Ewa Konopka, Adam Pelikant
Wydział Informatyki i Zarządzania
Wyższa Szkoła Informatyki i Umiejętności
ul. Rzgowska 17a 93-008 Łódź
email: adam.pelikant@p.lodz.pl

ZASTOSOWANIE METOD GRUPOWANIA W ANALIZIE SIECI SPOŁECZNOŚCIOWYCH

Streszczenie – Celem pracy jest omówienie różnych metod klasteryzacji (grupowania) w sieciach społecznych. Analizowane dane są wstępnie podzielone na klastry według miejsca zamieszkania członków sieci. Opracowany algorytm i bazująca na nim aplikacja dokonuje oceny jakości grupowania oraz umożliwia ponowny podział według różnych metod, a następnie porównanie wyników ich działania. Zaimplementowanych zostało wiele algorytmów, których działanie daje odmienne rezultaty. Aplikacja współpracuje z serwerem baz danych Microsoft SQL Server. Zastosowane zostały dwa typy użytkownika (UDT) w technologii CLR, które implementują obiekty odpowiadające składowym sieci-grafu [1]: osobę (SocNetPerson) i klaster (SocNetCluster).

Słowa kluczowe: Teoria grafów, analiza grafów, sieci społecznościowe, typy użytkownika CLR, grupowanie, ocena jakości grupowania, eksploracja danych, programowanie obiektowe w bazach danych

1 Wstęp

Sieci społeczne stanowią odzwierciedlenie relacji między ludźmi na różnych płaszczyznach: zawodowej, edukacyjnej, towarzyskiej, hobbystycznej. Istnieją różne rodzaje sieci społecznych: sieć kontaktów zawodowych, połączeń telefonicznych, osób o podobnych zainteresowaniach, sieć wysłanych emaili czy po prostu sieć znajomych, przyjaciół. Kolejnym przykładem takiej struktury są serwisy społecznościowe, grupujące ludzi np. o podobnych zainteresowaniach czy znających się w świecie realnym. Pozwalają one na prezentację własnej osoby, przesyłanie wiadomości, publikowanie informacji, także multimedialnej, śledzenie innych użytkowników, zawieranie znajomości i wiele innych działań.

Przeniesienie struktur sieci społecznych w sferę Internetu ułatwia ich analizę. W bazach danych zgromadzono do tej pory olbrzymie ilości informacji, co stanowi znakomity rezerwuar danych dla badaczy. Analizie poddaje się wiele różnych cech sieci społecznych. Jedną z właściwości, pozwalającą na wyodrębnienie względnie jednorodnych,

różniących się od siebie grup w ramach sieci, jest klasteryzacja. Umożliwia ona podział struktury na podsieci, charakteryzujące się podobieństwem wewnętrznym i zróżnicowaniem wobec siebie nawzajem. Identyfikacja grup jest istotnym elementem analizy sieci społecznych. Praca jest w całości poświęcona zagadnieniom klasteryzacji (inaczej grupowania, gronowania), metod jej oceny oraz sposobom implementacji w programie komputerowym.

2 Elementy teorii grafów

Sieć to abstrakcyjna struktura, reprezentująca system powiązań pomiędzy różnymi obiektami (elementami składowymi). Grafem G nazywamy uporządkowaną parę G = (V, E), gdzie V jest niepustym zbiorem wierzchołków, a E jest zbiorem jego krawędzi.

$$E \subseteq \{\{u,w\}: u,v \in V, u \neq v\ \} \tag{1}$$

Wierzchołki w grafie mogą być połączone ze sobą krawędziami, ale również możliwe jest połączenie wierzchołka z samym sobą. Do krawędzi grafów można przypisać wartości, które nazywamy wagami. Graf, który ma taką własność będziemy nazywać grafem ważonym lub grafem z wagami. Wagi mogą być zarówno dodatnie jak i ujemne. Grafem pustym nazywamy graf składający się tylko z wierzchołków, nie zawierający krawędzi.

Trasą w grafie G nazywamy skończony ciąg krawędzi postaci $v_0v_1, v_1v_2, \ldots, v_{m-1}v_m$, w którym każde dwie kolejne krawędzie są albo sąsiednie, albo identyczne. Trasę, w której wszystkie krawędzie są różne nazywamy ścieżką, a jeśli ponadto wierzchołki są różne, to ścieżkę nazywamy drogą [1]. Trasę, która ma początek i koniec w tym samym wierzchołku nazywany cyklem.

Grafem spójnym nazywamy graf, w którym dla każdej pary wierzchołków istnieje ścieżka, która je łączy, a grafem pełnym taki, w którym każdą parę wierzchołków łączy krawędź. Jest on także nazywany kliką. Gęstość grafu to stosunek liczby krawędzi do największej możliwej liczby krawędzi:

$$\frac{2|E|}{|V|(V|-1)} \tag{2}$$

Stopniem wierzchołka v w grafie G nazywamy liczbę krawędzi incydentnych (wychodzących lub wchodzących do wierzchołka) i oznaczamy deg(v).

Stopniem grafu G nazywamy liczbę:

$$\Delta(G) = \max_{v \in V} \deg(v) \tag{3}$$

Stopień grafu jest równy najwyższemu ze stopni jego wierzchołków.

Grafem regularnym nazywamy graf, w którym stopień każdego wierzchołka ma tę samą wartość. Podgrafem grafu G nazywamy graf powstały poprzez usunięcie części wierzchołków z H razem ze wszystkimi kończącymi się w nich krawędziami.

Grafem skierowanym lub inaczej digrafem nazywamy uporządkowaną parę (V, E), gdzie V jest skończonym niepustym zbiorem wierzchołków, a $E \subseteq VxV$ to zbiór jego łuków (krawędzi skierowanych posiadających zwrot). Multigrafem nazywamy graf, w którym możliwe są wielokrotne krawędzie łączące te same dwa wierzchołki, jak również wierzchołek z samym sobą (taką krawędź nazywamy pętlą). Grafem planarnym nazywamy taki, który można przedstawić na płaszczyźnie w taki sposób, aby jego krawędzie nie przecinały się. Drzewem nazywamy graf spójny bez cykli.

