Uniwersytet Ekonomiczny w Katowicach

Studia Ekonomiczne. Zeszyty Naukowe
Uniwersytetu Ekonomicznego w Katowicach
ISSN 2083-8611 Nr 368 · 2018

Barbara Probierz

Uniwersytet Ekonomiczny
Wydział Informatyki i Komunikacji
Katedra Inżynierii Wiedzy
barbara.probierz@ue.katowice.pl

Jan Kozak

Uniwersytet Ekonomiczny
Wydział Informatyki i Komunikacji
Katedra Inżynierii Wiedzy
jan.kozak@ue.katowice.pl

MECHANIZM PREDYKCJI FOLDERÓW W AUTOMATYCZNEJ KLASYFIKACJI WIADOMOŚCI E-MAIL W SKRZYNKACH POCZTOWYCH

Streszczenie: W tym artykule zaproponowana została nowa metoda sugerowania tworzenia folderów w skrzynkach pocztowych użytkowników poprzez zastosowanie sieci społecznych i algorytmów mrowiskowych. Celem pracy jest stworzenie mechanizmu predykcji nowych folderów w automatycznej klasyfikacji wiadomości e-mail do folderów w skrzynkach pocztowych. W proponowanym algorytmie zastosowano elementy analizy sieci społecznych służące do wyznaczania grup użytkowników posiadających podobną strukturę folderów w skrzynkach pocztowych, na podstawie których oparto mechanizm sugerowania zakładania nowych folderów dla użytkowników. Działanie proponowanej metody zostało przetestowane na ogólnodostępnym zbiorze danych Enron E-mail.

Słowa kluczowe: Enron E-mail, sieci społeczne, predykcja folderów, algorytmy mrowiskowe.

JEL Classification: C88.

Wprowadzenie

Historia powstania wiadomości e-mail zaczęła się prawie pół wieku temu, kiedy to w 1965 r. L. Pouzin, G. Schroeder i P. 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 R.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 składnie dla komunikacji e-mailowej: użytkownik@host, która funkcjonuje do dziś.

Obecnie każdego dnia przez Internet wysyłane są miliardy wiadomości, a w ciągu roku jest ich ponad 100 bilionów. Typowy użytkownik dostaje przeciętnie 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. Jednocześnie duża część przesyłanych wiadomości zawiera informacje zbędne, które powinny być filtrowane.

Celem pracy jest stworzenie sieci społecznej opartej na kontaktach pomiędzy nadawcami a odbiorcami wiadomości e-mail. Następnie podczas analizy i obserwacji sieci społecznej wyodrębniane są grupy użytkowników posiadających podobną strukturę folderów w skrzynkach pocztowych, na podstawie której opracowywany jest mechanizm sugerowania zakładania nowych folderów w skrzynkach pocztowych użytkowników. Proponowana metoda została zastosowana na zbiorach wiadomości e-mail z wybranych skrzynek pocztowych uzyskanych z ogólnodostępnego zbioru danych Enron E-mail.

1. Problem przypisania wiadomości do folderów

Jednymi z największych problemów użytkowników, zwłaszcza tych, dla których e-mail to podstawa komunikacji, są odpowiednie uporządkowanie poczty elektronicznej i przypisanie wiadomości do poszczególnych folderów. Szczególnie gdy proces ten ma się odbywać w sposób automatyczny. Z tego powodu jest coraz większe zainteresowanie tworzeniem systemów, które w sposób automatyczny mogą pomóc użytkownikom w zarządzaniu pocztą elektroniczną.

Niestety problem automatycznego sugerowania tworzenia folderów oraz przypisywania do nich wiadomości e-mail jest problemem bardzo spersonalizowanym, gdyż zależy od indywidualnych upodobań osób kontaktujących się ze sobą. Odzwierciedlenie tych upodobań można przedstawić w postaci sieci społecznych, których analiza pozwala lepiej zrozumieć zachowania użytkowników poczty elektronicznej. Natomiast poprzez zastosowanie algorytmów mrowiskowych możliwe jest poszukiwanie większej części przestrzeni rozwiązań z zastosowaniem eksploracji i eksploatacji.

2. Zbiór wiadomości elektronicznych Enron

Enron E-mail Dataset jest to zestaw danych zebrany i przygotowany przez Projekt CALO (*A Cognitive Assistant that Learns and Organizes*). Zawiera ponad 600 tys. wiadomości e-mail, które zostały wysłane lub odebrane przez 150 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 [Enron, b.r.].

Dzięki pracy zespołu z ośrodka SRI International, zwłaszcza Melinda Gervasio, dane zostały poprawione i udostępnione innym naukowcom do badań. Wiadomości te są przypisane do kont osobistych i podzielone na foldery. W zbiorze danych nie ma załączników do wiadomości e-mail, a niektóre wiadomości zostały usunięte ze względu na występowanie duplikatów w innych folderach. Brakujące informacje zostały w miarę możliwości uzupełnione na podstawie innych treści, jednak w przypadku, gdy nie było możliwości określenia odbiorcy, wprowadzono frazę no_address@enron.com.

Każda skrzynka pocztowa pracowników firmy Enron Corporation jest przechowywana w osobnym folderze i oznaczona nazwiskiem danego pracownika. W każdej skrzynce są foldery utworzone automatycznie przez system pocztowy (np. *sent mail*, *all documents*, *deleted items*) oraz foldery utworzone przez użytkowników. Wewnątrz tych folderów są kolejno ponumerowane wiadomości e-mail.

