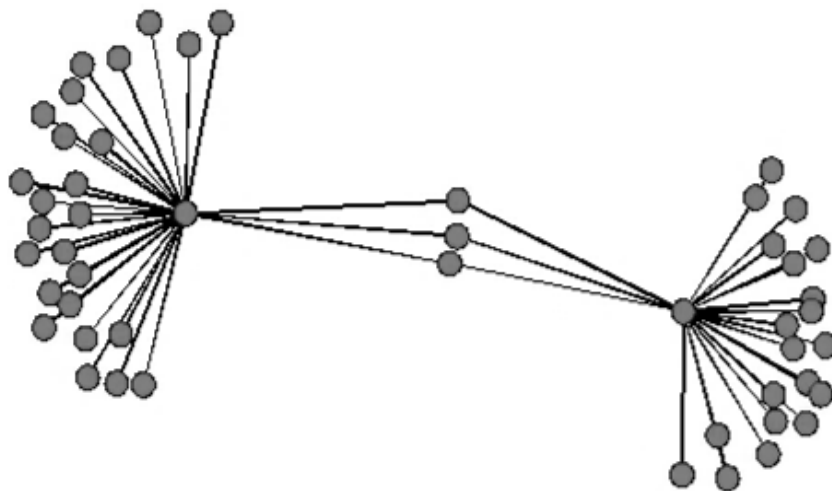
W badaniu przeanalizowano połączenia wiadomości (postów) 26 użytkowników portalu Facebook w pojedynczym wątku dyskusji. Dane zostały wyeksportowane za pomocą aplikacji modułu *NetWizz* portalu Facebook.



Rys. 2. Moduł *NetWizz* portalu Facebook

Źródło: opracowanie własne.

Z powodu konieczności zapewnienia ochrony danych osobowych nazwiska użytkowników zostały zastąpione przez moduł eksportujący ich identyfikatory w portalu. Graf powiązań użytkowników przedstawia rys. 3.



Rys. 3. Graf powiązań użytkowników w analizowanym wątku dyskusji

Źródło: opracowanie własne.

Tabela 3. Miary SNA dla analizowanych danych w wątku dyskusji na portalu Facebook

	Stopnie	Pośrednictwo	Bliskość
04966b5d50c213c2f989edcc09aad062c9890dc	23	-0,881201637	-0,1226
c8fce9477fe3cac4a5f24049384085031ec36ede	1	-19,50352993	-22,5661
42248d77a11cb9df6c275a305639c3477bb43bbd	1	-19,50352993	-22,5661
9e7ce9e7454fd3f914afc695c56aa63e9ae02c25	1	-19,50352993	-22,5661
93dee1bc36c95f0fd7e7db1fc0ecce4b581dd92	1	-19,50352993	-22,5661
99dab9e9b6d108c6088766eeab8d64b74377d5a	1	-19,50352993	-22,5661
...	1	-19,50352993	-22,5661
d8da3d437418492ace42a42dc3d7644d85512b42	1	-19,50352993	-22,5661
43d9175d3056f2ff1fcc7fc891f92b738c7afa09	1	-19,50352993	-22,5661
43b6f1742fe283761c2193ee497f40a80b8ff174	2	-19,37680269	-22,5661
4845889042620f2fa39a3baf8690d34197d0b58e	2	-19,37680269	-22,5661
b7d16c9c2b9cad456c51536b83c1765a0268939a	1	-19,50352993	-22,5661
98845bcce50aabc2edf2d4409e872eae001aec1a	1	-19,50352993	-22,5661
41201acaf89a53937c7474b55d6052fc5d77d354	1	-19,50352993	-22,5661
da5168673d53bbb31b7274eb25538465a4f61858	2	-19,37680269	-22,5661
c8f5fd611553eb786d025532e137265900a9848b	1	-19,50352993	-22,5661
8710a6152d01a2a64d722fd0be14938e66348500	1	-19,50352993	-22,5661
1d80aad2b0c8ef1eb9a21afa3634cd5ff636656	1	-19,50352993	-22,5661
c9cc319245ebda30b7024f4897e995c3199df7a	28	-0,733472401	0,287682
6200c42cb6adcf4dc8cea6a0ce93a5460112b03e	1	-19,56682679	-22,5661

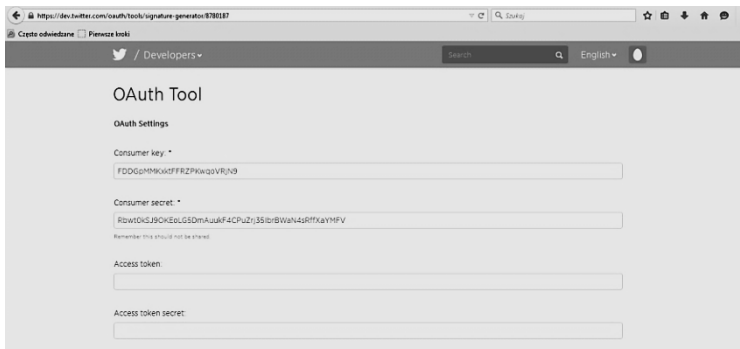
Źródło: opracowanie własne.