W pamięci komputera graf można reprezentować na kilka sposobów. Najpopularniejszymi są macierz i lista sąsiedztwa. Macierz sąsiedztwa – zbiór krawędzi reprezentowany jest jako tablica kwadratowa A[ ], o wymiarach n x n, gdzie n oznacza liczbę wszystkich wierzchołków w grafie. Elementy macierzy zdefiniowane są następująco:

$$A[x, y] = \begin{cases} 1, (x, y) \in E \\ 0, (x, y) \notin E \end{cases} \tag{4}$$

gdzie:

x, y - pary wierzchołków

E - zbiór krawędzi

W komórce na przecięciu wiersza w i kolumny k zapisana jest 1, jeśli istnieje krawędź między wierzchołkami w i k. W przeciwnym wypadku w komórce tablicy zapisywane jest 0. Jeśli graf jest ważony, można bezpośrednio w macierzy zapisać wagę. Należy jednak pamiętać o umownym sposobie reprezentacji braku krawędzi (0 może mieć przypisane takie znaczenie, jednakże w tym przypadku nie ma już możliwości zapisania krawędzi o koszcie 0).

Lista sąsiedztwa – zbiór krawędzi reprezentowany jest jako tablica jednowymiarowa L[n], gdzie n oznacza liczbę wierzchołków w grafie. Każdy element tablicy przedstawiony jest jako lista wierzchołków, z którymi dany wierzchołek jest połączony krawędzią.

3 Właściwości sieci

W [3], [4] zaproponowano następujące miary centralności wierzchołków.

Znormalizowany stopień dc_i (ang. degree) wierzchołka i:

$$dc_i = \frac{k_i}{N-1} \tag{5}$$

gdzie:

k_i - stopień wierzchołka i w grafie,

N - liczba wierzchołków w grafie.

Wierzchołek uzyskuje najwyższą wartość tego współczynnika, gdy ma najwyższy stopień, czyli ma najwięcej sąsiadów.

Promień rc_i (ang. radius) wierzchołka i:

$$rc_i = \frac{1}{\max\limits_{j \in V} d_{ij}} \tag{6}$$

gdzie:

d_{ij} - długość najkrótszej drogi miedzy wierzchołkami i oraz j. Wierzchołek uzyskuje najwyższą ocenę kiedy odległość, która dzieli go od najdalszego wierzchołka jest najmniejsza.

Bliskość cc_i (ang. closeness) wierzchołka i:

$$cc_i = \frac{N-1}{\sum\limits_{j \in V} d_{ij}} \tag{7}$$

gdzie:

d_{ij} - długość najkrótszej drogi miedzy wierzchołkami i oraz j,

N - liczba wierzchołków w grafie.

Miara ta opiera się na założeniu, że wierzchołek jest bardziej centralny, im jest bliżej innych wierzchołków.

Pośrednictwo bc_i (ang. betweenness) wierzchołka i:

$$bc_i = \frac{\sum\limits_{l \in V} \sum\limits_{k \neq l \in V} \frac{p_{l,i,k}}{p_{l,k}}}{(N-2)(N-1)} \tag{7}$$

gdzie:

$p_{l,i,k}$ - liczba dróg w grafie między wierzchołkami l oraz k, które przechodzą przez i, $p_{l,k}$ - liczba wszystkich dróg w grafie między wierzchołkami l oraz k.

Wartość miary normalizuje się biorąc pod uwagę maksymalną możliwą liczbę najkrótszych dróg w grafie pełnym.

Współczynnik gronowania (klasteryzacji) gc_i (ang. clusterization) wierzchołka i:

$$gc_i = \frac{2E_i}{k_i(k_i - 1)} \tag{8}$$

gdzie:

k_i - stopień wierzchołka i w grafie (wartość ki jest większa od 1),

E_i - liczba krawędzi między sąsiadami wierzchołka i.

Współczynnik klasteryzacji wierzchołka określa stosunek liczby krawędzi między sąsiadami tego wierzchołka do liczby wszystkich możliwych połączeń między nimi.

Współczynnik gronowania (średni) C (ang. clustering coefficient) dla całej sieci:

$$C = \frac{1}{N}\sum_{i \in V} gc_i \tag{9}$$

gdzie:

gc_i - współczynnik gronowania wierzchołka,

N - liczba wierzchołków w grafie.

Średnia odległość L (średnia długości najkrótszych dróg) w całej sieci:

$$L = \frac{\sum_{i \neq j \in V} d_{ij}}{N(N-1)} \tag{10}$$

gdzie:

d_{ij} - długość najkrótszej drogi miedzy wierzchołkami i oraz j,

N - liczba wierzchołków w grafie.

Potęgowy rozkład stopni wierzchołków P(k):

$$P(k) \sim k^{-\gamma} \tag{11}$$

gdzie:

k - stopień wierzchołka

γ - wykładnik stopnia wierzchołka

Sieci o takim rozkładzie stopni wierzchołków nazywamy sieciami bezskalowymi. Dla większości tych sieci parametr γ przyjmuje wartości z przedziału <2, 3>.

4 Modele sieci losowych

Najlepsze wyniki można uzyskać analizując duże sieci opisujące rzeczywiste powiązania. Jednak dostęp do nich jest w większości przypadków ograniczony. Nie dysponując tego rodzaju danymi źródłowymi, zmuszeni jesteśmy generować dane. Istnieje wiele różnych algorytmów tworzenia sieci losowych, o odmiennych założeniach, których rezultaty działania dają wyniki znacznie różniące się właściwościami powstałych sieci.

5 Model Erdős–Rényi

Jednym z podstawowych algorytmów tworzenia grafu losowego jest model Erdős– Rényi. Istnieją dwa, ściśle ze sobą powiązane, warianty tego modelu:

- Spośród wszystkich możliwych *k* grafów G o *n* wierzchołkach i *m* krawędziach wybierany jest jeden, z prawdopodobieństwem 1/*k*.
- Każda para wierzchołków jest łączona krawędzią z prawdopodobieństwem *p*. Rozkład P(k) stopni wierzchołków w tym modelu jest rozkładem dwumianowym, a przy małej liczbie krawędzi można przybliżyć go rozkładem Poissona. Większość wierzchołków ma w przybliżeniu taką samą liczbę krawędzi incydentnych. Współczynnik gronowania danego wierzchołka jest w praktyce niezależny od jego stopnia.