Wszystkie wiadomości w zbiorze Enron E-mail Dataset mają jednakową budowę. Są to pliki tekstowe zawierające w kolejnych liniach szczegółowe informacje, tj. identyfikator wiadomości, data wysłania, adres pocztowy nadawcy, adres pocztowy odbiorcy, temat wiadomości, odbiorcy, do których wysłano kopie wiadomości, imię i nazwisko nadawcy wiadomości, imię i nazwisko odbiorcy wiadomości, nazwa folderu, w którym jest wiadomość, nazwa skrzynki pocztowej, w której jest wiadomość, treść wiadomości.

Tabela 1. Parametry wybranych skrzynek pocztowych ze zbioru danych Enron E-mail

Skrzynka pocztowa	Liczba obiektów	Liczba Klas	Liczba wiadomości w folderze		
			średnia	minimalna	maksymalna
lokay-m	2493	11	226,64	6	1159
sanders-r	1188	30	39,60	4	420
shackleton-s	1001	53	18,89	3	259
steffes-j	625	23	27,17	3	317
symes-k	770	12	64,17	3	254
williams-w3	2769	18	153,83	3	1398
farmer-d	3672	25	146,88	5	1192
beck-s	1971	101	19,51	3	166

Źródło: Opracowanie własne.

W tabeli 1 przedstawione zostały parametry dotyczące wybranych skrzynek pocztowych ze zbioru Enron E-mail. Zawarte są tam dane dotyczące liczby folderów oraz liczby wiadomości e-mail zawartych w skrzynkach pocztowych, jak również dane statystyczne, dotyczące występowania wiadomości w folderach.

3. Sieci społeczne

Sieć społeczna (ang. *Social Network*) to wielowymiarowa struktura złożona ze zbioru jednostek społecznych oraz połączeń miedzy nimi. Jednostki społeczne to osoby funkcjonujące w danej sieci, natomiast połączenia odwzorowują różnorodne relacje społeczne pomiędzy poszczególnymi osobami. Pierwsze badania sieci społecznych przeprowadził w 1923 r. Jacob L. Moreno, który uznawany jest za jednego z założycieli dyscypliny analizy sieci społecznych [Moreno, 1934]. Jest to gałąź socjologii, która zajmuje się ilościową oceną roli jednostki w grupie lub społeczności przez analizę sieci powiazań miedzy 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 [por. Moreno, 1978].

Sieć społeczną przedstawia się w postaci grafu, takiego, że:

$$\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E}) \tag{1}$$

gdzie V jest skończonym zbiorem wierzchołków grafu $V = \{1, \ldots, n\}$, przy czym $\overline{\overline{V}} \geq 2$, natomiast E jest skończonym zbiorem wszystkich dwuelementowych podzbiorów zbioru V zwanych krawędziami, łączącymi poszczególne wierzchołki, takim, że:

$$\boldsymbol{E} \subseteq \{\{\boldsymbol{u}, \boldsymbol{v}\} : \boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{V}, \boldsymbol{u} \neq\} \tag{2}$$

Krawędzie w sieci społecznej reprezentują interakcje, przepływ informacji i dóbr, podobieństwo, afiliacje lub związki społeczne. Miarami siły powiazania są częstotliwość, wzajemność oraz rodzaj interakcji lub przepływu informacji, ale także siła powiazania, zależna od atrybutów łączonych węzłów (np. stopień pokrewieństwa) oraz struktury sąsiedztwa tych węzłów (np. liczba wspólnych sąsiadów).

Głównymi wskaźnikami charakteryzującymi daną sieć społeczną są stopnie wierzchołków oraz centralność według tych stopni. Stopień wierzchołka *v* (stopień wejściowy, stopień wyjściowy) to liczba krawędzi wchodzących lub wychodzących z danego węzła i określana jest wzorem:

$$\deg(v) = \sum_{u=1}^{n} k_{v,u} \tag{3}$$

gdzie $k_{v,u}$ to krawędź miedzy wierzchołkiem *v*, a wierzchołkiem *u*.

Natomiast centralność według stopni wierzchołków służy do określania, które węzły są kluczowe z punktu widzenia rozprzestrzeniania informacji lub wpływania na węzły położone w bezpośrednim sąsiedztwie. Najczęściej centralnym wierzchołkiem określa się te wierzchołki, które mają najwięcej relacji z innymi wierzchołkami (posiadają największą liczbę krawędzi). Maksymalny stopień wierzchołka *v* w sieci *G* określony jest wzorem:

$$\Delta(\mathbf{G}) = \mathbf{max}\{\mathbf{deg(v)}: \mathbf{v} \in \mathbf{V(G)}\} \tag{4}$$

Wśród dodatkowych wskaźników charakteryzujących sieć społeczną wyróżnić można takie wskaźniki, jak promień, średnia odległość w sieci czy grupowanie obiektów. Promień r_{cv} (ang. *radius*) wierzchołka *v*:

$$rc_v = \frac{1}{max_{v \in V} d_{v,u}} \tag{5}$$

gdzie $d_{v,u}$ to długość najkrótszej drogi w grafie *G* między wierzchołkami *v* oraz *u* (długość drogi w grafie między wierzchołkami *v* oraz *u* równa jest liczbie krawędzi na drodze z *v* do *u*). Najwyższą ocenę uzyskuje wierzchołek, który jest możliwie najbliżej wszystkich najbardziej wysuniętych wierzchołków sieci (odległość dzieląca go od najdalszego wierzchołka jest najmniejsza).