W tabeli 3 znajdują się liczby stopni oraz miary pośrednictwa i bliskości aproksymowanych regresją logistyczną dla poszczególnych węzłów.

Miary wskazują wyróżniające się znaczenie osób c9cc3192455ebda30b-7024f4897e995c3199df7a i 04966b5d50c213c2f989eddc09aad062c9890dc, których miary pośrednictwa są zdecydowanie wyższe niż pozostałych oraz osób 43b6f1742fe283761c2193ee497f40a80b8ff174, 4845889042620f2fa39a3-baf8690-d34197d0b58e i da5168673d53bbb31b7274eb25538465a4f61858.

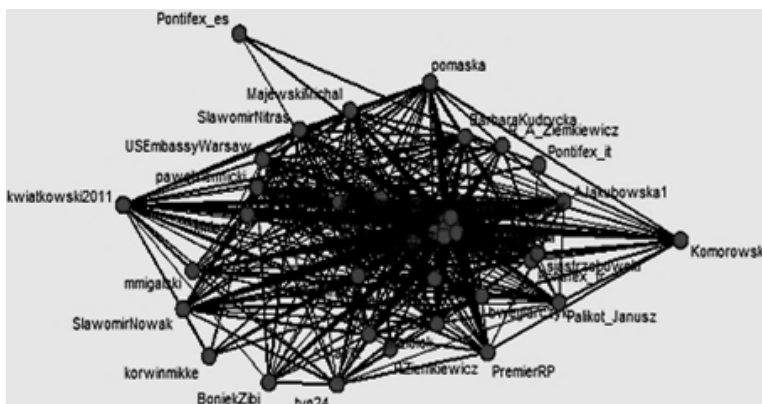
5. Integracja z serwisem Twitter – przykład

W celu integracji z portalem Twitter konieczne było założenie konta typu developer, pobranie klucza prywatnego i publicznego oraz właściwe pobranie danych z wykorzystaniem skryptu w języku *Python*.



Rys. 4. Export danych SNA z portalu Twitter

Źródło: opracowanie własne.



Rys. 5. Graf powiązań użytkowników portalu Twitter subskrybowanych przez konto testowe

Źródło: opracowanie własne.

Analizowane dane dotyczyły stu wybranych osób publicznie znanych, polityków, publicystów, komentatorów sportowych (oraz konto *pontifex* – papieskie). Za powiązanie w grafie uznawano to, iż jedna osoba jest subskrybentem (*follower*) drugiej. Powiązania więc nie musiały być symetryczne. Rysunek 5 przedstawia graf powiązań.

Tabela 4 przedstawia liczbę połączeń, stopnie pośrednictwa, bliskości oraz miarę centralności estymowaną wektorami własnymi dla poszczególnych węzłów.

Tabela 4. Miary SNA dla analizowanych połączeń portalu *Twitter*

	Stopnie	Pośrednictwo	Bliskość	Wektory własne
kwiatkowski2011	85	14,5502	0,145055	0,243622
SlawomirNowak	64	10,18665	0,093549	0,182322
ProtasiewiczJ	84	14,15217	0,146735	0,494696
Pontifex_pl	31	84	1,56E-13	-0,30538
R_A_Ziemkiewicz	33	0,778215	0,022143	-0,36772
TomaszSiemoniak	90	22,17916	0,159017	0,494696
Pontifex_es	6	0	1,56E-13	-3,04452
Pontifex_fr	6	0	1,56E-13	-3,04452
trzaskowski_	74	7,365014	0,128951	0,243622
pawlakwaldemar	31	0	0	-0,12136
Polsport	38	8,813546	0,024595	-0,24362
skorzynski	87	15,36686	0,15452	0,625706
KrzysztofLisek	89	21,52915	0,176369	0,060625
LeszekMiller	86	19,51107	0,155669	0,494696
Rastastacja	0	0	0	-16,5661
bweglarczyk	70	9,714886	0,115375	0,367725
agielewska	78	6,847203	0,152939	0,121361
tvn24	38	0	0	0,305382
kmleski	51	5,130286	0,091717	-0,4947
KurskiPL	79	13,29188	0,138935	0,182322
gnapieralski	73	7,702563	0,14131	0,060625
michalkobosko	86	19,04573	0,1529	0,559616
Platforma_org	76	12,12626	0,143585	0,182322
Palikot_Janusz	73	13,14377	0,119717	0,430783
jbrudzinski	69	5,128856	0,130415	-0,18232
KonradPiasecki	99	45,17611	0,164498	1,058607
MajewskiMichal	83	28,60656	0,142123	0,693147
TygodnikWPROST	86	52,33176	0,160943	0,243622
Komorowski	36	0,954358	0,023086	-0,24362
Vevo_PL	7	2,290147	0,006379	-3,04452
mackab	0	0	0	-16,5661
jmiziolek	84	10,00179	0,173692	0,121361

Źródło: opracowanie własne.