Prawdopodobieństwo przyłączenia kolejnego wierzchołka w sieci losowej maleje wykładniczo dla wierzchołków o coraz wyższym stopniu. Sieci losowe charakteryzują się więc niskim współczynnikiem gronowania i małą średnią wartością długości dróg między wierzchołkami.

Algorytm tworzenia sieci w oparciu o model Erdős–Rényi w wariancie b przedstawia metakod:

1. Ustal parametr p.
2. Stwórz N wierzchołków grafu.
3. Dla każdego wierzchołka $v_i \in N, i \leq N$ wykonaj:
4. Dla każdego wierzchołka $v_j \in N, i < j \leq N$ wykonaj:
5. Wylosuj liczbę x z przedziału (0;1).
6. Jeśli x < p, to połącz wierzchołki vi i vj krawędzią.

6 Sieci Małego Świata

W 1998 roku dwaj matematycy Duncan Watts i Steven Strogatz w opublikowanej pracy [5] zaproponowali model sieci „małego świata”

(ang. Small World). Według nich model powstaje z sieci regularnej, w której losowo wybranym krawędziom zamieniamy wierzchołki startowe. W związku z tym drogi między odległymi parami wierzchołków są krótsze, a więc średnia długość dróg jest mała. Według spostrzeżeń autorów, sieć małego świata powinna charakteryzować się małą średnią odległością między wierzchołkami i wysokim współczynnikiem klasteryzacji.

7 Model Barabasi-Albert

W 1999 roku dwaj fizycy, Albert-László Barabási i Réka Albert, w swojej pracy [6] zaproponowali model sieci „bezskalowej" (ang. Scale Free). Służy on do generowania losowych grafów. Model B-A charakteryzuje się stałym wzrostem rozmiaru sieci, preferencyjnym dołączaniem nowych wierzchołków do sieci i potęgowym rozkładem stopni wierzchołka. Rozkład ten w sieci bezskalowej spełnia prawo siły, która jest matematyczną zależnością między dwiema wielkościami, mówiącą o tym, że częstotliwość występowania pewnego zdarzenia zmienia się tak, jak siła jakiejś cechy tego zdarzenia. W wypadku grafów bezskalowych przekłada się to na niewielką liczbę wierzchołków o wysokim stopniu i bardzo dużą liczbę wierzchołków o stopniu niskim. Zasada wzrostu rozmiaru sieci oznacza, że sieć rośnie z upływem czasu. Zasada preferencyjnego przyłączania mówi, że nowe wierzchołki przyłączają się z większym prawdopodobieństwem do wierzchołków o wysokim stopniu niż do wierzchołków ze stopniem niskim. Wierzchołki o wyższym stopniu mają większą zdolność przyciągania nowych wierzchołków, a więc im wyższy stopień ma wierzchołek w sieci (więcej znajomych), tym większe prawdopodobieństwo, że nowy wierzchołek zostanie z nim połączony. W kontekście sieci społecznych zasada ta jest opisywana określeniem „bogaci stają się coraz bogatsi" [6], [7]. Opisany model Barabási-Albert został wykorzystany do wygenerowania losowego grafu reprezentującego sieć społeczną potrzebną do przeprowadzenia dalszych analiz, które są przedmiotem tej pracy. Algorytm generowania grafu reprezentującego sieć społeczną w oparciu o ten model przebiega następująco:

1. Generujemy n wierzchołków grafu.
2. Wybieramy losowo 2 wierzchołki: j i k, a następnie łączymy je krawędzią i przyłączamy do sieci S.
3. Dla każdego nowo utworzonego wierzchołka n wykonaj:
4. Dla każdego wierzchołka s w sieci S wykonaj:
5. Wylosuj liczbę rzeczywistą R z zakresu (0;1)
6. Jeśli liczba R jest mniejsza niż stosunek stopnia wierzchołka s do całkowitej sumy stopni w grafie

$$\frac{\deg(s)}{\sum_{r \in S} \deg(r)}$$

```
gdzie:
  deg(s) – stopień wierzchołka rozpatrywanego (s),
  ∑r∈S deg(r) – suma stopni wszystkich wierzchołków
sieci S
to połącz krawędzią wierzchołki n i s.
```

Ze wzoru zamieszczonego w metakodzie wynika, że im wyższy stopień ma wierzchołek w sieci (więcej znajomych), tym większe prawdopodobieństwo, że nowy wierzchołek zostanie z nim połączony.

8 Klasteryzacja

Klasteryzacja (grupowanie, ang. clustering), jest metodą nienadzorowanej analizy danych. Jej celem jest podział danych na klastry (grupy) w taki sposób, aby każdy z nich zawierał obiekty najbardziej ze sobą powiązane (podobne do siebie), a równocześnie elementy należące do różnych klastrów powinny jak najbardziej różnić się między sobą. Do oceny podobieństwa obiektów mogą być stosowane różne kryteria. Najczęściej wykorzystuje się do tego celu różne miary odległości. W procesie klasteryzacji nie dysponujemy danymi wzorcowymi, a co za tym idzie proces ten jest uczeniem „bez nauczyciela".

Wierzchołek grafu zostanie przypisany do klastra, w którego skład wchodzą wierzchołki najbardziej do niego podobne. Przyporządkowanie wierzchołka do klastra może być twarde (ang. hard) lub rozmyte (ang. fuzzy). Przyporządkowanie twarde przypisuje dany wierzchołek dokładnie do jednego klastra, podczas gdy przynależność rozmyta umożliwia przypisanie wierzchołka do więcej niż jednego klastra, z uwzględnieniem współczynnika przynależności z przedziału <0,1>. Należy zaznaczyć, że suma współczynników dla każdego wierzchołka musi wynosić 1.

Klasteryzację stosuje się w procesie rozpoznawania obrazów, wzorców, ale także w marketingu czy badaniach rynku. Dobre metody klasteryzacji cechują:

- umiejętność tworzenia klastrów o różnych kształtach,
- mała wrażliwością na zaburzenia danych (pomijanie szumu),
- duża skalowalność,
- niska złożoność obliczeniowa.