Średnia odległość *L* (średnia długość dróg najkrótszych) w sieci:

$$L = \frac{\sum_{u \neq v \in V} d_{v,u}}{V(n-1)} \tag{6}$$

gdzie $d_{v,u}$ to długość najkrótszej drogi w grafie *G* miedzy wierzchołkami *v* oraz *u*.

Sposób grupowania (ang. *Clustering*) interpretowany jest jako rozmieszczenie blisko siebie obiektów, w jaki sposób powiązanych ze sobą, a powstała struktura określana jest jako klaster lub grono. Prawdopodobieństwo, że najbliżsi sąsiedzi wierzchołka v są również swoimi najbliższymi sąsiadami, określa współczynnik gronowania $g_{c,v}$ wierzchołka v takim, że:

$$g_{cv} = \frac{2E_v}{k_v(k_v - 1)}, kv > 1 \quad (7)$$

gdzie E_v to liczba krawędzi k_v między sąsiadami wierzchołka v. Jest to iloraz liczby krawędzi pomiędzy sąsiadami danego wierzchołka do liczby krawędzi, jaki miałby graf pełny składający się z tych sąsiadów. Współczynnik gronowania (klasteryzacji) służy do szacowania, ilu sąsiadów danego wierzchołka jest połączonych każdy z każdym [Zhang i in., 2008].

4. Algorytmy mrowiskowe

Algorytmy mrowiskowe (ang. *Ant Colony Optimization* – ACO) stanowią metaheurystyczne podejście do rozwiązywania wielu problemów optymalizacyjnych, wykorzystując idee zachowań komunikacyjnych występujące w koloniach mrówek. Wiele gatunków mrówek znajduje najkrótszą ścieżkę między mrowiskiem a źródłem pokarmu za pomocą feromonu jako mechanizmu komunikacji.

Podczas poszukiwania pożywienia mrówki tworzą ścieżki, na których odkładają ślad feromonowy. Pozwala im to na szybki powrót do mrowiska i przekazanie informacji innym mrówkom o miejscu, w którym znajduje się pożywienie. Stężenie feromonów na ścieżce wpływa na wybór dokonywany przez mrówki. Im większe stężenie feromonów, tym bardziej atrakcyjna ścieżka. Ostatecznie dzięki oddziaływaniu sprzężenia zwrotnego tworzone są najkrótsze ścieżki łączące mrowisko z pożywieniem, na których odłożona jest duża wartość śladu feromonowego.

Pierwszą inspiracją do powstania algorytmów mrowiskowych (ACO) była chęć poznania, w jaki sposób mrówki są w stanie odnaleźć najkrótszą drogę z mrowiska do pożywienia [Verhaeghe, Deneubourg, 1983]. Badania i eksperymenty wykonywane przez S. Goss, J.L. Deneubourg i innych, opisane w pracach [Beckers i in., 1989; Goss i in., 1990], dotyczące zrozumienia sposobu realizacji tego zadania przez naturę, były pierwszym krokiem do zaimplementowania tego rozwiązania w algorytmice. Jednak dopiero podjęte przez M. Dorigo [Dorigo, Maniezzo, Colorni, 1996; Dorigo, Di Caro, 1999; Dorigo, Di Caro, Gambardel-

la, 1999] próby stworzenia sztucznego systemu mrówkowego oraz zastosowania go do znalezienia najkrótszej drogi pomiędzy wierzchołkami dla zadanego grafu były kluczowym krokiem do powstania algorytmów ACO [Dorigo, Birattari, Stutzle, 2006; Dorigo i in., red., 2008; Doerner, Merkle, Stutzle, 2009].

Jednym z algorytmów mrowiskowych stosowanych w eksploracji danych jest algorytm mrowiskowy do konstruowania drzew decyzyjnych (ang. *Ant Colony Decision Tree* – ACDT). Algorytm ten łączy idee algorytmów mrowiskowych oraz algorytmu CART, dzięki czemu osiąga jakościowo dobre klasyfikatory dla wielu standardowych problemów z dziedziny eksploracji danych [Boryczka, Kozak, 2010].

Wykonywanie algorytmu polega na wyborze testu dla każdego węzła na podstawie dwóch czynników. Pierwszym czynnikiem jest maksymalna wartość zgodna z kryterium podziału algorytmu CART, a drugim dodatkowa informacja zapisana w postaci śladu feromonowego [Boryczka, Kozak, 2011]. Podczas pracy algorytmu każdy agent-mrówka w populacji konstruuje drzewo decyzyjne.

Feromon odkładany jest dla każdego podziału wybranego podczas konstruowania drzewa wraz z informacją o podziale dokonanym w węźle nadrzędnym. Dzięki takiemu zastosowaniu algorytm stara się budować kolejne drzewa z uwzględnieniem struktury poprzednich drzew, modyfikując pojedyncze węzły. Wynikiem pracy algorytmu jest najlepsze drzewo decyzyjne.

Wartość funkcji heurystycznej wyznaczana jest na podstawie kryterium podziału stosowanego w algorytmie CART, zgodnie ze wzorem:

$$\operatorname{argmax}_{a_j \le a_j^R, j=1,\ldots,M} \left(\frac{P_l P_r}{4} \left[\sum_{k=1}^{K} |p(k|m_l) - p(k|m_r)| \right]^2 \right) \tag{8}$$

gdzie:
$p(k|m_l)$ – prawdopodobieństwo wystąpienia klasy decyzyjnej k w węźle m_l,
$p(k|m_r)$ – prawdopodobieństwo wystąpienia klasy decyzyjnej k w węźle m_r,
P_l – prawdopodobieństwo przejścia obiektu do węzła m_l (lewego poddrzewa),
P_r – prawdopodobieństwo przejścia obiektu do węzła m_r (prawego poddrzewa),
K – klasy decyzyjne.