Dane z tab. 4 były punktem wyjścia do procedury analizy skupień. Zastosowano metodę *partitioning around medoids* [Kaufman Rousseeuw 1990]. Do ustalenia liczby klas wykorzystano index GAP [Tibshirani, Walther, Hastie 2001].

Otrzymano następujący podział na skupienia:

Klasa 1: pontifex_pl, borek_mati;

Klasa 2: SlawomirNowak, trzaskowski, bweglarczyk, agielewska, kmlesk, gnapieralski, Platforma_org, Palikot_Janusz, jbrudzinski, PopPolityk, nmaliszewski, gazeta_wyborcza, Piechocinski, PremierRP, AzraelK, AndZyberto, ArlenaSokalska, BarbaraKudrycka, RyszardKalisz, lis_tomasz, USEmbassyWarsaw, AJakubowski, mmigalski

Klasa 3: R_A_Ziemkiewicz, pawlakwaldemar, Polsport, Rastastacja, tvn24, Komorowski, BoniekZibi,, korwinmikke, KR_PL, MTVPolska, InstytutHH

Klasa 4: kwiatkowski2011, ProtasiewiczJ, TomaszSiemoniak, skorzynski, KrzysztofLisek, LeszekMiller, KurskiPL, michalkobosko, KonradPiasecki, MajewskiMichal, TygodnikWPROST, jmiziolek, michal_kolanko, GoskaSerafin, Jaroslaw_Gowin, johngodson, AdamBielan, tvp_info, dominikuhlig, hytrekprosiecka, sjastrzebowski, michalkarnowski, SlawomirNitras, pomaska, pawelsienicki,

Przy tym charakterystyki poszczególnych skupień (średnie miar) przedstawiają się następująco:

[1]	36.50000000	71.82116794	0.02834481	-0.56914539
[2]	87.9200000	20.3722848	0.1619148	0.3786684
[3]	65.739130435	9.831676243	0.115588135	0.006643192
[4]	17.705882353	0.907175945	0.005442009	-3.277094996

Z pewną rezerwą można uznać, że skupienie pierwsze, mające bardzo wysoki średni współczynnik pośrednictwa, to konta łączące różne grupy użytkowników. Polityków należących do skupienia drugiego należy uznać za istotnych uczestników życia politycznego, polityków i publicystów ze skupienia trzeciego za mniej znaczących, choć w dalszym ciągu prominentnych, a polityków ze skupienia ostatniego za „zachodzące gwiazdy”.

6. Zakończenie

W artykule przedstawiono zastosowania metod *Social Network Analysis* analizy sieci społecznościowych dla rzeczywistych zbiorów danych. Metody te mogą znaleźć zastosowanie w naukach ekonomicznych np. w:

- analizie danych dotyczących zależności pomiędzy graczami na rynkach,
- analizie zależności pomiędzy kontrahentami oraz odbiorcami towarów,
- konstruowaniu scenariuszy odejścia kluczowych kontrahentów,
- znajdowaniu liderów opinii w sieciach społecznościowych.

Popularność metod *Social Network Analysis* warunkowana jest coraz większą dostępnością narzędzi softwarowych umożliwiających dokonywanie odpowiednich analiz oraz coraz większą dostępnością danych i faktu, że wiele popularnych portali społecznościowych, takich jak Facebook czy Twitter, umożliwia eksport danych dla metod analizy sieci społecznościowych.

Metody analizy sieci społecznościowych mogą być również używane jako metody wstępne do przygotowania danych dla innych metod wielowymiarowej analizy statystycznej.

Literatura

- Butts C.T., 2014., *sna: Tools for Social Network Analysis. R package version 2.3-2.*, <http://CRAN.R-project.org/package=sna>.
- Cartwright D., Harary F., 1956, *Structural balance: A generalization of Heider's theory*, *Psychological Review*, vol. 63, s. 277–293.
- Kaufman L., Rousseeuw P.J., 1990, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York.
- Scott J., 2013, *Social Network Analysis*, Sage, London.
- Tibshirani R., Walther G., Hastie T., 2001, *Estimating the number of clusters in a data set via the gap statistic*, *Journal of the Royal Statistical Society, ser. B*, vol. 63, part 2, s. 411–423.
- Wasserman S., Faust K., 1994, *Social Network Analysis: Methods and Applications*, Cambridge: Cambridge University Press.