Istnieje bardzo wiele metod i algorytmów grupowania. Podziału możemy dokonać ze względu na typy danych (np. liczby, dane tekstowe, obrazy itp.), mechanizm generowania klastrów (algorytmy deterministyczne i probabilistyczne), czy sposób otrzymania klastrów (algorytmy hierarchiczne i płaskie). Wyróżniamy następujące grupy metod klasteryzacji:

1. Metody hierarchiczne budują klastry uporządkowane i umożliwiają obserwację działania na różnych poziomach szczegółowości, ponieważ przedstawiają strukturę klasteryzacji w postaci drzewa (dendrogramu). Dzielimy je na
   a) metody aglomeracyjne (ang. agglomerative),
   b) metody dzielące (ang. divisive).

W metodach aglomeracyjnych każdy obiekt stanowi osobny klaster (skupienie), następnie w kolejnych iteracjach klastry łączone są w większe skupienia. Algorytm w metodach aglomeracyjnych działa następująco:

1. Utwórz n klastrów zawierających pojedyncze obiekty.
2. Oblicz wartości miary podobieństwa (odległości) dla wszystkich par obiektów.
3. Połącz dwie grupy najbardziej podobne.
4. Jeżeli wszystkie obiekty należą do jednego klastra, to zakończ pracę.
5. W przeciwnym wypadku przejdź do punktu 2.

W metodach dzielących (deaglomeracyjnych) początkowo wszystkie obiekty przypisujemy do jednego klastra, a następnie w kolejnych iteracjach jest on dzielony na coraz mniejsze grupy tak długo, aż osiągniemy ich zadaną liczbę lub każdy obiekt będzie osobnym klastrem. Algorytm w metodach podziałowych działa następująco:

1. Włóż wszystkie obiekty do jednego klastra.
2. Oblicz wartości funkcji kryterium dla wszystkich możliwych podziałów klastra na dwie lub więcej podgrupy.
3. Wybierz najlepszy podział i podziel klaster.
4. Jeżeli każdy obiekt znajduje się w oddzielnym klastrze, zakończ pracę.
5. W przeciwnym razie idź do punktu 2.

Najczęściej stosowane miary odległości miedzy grupami (klastrami) w metodach hierarchicznych to:

– metoda pojedynczego połączenia (najbliższego sąsiada - ang. single link) – wyrażana przez najmniejszą odległość między dwoma najbliższymi punktami z różnych skupień

$$d_{\min}\left(C_i, C_j\right) = \min_{p \in C_i, p' \in C_j} \left\| p - p' \right\| \tag{12}$$

– metoda pełnego połączenia (najdalszego sąsiada, ang. complete link) - brana jest pod uwagę największa odległość między dwoma najbardziej oddalonymi punktami, po jednym z każdego klastra

$$d_{\max}\left(C_i, C_j\right) = \max_{p \in C_i, p' \in C_j} \left\| p - p' \right\| \tag{13}$$

– metoda odległości miedzy środkami (ang. mean distance) - odległość zdefiniowana jest między dwoma środkami ciężkości danych klastrów

$$d_{mean}\left(C_i, C_j\right) = \left\| m_i - m_j \right\| \tag{14}$$

– metoda średniej odległości (ang. average distance) - średnia odległość pomiędzy wszystkimi parami elementów należących do obu klastrów.

$$d_{ave}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} \left\| p - p' \right\| \tag{15}$$

gdzie dla wzorów (12) –(15):

p, p’ - dowolne obiekty (punkty);
C_i, C_j – klastry;
m_i, m_j - środki klastrów;
n_i, n_j - liczba obiektów (punktów);

2. Metody partycjonujące polegają na konstruowaniu zadanej liczby klastrów z n obiektów. W każdym kroku obiekt może być przydzielony do innego klastra, przez co uzyskuje się poprawę jakości grupowania. Wyróżniamy następujące metody:

– probabilistyczne, które bazują na określeniu prawdopodobieństwa na podstawie rozkładu Gaussa.
– metoda k-średnich (ang. k-means) jest jedną z najprostszych metod, która dokonuje klasteryzacji. Do obliczania odległości między dowolnym punktem, a odpowiadającym mu środkiem klastra najczęściej stosuje się miarę euklidesową (L2). Można ją opisać następującym metakodem:

```
1. Dokonaj losowego wyboru n obiektów jako początkowe
   środki (centroidy) k klastrów,
2. Przypisz elementy do najbliższych klastrów,
3. Wylicz środki nowych klastrów,
4. Powtarzaj kroki 2-3 tak długo, jak długo występują
   zmiany przydziału obiektów do klastrów, lub nie
   zostanie wykonana określona liczba iteracji.
```

Algorytm k-średnich jest prosty i szybki, ma niewielką złożoność obliczeniową, ale ma również pewne ograniczenia. Wynik działania algorytmu zależy od początkowego podziału obiektów na klastry. Aby zwiększyć szansę znalezienia optymalnego rozwiązania należy uruchomić algorytm kilkakrotnie dla różnych podziałów początkowych. Minusem tej metody jest negatywny wpływ odległych obiektów na położenie środka ciężkości i niezdolność wychwycenia szumu.

Zaproponowana w pracy metoda liderów przyjmuje, że klastrami staną się liderzy, tj. te osoby, których liczba znajomych jest większa lub równa podanemu parametrem progowi. Liczba klastrów nie jest zatem narzucana z góry, a zależna od liczby osób „popularnych" w sieci. Klastry tworzone są w miejscu zamieszkania liderów, a następnie poddawane są dokładnie takim samym działaniom jak w metodzie k-średnich. Przebieg algorytmu:

1. Utwórz k klastrów w miejscach, w których znajdują się osoby o liczbie znajomych >= z,
2. Przypisz elementy do najbliższych klastrów,
3. Wylicz środki nowych klastrów,
4. Powtarzaj kroki 2-3 tak długo, jak długo występują zmiany przydziału obiektów do klastrów, lub nie zostanie wykonana określona liczba iteracji.

Cechą charakterystyczną algorytmu jest fakt, że liderzy mogą w pewnym momencie zostać przesunięci do innego klastra niż ten, który określali na początku jego działania.