Natomiast prawdopodobieństwo wyboru testu w węźle jest standardowym prawdopodobieństwem wykorzystywanym w systemach mrowiskowych, obliczanym zgodnie ze wzorem:

$$p_{i,j} = \frac{\tau_{m,m_{L(i,j)}}(t)^{\alpha} \cdot \eta_{i,j}^{\beta}}{\sum_i^a \sum_j^{b_i} \tau_{m,m_{L(i,j)}}(t)^{\alpha} \cdot \eta_{i,j}^{\beta}} \tag{9}$$

gdzie:
$\eta_{i,j}$ – współczynnik informacji heurystycznej dla testu atrybutu *i* o wartości *j*,
$\tau_{m,m_{L(i,j)}}$ – ślad feromonowy w czasie *t* dla krawędzi prowadzącej od węzła *m* do węzła m_L (dla testu atrybutu *i* o wartości *j*),
α i β – parametry określające względną wagę funkcji heurystycznej i śladu feromonowego.

Najważniejszymi regułami zachowania agenta-mrówki są reguły aktualizacji śladu feromonowego i funkcja przejścia między stanami. Każda decyzja dotycząca wyboru kolejnego kroku podejmowana jest przez sztuczną mrówkę zgodnie ze wzorem:

$$j = \begin{cases} \arg max_{r \in J_i^k} \left\{ [\tau_{i,r}(t)] \cdot \left[\eta_{i,r}\right]^{\beta} \right\}, jeśli\ q \leq q_0 \\ p_{i,j}^k(t), \text{ w przeciwnym razie} \end{cases} \quad (10)$$

gdzie:
$\eta_{i,r}$ – wartość heurystycznie oszacowanej jakości przejścia ze stanu i do stanu *r*,
$\tau_{i,r}$ – wartość nagrody, czyli stopień użyteczności branej pod uwagę decyzji,
β – parametr określający ważność wartości $\eta_{i,r}$,
$p_{i,j}^k(t)$ – kolejny krok (decyzja).

5. Mechanizm predykcji folderów

Zaproponowany mechanizm umożliwia sugerowanie zakładania nowych folderów dla użytkowników na podstawie struktury folderów innych użytkowników wyznaczonych przez stworzoną sieć społeczną. Proponowana metoda oparta jest na analizie macierzy śladu feromonowego tworzonej podczas klasyfikowania wiadomości do folderów.

Same mechanizmy sugerowania przypisywania wiadomości do nowych folderów nie są obecnie nowością, gdyż są w praktyce stosowane w niektórych systemach pocztowych. Jednak należy zwrócić uwagę przede wszystkim na ich zawężoną tematykę i sposób działania. Nowe foldery dotyczą w szczególności wiadomości generowanych automatycznie lub rozpoznawanych za pomocą programu pocztowego jako wiadomości związanych z forami dyskusyjnymi, ofertami handlowymi czy serwisami społecznościowymi. Nie sposób jednak znaleźć algorytmów, przy pomocy których możliwe byłoby sugerowanie bardziej nietypowych folderów dla wiadomości, które nie są generowane automatycznie. Au-

torska metoda nie tylko związana jest z sugestią nowych folderów, ale dodatkowo bazuje na możliwościach, jakie dają algorytmy mrowiskowe oraz sieci społeczne.

Algorytm do automatycznego przypisywania wiadomości do folderów wraz z mechanizmem sugerującym użytkownikom tworzenie nowych folderów w swoich skrzynkach pocztowych polega na:

- przeprowadzeniu analizy dotychczas odebranych wiadomości e-mail pod względem kontaktów użytkowników,
- stworzeniu sieci społecznej opartej na kontaktach pomiędzy nadawcą a odbiorcami wiadomości (rys. 1 – krok 1),
- wyodrębnieniu grupy użytkowników posiadających podobną strukturę społeczną na podstawie analizy i obserwacji sieci społecznej (rys. 1 – krok 2),
- przetworzeniu zbioru danych do postaci tabeli decyzyjnej w obrębie danej grupy (rys. 1 – krok 3),
- zastosowaniu algorytmu opartego na rozwiązaniach znanych z algorytmów mrowiskowych (rys. 1 – krok 4),
- przedstawieniu mechanizmu predykcji folderów dla użytkowników na podstawie analizy macierzy klasyfikacji wiadomości do folderów (rys. 1 – kroki 5 i 6).

Zasadniczym aspektem jest w tym przypadku wyodrębnienie grupy kontaktów dla użytkownika, któremu mają zostać zasugerowane nowe foldery. W tym celu, zgodnie z utworzoną siecią kontaktów, należy ustalić najbliższych sąsiadów tego użytkownika (traktowanego jako użytkownika centralnego), a następnie na podstawie preferencji tych użytkowników dokonać sugestii stworzenia nowych folderów. Utworzone grupy kontaktów użytkownika kluczowego z najbliższymi sąsiadami przedstawiono w tabeli 2.