Metoda k-medoids (ang. k-medoid) polega na tworzeniu medoid, czyli obiektów z grupy, które w danym klastrze są najbardziej centralne, to znaczy że ich odległość od wszystkich pozostałych w skupieniu jest najmniejsza. Przykładowymi algorytmami są PAM, CLARA i CLARANS. Przebieg algorytmu PAM:

1. Dokonaj wyboru k obiektów jako początkowe medoidy,
2. Przypisz każdy z pozostałych obiektów do najbliższego medoidu,
3. Zamieniaj każdy z medoidów z nie-medoidem do chwili, gdy nie ma już zmian w zawartości grupowania wyliczając koszt zmiany,
4. Powtarzaj kroki 2-3 tak długo, jak długo występują zmiany przydziału w grupowaniu do medoidów.

Algorytm k-medoids dobrze radzi sobie z odległymi obiektami, a początkowy wybór medoidów nie ma wpływu na wyniki, jednak nie radzi sobie z dużymi zbiorami danych.

Metoda rozmytej analizy skupień (ang. fuzzy clustering) polega na przydzielaniu elementów do więcej niż jednego klastra z tak zwanym prawdopodobieństwem przynależności. Nie jest to klasyczna klasteryzacja, ponieważ grupowanie nie jest rozłączne. Grupowanie rozmyte jest pomocnew przypadku, gdy nie istnieje wyraźna granica rozdzielająca grupy obiektów. Jedną z metod jest rozmyta metoda k-średnich (fuzzy c-means). Podstawowym celem metody jest znalezienie takich środków ciężkości klastrów, które minimalizują funkcję określoną wzorem:

$$J_m = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m \left\| d_{ij} \right\|^2 \tag{16}$$

gdzie:

m - parametr rozmycia większy od 1

u_{ij} - stopień przynależności j- tego obiektu do i-tego klastra

d_{ij} - odległość euklidesowa między środkiem ciężkości i-tego klastra a j-tym obiektem.

Algorytm rozmytej metody k-średnich zbudowany jest z następujących kroków:

1. Dokonaj losowego wyboru n obiektów jako początkowe środki (centroidy) k klastrów,
2. Wybierz losowo stopnie przynależności do klastrów dla wszystkich obiektów,
3. Dopóki zmiana stopni przynależności przekracza zadaną wartość ε, wykonaj:
   a) Oblicz centroidy klastrów ze wzoru:

$$c_k = \frac{\sum_x w_k(x)^m x}{\sum_x w_k(x)^m}$$

gdzie:
wk(x) – stopnień przynależności obiektu x do klastra k,
m – współczynnik > 1

b) Dla każdego obiektu oblicz stopnie przynależności do klastrów ze wzoru:

$$w_k(x) = \frac{1}{\sum_j \left(\frac{d(center_k, x)}{d(center_j, x)} \right)^{\frac{2}{m-1}}}$$

Wadą tego algorytmu jest kulisty kształt klastrów.

Metody grupowania gęstościowego polegają na wyszukaniu punktów gęsto ułożonych. Tworzą klastry o różnych kształtach. Dobrze radzą sobie z oddalonymi punktami. Podstawowymi algorytmami są DBSCAN i OPTICS.

DBSCAN (ang. Density Based Spatial Clustering of Applications with Noise). Klaster zdefiniowany jest jako maksymalny zbiór gęstościowo połączonych punktów, natomiast punkty niepołączone z nim są punktami oddalonymi. Przed uruchomieniem algorytmu potrzebne jest zdefiniowanie dwóch parametrów: E - maksymalny promień sąsiedztwa i minPts - minimalna liczba punktów w sąsiedztwie. Sąsiedztwo określone jest jako:

$$N_{\varepsilon} = \left| y \in X; d(x, y) \leq \varepsilon \right| \tag{17}$$

Algorytm przedstawia się następująco:

```
Dla każdego nieodwiedzonego punktu P wykonaj:
    1. Zaznacz P jako odwiedzony
    2. SasiedziPTS = WybierzSasiadow(P, E)
    3. Jeśli liczność(SasiedziPTS) < minPts, to oznacz P
       jako szum
    4. W przeciwnym wypadku:
    5. K = nowy klaster
    6. RozszerzKlaster(P, SasiedziPTS, K, E, minPts)

Funkcja RozszerzKlaster(P, SasiedziPTS, K, E, minPts)
    1. Dodaj P do klastra K
    2. Dla każdego punktu P' z SasiedziPTS wykonaj:
    3. Jeśli P' jest nieodwiedzony
    4. Zaznacz P' jako odwiedzony
    5. SasiedziPTS' = WybierzSasiadow(P', E)
    6. Jeśli liczność(SasiedziPTS') >= minPts, to dołącz
       SasiedziPTS' do SasiedziPTS
    7. Jeśli P' nie jest przyłączony do żadnego klastra,
       to przypisz go do K

 Funkcja WybierzSasiadow(P, E)
    1. zwróć listę punktów oddalonych od P nie więcej niż
       o E
```

Metody gridowe używają siatkowych struktur danych o wielu poziomach dokładności. Przykładowymi algorytmami są STING i WaveCluster, metoda górska [8].

9 Miary odległości w metodach klasteryzacji

Bardzo istotnym czynnikiem, który ma wpływ na podział danych, jest sposób obliczania odległości pomiędzy nimi. Najczęściej stosowaną miarą odległości jest odległość potęgowa - inaczej nazywana odległością Minkowskiego

$$d(x,y) = \left(\sum_i |x_i - y_i|^q\right)^{1/q} \tag{18}$$

gdzie q to parametr zdefiniowany przez użytkownika.

Dla q = 1 otrzymujemy odległość Manhattan (city-block), która jest wartością bezwzględną z różnicy miedzy wartościami i-tych cech badanych obiektów.

$$d(x,y) = \sum_i \left(|x_i - y_i|\right) \tag{19}$$

Dla q = 2 uzyskujemy odległość Euklidesową, geometrycznie określaną jako odległość między dwoma punktami w przestrzeni n-wymiarowej i wyrażonej wzorem:

$$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2} \tag{20}$$

Odległość Czebyszewa - jest to największa różnica między dwoma badanymi punktami x i y.

$$d(x,y) = \max\left(|x_i - y_i|\right) \tag{21}$$

gdzie dla powyższych wzorów:

x_i, y_i - współrzędne badanych obiektów

Odległość Mahalanobisa - jest to odległość między dwoma punktami w n-wymiarowej przestrzeni, która różnicuje udział poszczególnych składowych oraz wykorzystuje korelacje między nimi. Definiujemy ją jako:

$$d(x,y) = \sqrt{(x-y)C^{-1}(x-y)^T} \tag{22}$$

gdzie:

x, y - wektory losowe

C - symetryczna, dodatnio określona macierz

Współczynnik korelacji liniowej Pearsona jest to bezwymiarowy współczynnik podobieństwa, którego wartość leży w zakresie <-1,1> i odzwierciedla stopień zależności między badanymi punktami (obiektami, cechami). Wyrażamy go wzorem:

$$r_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} \tag{23}$$

gdzie:

x_i, y_i - kolejne pary współrzędnych;

$\bar{x}$, $\bar{y}$ - średnie arytmetyczne

Odległość kątowa jest również współczynnikiem podobieństwa wyrażonym wzorem:

$$p_{xy} = \frac{\sum_i x_i y_i}{\sum_i x_i^2 \sum_i y_i^2} \tag{24}$$

10 Funkcje oceny klasteryzacji

Po przeprowadzeniu klasteryzacji należy dokonać jej oceny. Najczęściej używane funkcje uwzględniają odległości obiektów od środków (lub centroidów) klastrów w stosunku do odległości pomiędzy klastrami. W procesie klasteryzacji oczekujemy, że powstaną klastry maksymalnie zwarte i maksymalnie rozłączne. Do oceny wyników grupowania mogą posłużyć poniższe miary.

Odchylenie wewnątrzklastrowe wc(C), które jest sumą odległości obiektów od środków klastrów, do których obiekty należą [4].

$$wc(C) = \sum_{j=1}^{k} wc\left(C_j\right) = \sum_{j=1}^{k} \sum_{x(i)\in C_j} d\left(x(i), r_j\right)^2 \tag{25}$$

Przyjęcie tego odchylenia jako miary zwartości klastrów, prowadzi do generowania klastrów sferycznych (algorytm k-średnich). Odchylenie wewnątrzklastrowe możemy zdefiniować również, jako odległość punktu w klastrze do najbliższego obiektu w tym klastrze i wybieramy maksimum z tych odległości. Powyższa miara prowadzi do generowania klastrów podłużnych.

Odchylenie międzyklastrowe bc(C) jest sumą odległości środków wszystkich par klastrów.

$$bc(C) = \sum_{1 \le i < j \le k} d\left(r_i, r_j\right)^2 \tag{26}$$

gdzie dla powyższych wzorów:

d - odległość,
x(i) - obiekty,
r_i, r_j - środki klastrów,
k - liczba klastrów,
C - podział obiektów pomiędzy k klastrów.

Można przyjąć jako miarę jakości klasteryzacji stosunek odchylenia międzyklastrowego przez odchylenie wewnątrzklastrowe bc/wc. Obliczanie funkcji kryterialnej wymaga przejrzenia całego zbioru obiektów dla każdego pojedynczego podziału. Funkcja kryterialna, która za miarę przyjmuje tylko odchylenie wewnątrzklastrowe z pominięciem odchylenia międzyklastrowego, nazywamy funkcją błędu średniokwadratowego.

Indeks Daviesa-Bouldin'a [10] wprowadzony został jako metryka oceny algorytmów grupowania. Indeks ten uwzględnia rozproszenie wewnętrzne i odległość między klastrami. Rozproszenie wewnątrz klastra można wyrazić:

$$S = \left(\frac{1}{N} \sum_{i=1}^{N} \left| X_i - r \right|^q \right)^{1/q} \tag{27}$$

gdzie :

X_i - n wymiarowy wektor cech (np. odległości);
r - środek ciężkości;
N - liczba obiektów w klastrze;
q - parametr determinujący użytą metrykę.

Wskaźnik Daviesa-Bouldin'a jest definiowany jako:

$$R_{ij} = \frac{S_i + S_j}{M_{ij}} \tag{28}$$

gdzie :

S_i, S_j - rozproszenie wewnątrz klastra;
M_{ij} - odległość między klastrami.

Do obliczenia odległości między klastrami Ci i Cj najczęściej stosuje się jedną z miar odległości Minkowskiego, które opisano wcześniej.

Miara DB wyrażona jest wzorem:

$$DB_{ij} = \frac{1}{N}\sum_{i=1}^{N} \max_{j, j\neq i} R_{ij} \tag{29}$$

gdzie :

N – liczba klastrów

R_{ij} – wskaźnik Daviesa-Bouldin'a

Szukamy parametrów, które minimalizują miarę DB, aby uzyskać klastry o małym rozproszeniu wewnętrznym i leżące daleko od siebie. Niska wartość DB będzie wskazywała na dobry podział.

Indeks Dunna [11] jest metryką oceny algorytmów klastrowania. Miara ta jest podobna do miary DB i tak samo jest wykorzystywana do oceny w oparciu o kryterium wewnętrzne. Ma na celu identyfikację klastrów gęstych i dobrze oddzielonych. Wskaźnik ten definiowany jest jako stosunek minimalnej odległości między klastrami, do maksymalnej odległości wewnątrz klastra:

$$DI = \min_{1\leq i\leq n}\left\{ \min_{1\leq j\leq n, i\neq j}\left\{ \frac{d(i,j)}{\max_{1\leq k\leq n} d'(k)} \right\} \right\} \tag{30}$$

gdzie :

d(i,j) – odległość między klastrami,

d’(k) – odległość wewnątrz klastra k

Chcąc określić rozmiar lub średnicę klastra możemy zastosować różne miary odległości wewnątrz klastra (np. maksymalna odległość między dwoma punktami, średnia odległość między wszystkimi parami). Ponieważ kryterium wewnętrzne poszukuje klastrów o wysokim podobieństwie wewnątrz klastra i niskim między klastrami, to algorytmy, które produkują klastry o wysokim indeksie Dunna (większym niż 1) dają najlepsze rezultaty.