Główna idea rozwiązania bazuje na analizie wspólnej macierzy śladu feromonowego dla wszystkich użytkowników w grupie. W klasycznej wersji proponowanego algorytmu, opisanego w pracy [Boryczka, Probierz, Kozak, 2016], pomimo zastosowania grupy użytkowników, jako dostępne wartości atrybutu decyzyjnego dopuszczalne są jedynie te, które pierwotnie występują u użytkownika, dla którego wykonywana jest predykcja. Wiąże się to m.in. z tym, że wszystkie wiadomości, które pozostali użytkownicy przechowują we własnych, unikalnych względem użytkownika centralnego folderach, zostają pominięte.

Tabela 2. Wybrane grupy użytkowników

Nazwa grupy	Kluczowy użytkownik	Najbliżsi sąsiedzi kluczowego użytkownika
Grupa 1	lokay-m	hyatt-k, mcconnell-m, schoolcraft-d, scott-s, watson-k
Grupa 2	sanders-r	cash-m, dasovich-j, haedicke-m, kean-s, sager-e, steffes-j
Grupa 3	shackleton-s	jones-t, mann-k, stclair-c, taylor-m, ward-k, williams-j
Grupa 4	steffes-j	dasovich-j, gilbertsmith-d, presto-k, sanders-r, shapiro-r
Grupa 5	symes-k	scholtes-d, semperger-c, williams-w3
Grupa 6	williams-w3	mann-k, semperger-c, solberg-g, symes-k
Grupa 7	farmer-d	bass-e, beck-s, griffith-j, nemec-g, perlingiere-d, smith-m
Grupa 8	beck-s	buy-r, delainey-d, hayslett-r, kaminski-v, kitchen-l, may-l, mcconnell-m, shankman-j, white-s

Źródło: Opracowanie własne.

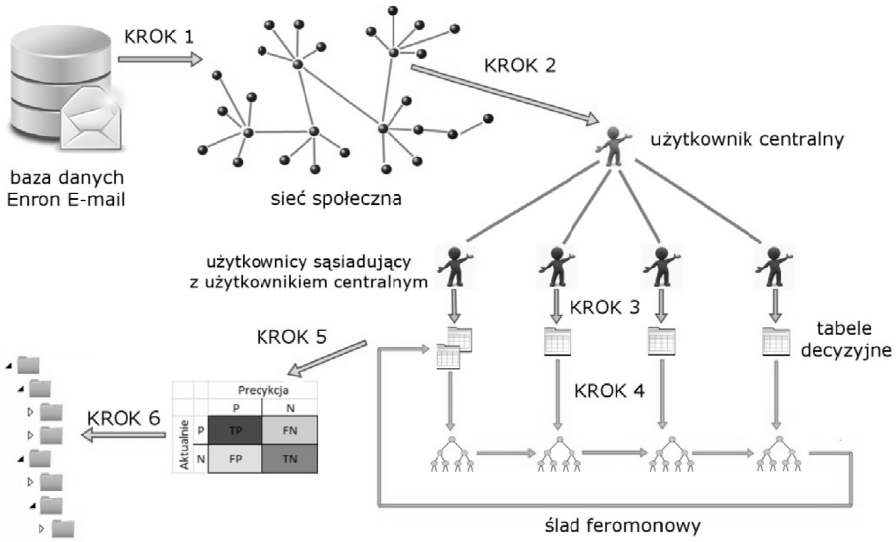
W tym przypadku dopuszczalne wartości atrybutu decyzyjnego są sumą wartości atrybutów decyzyjnych wszystkich użytkowników w grupie (nie tylko centralnego), zgodnie ze wzorem:

$$D = D1 \cup D2 \cup \ldots \cup Dn \tag{11}$$

gdzie:
D_i – zbiór wartości atrybutów decyzyjnych i-tego użytkownika,
n – liczba użytkowników w grupie.

W znacznym uproszczeniu można stwierdzić, że jeśli wiadomość jest o zbliżonych cechach (atrybutach), a pozostali użytkownicy w grupie będą ją przechowywali w folderze, którego użytkownik centralny nie ma, to zostanie mu zasugerowane utworzenie nowego folderu. Jak można zauważyć, w tym przypadku duże znaczenie ma wstępne przetworzenie danych i dostosowanie nazw folderów do zbliżonych, aby różnice wynikające np. z zapisu nazwy folderu nie sugerowały różnicy pomiędzy folderami.

Schemat działania proponowanego algorytmu z mechanizmem predykcji folderów został przedstawiony na rysunku 1. Na podstawie algorytmu opisanego w pracy [Boryczka, Probierz, Kozak, 2016], po przejściu przez kroki 1-4, otrzymywany jest najlepiej zbudowany klasyfikator, którego działanie weryfikowane jest na podstawie danych testowych. Podczas pracy algorytmu tworzona jest macierz śladu feromonowego (rys. 1 – krok 5), której analiza pozwala na zasugerowanie użytkownikowi utworzenie nowych folderów (rys. 1 – krok 6).



Rys. 1. Schemat algorytmu z mechanizmem predykcji folderów

Źródło: Opracowanie własne.

W wyniku pracy tak zaprojektowanego algorytmu, a także pewnej formy pamięci (decyzji innych użytkowników) poprzez ślad feromonowy, możliwe jest określenie wagi nowej sugestii. W tym celu zaproponowana została macierz śladu feromonowego będąca analogią do klasycznej macierzy błędu, która w tym przypadku ma na celu zobrazowanie powstałego rozwiązania, a nie określenie błędów klasyfikacji.