Metoda sylwetki (ang. silhouette clustering) po raz pierwszy została opisana przez Petera J. Rousseeuw w 1986 roku [12]. Jest ona syntetycznym wskaźnikiem jakości grupowania. Miara sylwetki porównuje średnią odległość do elementów w tym samym klastrze ze średnią odległością do elementów w innym klastrze. Im wyższa wartość tego parametru, tym lepsza klasteryzacja. Indeks sylwetki dobrze się sprawdza m.in. w metodzie k-means, gdzie znajduje zastosowanie do określania optymalnej liczby klastrów. Każdemu obiektowi przyporządkowana jest miara:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i); b(i)\}} \tag{31}$$

gdzie :

a(i) - średnie niepodobieństwo (np. miary odległości) pomiędzy i a pozostałymi obiektami w klastrze

$$a(i) = \frac{1}{m_j - 1} \sum_{j \in X_j, j \neq j} d(i, j) \tag{32}$$

Można interpretować a(i) jako miarę, która określa na ile dobrze i-ty element jest przypisany do klastra. Im mniejsza jest jego wartość, tym przyporządkowanie jest lepsze.

b(i) - najmniejsza średnia odległość pomiędzy i-tym elementem, a każdym z pozostałych klastrów

$$b(i) = \min_{X_k \neq X_j} \frac{1}{m_k} \sum_{j \in X_k} d(i, j) \tag{33}$$

Klaster z najmniejszą wartością nazwiemy klastrem sąsiednim do i-tego.

Miara s(i) przyjmuje wartości z przedziału <-1,1>. Wartość bliska 1 oznacza, że element jest prawidłowo przyporządkowany do klastra, natomiast bliska -1 mówi nam, że lepsze byłoby przypisanie i-tego elementu do klastra sąsiedniego. s(i) bliskie 0 oznacza, że obiekt leży na granicy dwóch klastrów i pasuje do nich w równym stopniu.

Dla każdego klastra oblicza się średnią wartość miary s(i) obiektów wchodzących w jego skład i oznacza się SI(c), gdzie c jest numerem danego klastra. Określa ona jak gęsto są zgrupowane obiekty leżące wewnątrz klastra. SI(c) dla całego zestawu danych jest miarą jakości klasteryzacji. Jeśli klastrów jest za dużo lub za mało, to część z nich będzie miała zbyt wąską sylwetkę w stosunku do pozostałych.

11 Funkcjonalność aplikacji

Aplikacja wczytuje z bazy danych przygotowany wcześniej zestaw danych, składający się z osób, które są członkami sieci społecznościowej, przyporządkowanych do klastrów. W bazie zostały utworzone m.in. dwie tabele: Persons i Clusters. W pierwszej z nich zapisano dane wygenerowanych losowo osób (nazwisko, imię, współrzędne geograficzne miejsca zamieszkania i inne), w drugiej definicje klastrów (m.in. nazwa, współrzędne geograficzne). Klastrami zostały miasta o liczbie ludności przekraczającej określony próg. Każda osoba została przypisana do najbliższego geograficznie klastra.

Tabele zbudowano z wykorzystaniem typów użytkownika SocNetCluster (implementujący klaster) i SocNetPerson (przechowujący dane

osoby) stworzonych w CLR [1], [2]. Dzięki temu wszystkie istotne dane zostały umieszczone w obiekcie w postaci binarnej i zapisane w jednej kolumnie tabeli.

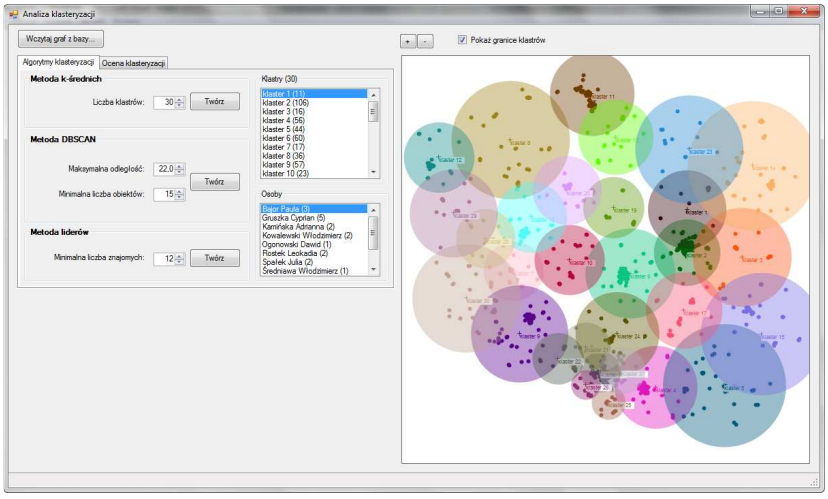

Fig. 1. Wygląd okna głównego aplikacji

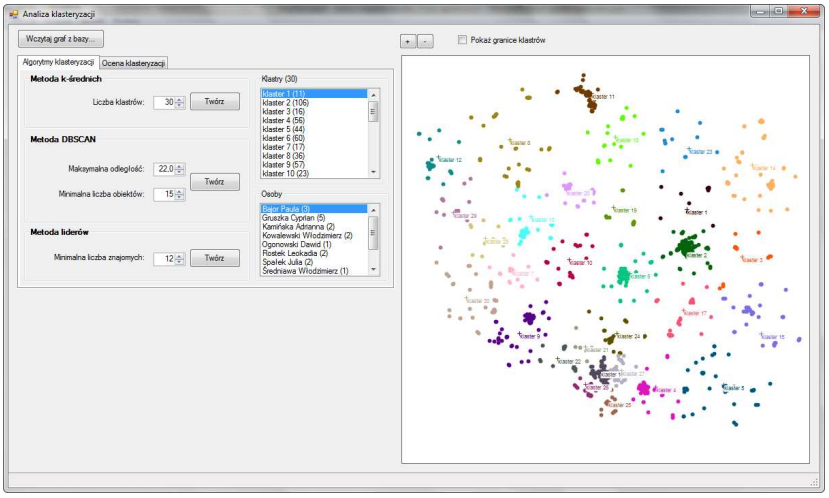

Fig. 2. Okno główne aplikacji (ukryte obszary klastrów)

Aplikacja została napisana w języku C#, w środowisku Microsoft Visual Studio Professional Edition, z zastosowaniem interfejsu Windows Form. Program składa się z jednego okna głównego oraz okna pomoc-

niczego służącego do wpisania parametrów połączenia z serwerem bazy danych. Główne okno aplikacji pozwala na wczytanie danych z bazy SQL Server, przeprowadzenie ponownej klasteryzacji grafu wybranym algorytmem oraz obliczenie różnych miar klasteryzacji. Przedstawione są także: aktualna lista klastrów i osoby przynależące do klastra zaznaczonego na liście. W celu ułatwienia wizualnej oceny jakości klasteryzacji, zaimplementowano także uproszczoną mapę schematyczną (panel), na której poszczególne osoby reprezentowane są jako okrągłe obiekty w kolorze zależnym od klastra przyporządkowania. Wygląd okna głównego przedstawiają rysunki (1-3).