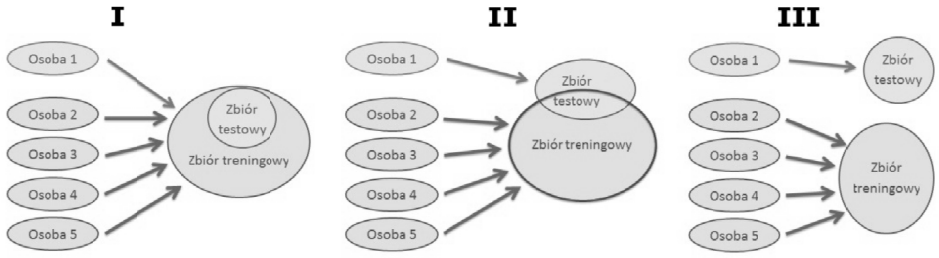
Macierz błędów (ang. *confusion matrix*) to narzędzie stosowane do oceny jakości modeli klasyfikacyjnych, które przedstawia zależność dokładności klasyfikacji każdej z klas oraz błędów wskazujących obiekty zaklasyfikowane do innej klasy. Wiersze w takiej macierzy odpowiadają poprawnym klasom decyzyjnym, a kolumny decyzjom przewidywanym przez klasyfikator. Dokładność klasyfikacji poszczególnych klas odczytujemy na podstawie przecięcia wierszy z kolumnami. Definicja macierzy błędów przedstawiona jest w tabeli 3.

Tabela 3. Definicja macierzy błędów

		Klasa przewidywana	
		Pozytywna	Negatywna
Klasa rzeczywista	Pozytywna	Prawdziwie pozytywna	Fałszywie negatywna
	Negatywna	Fałszywie pozytywna	Prawdziwie negatywna

Źródło: Opracowanie własne.

Przeprowadzone badania zostały podzielone na trzy oddzielne etapy ze względu na podział danych na zbiór treningowy i testowy. W każdym z trzech etapów zbiór treningowy stanowią skrzynki pocztowe osób sąsiadujących z użytkownikiem centralnym (kluczowym), natomiast zbiór testowy to skrzynka pocztowa użytkownika centralnego. W zależności od etapu badań zbiór testowy może w całości zawierać się w zbiorze treningowym (etap I), jest w 50% zawarty w zbiorze treningowym (etap II) lub stanowi zupełnie nowe dane dla klasyfikatora, gdyż nie ma wspólnych elementów w stosunku do zbioru treningowego (etap III). Charakterystyka podziału zbiorów została także przedstawiona na rysunku 2.



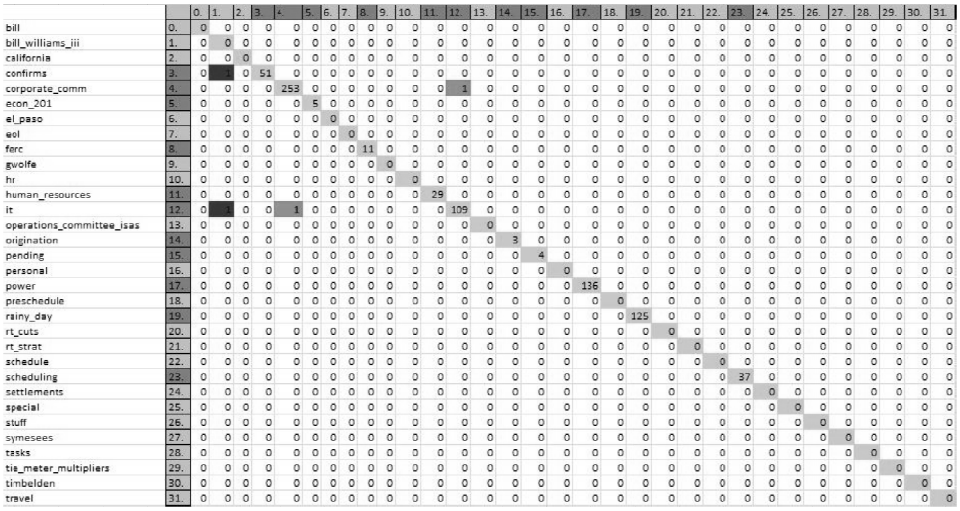
Rys. 2. Podział danych na zbiory treningowe i testowe

Źródło: Opracowanie własne.

Dla użytkowników kluczowych grup z tabeli 2, utworzonych na podstawie sieci społecznej opartej na kontaktach pracowników z firmy Enron Corporation, stworzono i przeanalizowano macierze śladu feromonowego odkładanego podczas klasyfikowania wiadomości do folderów. Dla każdego użytkownika stworzono trzy macierze śladu feromonowego, zgodnie z trzema etapami badań. Sugerowane foldery, które powinny zostać utworzone dla danego użytkownika, to te, do których w II lub III etapie zostało sklasyfikowanych wiele wiadomości, natomiast nie były one przypisane do tych folderów podczas etapu I.

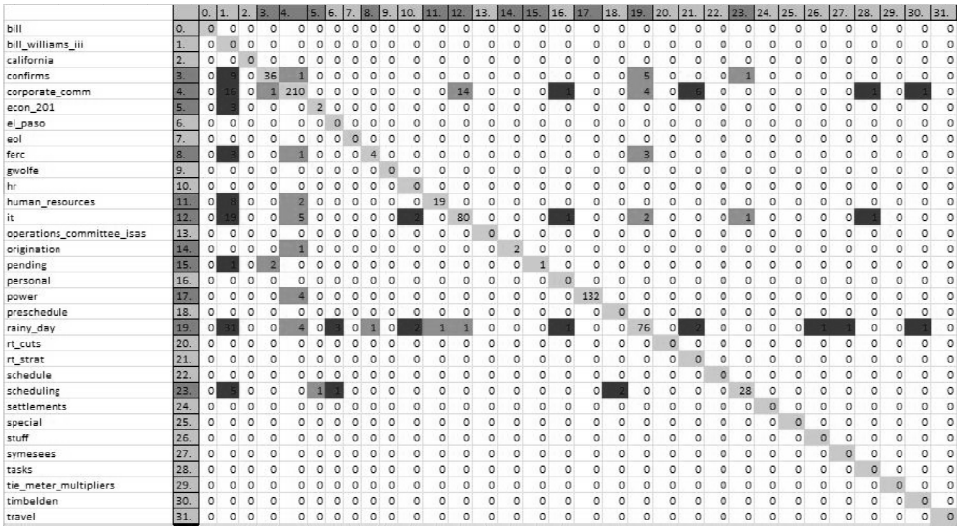
Na podstawie przeprowadzonej analizy macierzy śladu feromonowego odkładanego podczas pracy klasyfikatora otrzymano wyniki przypisania wiadomości do folderów dla wszystkich użytkowników kluczowych z tabeli 2, wybranych zgodnie z utworzoną siecią społeczną. Foldery, do których zostały przypisane wiadomości ze skrzynki użytkownika kluczowego, to wszystkie foldery, które zawierały się w skrzynkach pocztowych wszystkich użytkowników w grupie zgodnie ze wzorem (11), bez względu na występowanie tych folderów w skrzynce pocztowej użytkownika kluczowego.

Na rysunku 3 przedstawiono macierz dla I etapu dla przykładowego użytkownika *symes-k*. Natomiast rysunki 4 i 5 przedstawiają macierze śladu feromonowego odpowiednio dla II i III etapu dla tego samego użytkownika.



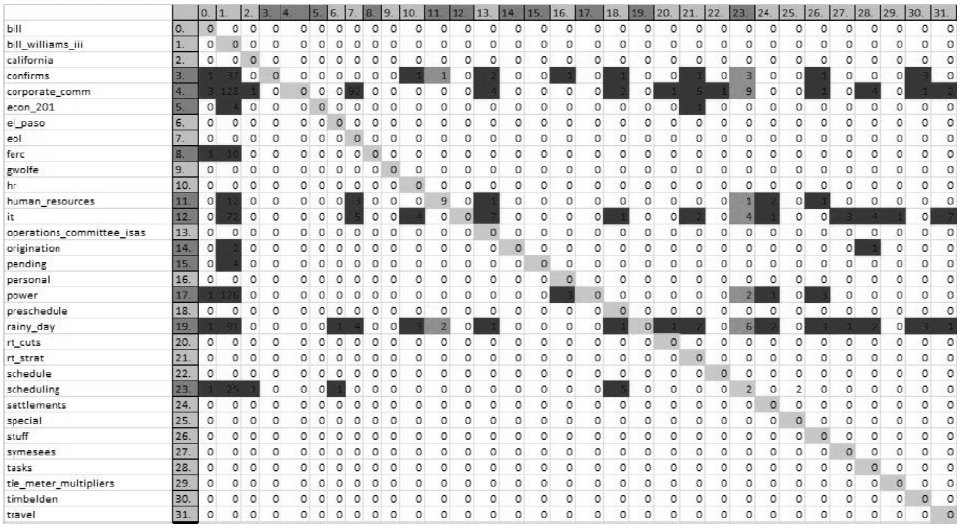
Rys. 3. Macierz śladu feromonowego dla symes-k – etap I

Źródło: Opracowanie własne.



Rys. 4. Macierz śladu feromonowego dla symes-k – etap II

Źródło: Opracowanie własne.



Rys. 5. Macierz śladu feromonowego dla symes-k – etap III

Źródło: Opracowanie własne.

W przedstawionych macierzach na przekątnej wykazana jest liczba wiadomości prawidłowo przypisanych do folderów, których wskazanie przez klasyfikator jest identyczne z folderami określonymi przez użytkownika w skrzynce pocztowej. Natomiast liczby występujące poza przekątną to wiadomości, których przypisanie do folderów oparte jest na strukturach folderów innych użytkowników będących najbliższymi sąsiadami kluczowego użytkownika, którzy razem tworzą grupy użytkowników. Po odrzuceniu folderów, do których nie została przypisana żadna wiadomość oraz tych, które występowały w skrzynce pocztowej użytkownika kluczowego, otrzymywana jest lista folderów, które zostają zasugerowane do utworzenia dla tego użytkownika.

Dla użytkowników kluczowych z tabeli 2, po przeanalizowaniu otrzymanych macierzy śladu feromonowego, w tabeli 4 przedstawiono wyniki mechanizmu predykcji folderów, określając, że minimalna liczba przypisanych wiadomości do folderu wynosi 10. Odpowiednio dla etapu II oraz etapu III przedstawiono liczbę folderów zasugerowanych do utworzenia w skrzynce pocztowej użytkownika kluczowego ze względu na przypisanie wiadomości do takich folderów przez innych użytkowników z grup określonych przez utworzoną sieć społeczną. Dodatkowo wykazano liczbę wiadomości, które zostały przypisane do nowo utworzonych folderów.

Tabela 4. Liczba sugerowanych folderów dla wybranych skrzynek pocztowych

Kluczowy użytkownik	Etap II		Etap III	
	liczba folderów	liczba wiadomości	liczba folderów	liczba wiadomości
lokay-m	7	124	3	40
sanders-r	4	849	14	894
shackleton-s	7	262	12	526
steffes-j	5	141	9	286
symes-k	3	550	8	750
williams-w3	2	1151	11	2250
farmer-d	27	564	59	2632
beck-s	12	268	25	727

Źródło: Opracowanie własne.

Podsumowanie

Otrzymane wyniki zależą nie tylko od dużej częstotliwości kontaktów pomiędzy poszczególnymi osobami, ale przede wszystkim od subiektywnie utworzonych struktur folderów innych osób. Jednocześnie stworzone macierze pozwalają na zaobserwowanie rzeczywistego rozwiązania – często bardzo duża liczba wiadomości przypisanych do folderów utworzonych przez innych użytkowników (dla przypadków zaznaczonych pogrubieniem w macierzy) w stosunku do liczby pozostałych wiadomości oznacza, że proponowana sugestia utworzenia folderu posiada duże wsparcie w przypadku grupy użytkowników. Natomiast im mniejsza wartość, tym słabsze wsparcie sugestii.

Literatura

Beckers R., Goss S., Deneubourg J.L., Pasteels J.M. (1989), *Colony Size, Communication and Ant Foraging Strategy*, "Psyche", Vol. 96, s. 239-256.

Boryczka U., Kozak J. (2010), *Ant Colony Decision Trees – A New Method for Constructing Decision Trees Based on Ant Colony Optimization*, "Computational Collective Intelligence. Technologies and Applications", LNCS, Springer, s. 373-382.

Boryczka U., Kozak J. (2011), *An Adaptive Discretization in the ACDT Algorithm for Continuous Attributes*, "Computational Collective Intelligence. Technologies and Applications", LNCS, Springer, s. 475-484.

Boryczka U., Probierz B., Kozak J. (2016), *Automatic Categorization of Email into Folders by Ant Colony Decision Tree and Social Networks* [w:] I. Czarnowski, A.M. Caballero, R.J. Howlett, L.C. Jain (eds.), *Intelligent Decision Technologies*

2016: Proceedings of the 8th KES International Conference on Intelligent Decision Technologies (KES-IDT 2016), Part II, Springer, Tenerife, Spain, s. 71-81.

Doerner K.F., Merkle D., Stutzle T. (2009), *Special Issue on Ant Colony Optimization*, "Swarm Intelligence", Vol. 3(1), s. 1-2.

Dorigo M., Birattari M., Blum Ch., Clerc M., Stutzle T., Winfield A., eds. (2008), *Ant Colony Optimization and Swarm Intelligence*, 6th International Conference, ANTS 2008, Vol. 5217, LNCS, Springer, Berlin, Heidelberg.

Dorigo M., Birattari M., Stutzle T. (2006), *Ant Colony Optimization – Artificial Ants as a Computational Intelligence Technique*, "IEEE Computational Intelligence Magazine", Vol. 1, s. 28-39.

Dorigo M., Di Caro G. (1999), *New Ideas in Optimization*, McGraw–Hill, London, UK.

Dorigo M., Di Caro G., Gambardella L. (1999), *Ant Algorithms for Distributed Discrete Optimization*, "Artifical Life", Vol. 5(2), s. 137-172.

Dorigo M., Maniezzo V., Colorni A. (1996), *The Ant System: An Autocatalytic Optimization Process*, Raport instytutowy 91-016, Department of Electronics, Politecnico di Milano.

Enron (b.r.), *E-mail Dataset*, https://www.cs.cmu.edu/~./enron/ (dostęp: 25.02.2017).

Goss S., Gervet J., Deneubourg J.L., Theraulaz G. (1990), *Swarm Intelligence in Wasps Colonies: An Example of Task Assignment in Multiagents Systems* [w:] A. Meystel, J. Herarth, S. Gray (eds.), *Proceedings of the 1990 IEEE International Symposium on Intelligent Control*, IEEE Computer Society Press, Los Alamitos, CA, s. 135-143.

Moreno J.L. (1934), *Who Shall Survive? A New Approach to the Problem of Human Interrelations*, Vol. 58, Nervous and Mental Disease Publishing Co., Washington, DC.

Moreno J.L. (1978), *Who Shall Survive? Foundations of Sociometry*, Group Psychotherapy and Sociodrama, Beacon House, Beacon, NY.

Verhaeghe J.C., Deneubourg J.L (1983), *Experimental Study and Modelling of Food Recruitment in the Ant Tetramorium Impurum*, "Insectes Sociaux", Vol. 30, s. 347-360.

Zhang P., Wang J., Xiaojia L., Menghui L., Zengru D., Ying F. (2008), *Clustering Coefficient and Community Structure of Bipartite Networks*, "Physica A: Statistical Mechanics and its Applications", Vol. 387(27), s. 6869-6875.

MECHANISM TO SUGGEST THE CREATION OF NEW FOLDERS IN AUTOMATIC E-MAIL MESSAGE CLASSIFICATION IN E-MAIL BOXES

Summary: This article has been proposed a new method of suggesting the creation of folders in users' mailboxes through the use of social networks and Ant Colony Optimization. The aim of the work is to create new folders in the prediction of automatic classification of e-mail folders in mailboxes. In the proposed algorithm, social network analy-

sis elements were used to designate groups of users having a similar folder structure in mailboxes, based on which the mechanism of suggesting the creation of new folders for users was based. The proposed method has been tested on the public Enron E-mail data set.

Keywords: Enron E-mail, Social Networks, folder prediction, Ant Colony Optimization.