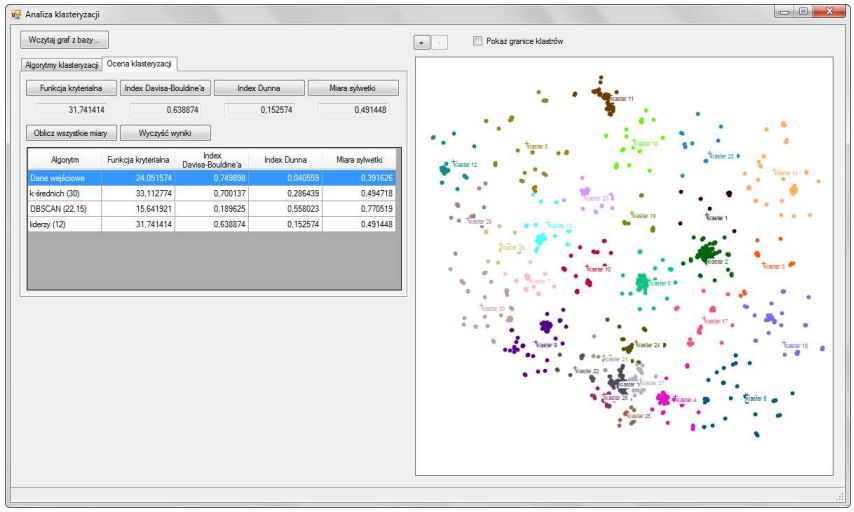

Fig. 3. Okno główne aplikacji (karta z wynikami funkcji oceny)

12 Testy i analiza otrzymanych wyników

Przeprowadzona została seria testów z użyciem stworzonej aplikacji, bazująca na losowym modelu sieci społecznościowej wczytanym z bazy danych. Jako dane referencyjne przyjęty został podział na klastry zdefiniowany w bazie, na których uruchomione zostały wszystkie cztery zaimplementowane funkcje oceny. Następnie przeprowadzone zostały dwie serie ponownych klasteryzacji, z zastosowaniem metod:

1. k-średnich – 10 grupowań, dla liczby klastrów od 10 do 100 ze skokiem wynoszącym 10.
2. liderów - 7 grupowań, dla minimalnej liczby znajomych od 10 do 40 ze skokiem wynoszącym 5.

Otrzymane wyniki przedstawione zostały w poniższych tabelach. Zaprezentowano również wykresy wartości poszczególnych miar oceny. Ze względu na znaczą rozbieżność zakresu wartości funkcji kryterialnej w stosunku do pozostałych miar, przebieg tej funkcji został przedstawiony na odrębnym wykresie.

Tabela. 1. Wyniki funkcji oceny dla danych wejściowych

Liczba klastrów	Funkcja kryterialna	Miara Daviesa-Bouldine'a	Miara Dunna	Miara sylwetki
32	24,051574	0,749898	0,040559	0,391626

Tabela. 2. Wyniki funkcji oceny dla metody k-średnich

Liczba klastrów	Funkcja kryterialna	Miara Daviesa-Bouldine'a	Miara Dunna	Miara sylwetki
10	0,9752	0,6062	0,4443	0,5074
20	8,2838	0,6386	0,3183	0,5033
30	26,7778	0,6591	0,3076	0,4950
40	69,1359	0,6639	0,2027	0,4788
50	145,7095	0,6710	0,1672	0,4587
60	248,0074	0,6607	0,1354	0,5106
70	340,6661	0,6744	0,0790	0,5064
80	475,9248	0,6527	0,1657	0,4812
90	783,3863	0,6215	0,0589	0,4984
100	960,7108	0,6573	0,0221	0,5007

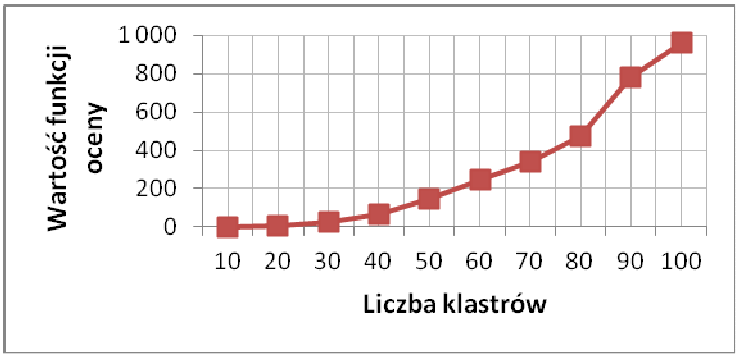

Fig. 4. Wykres funkcji kryterialnej dla metody k-średnich

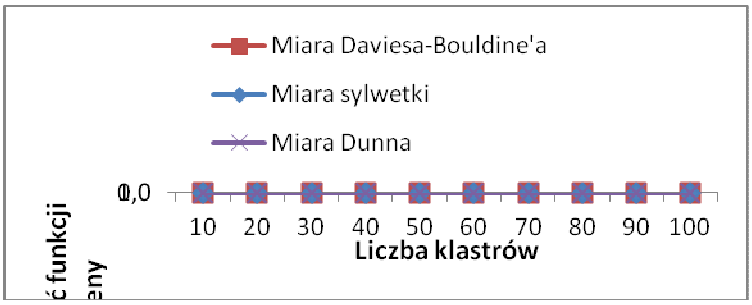

Fig. 5. Wykres trzech funkcji oceny dla metody k-średnich

Tabela. 3. Wyniki funkcji oceny dla metody liderów

Minimalna liczba znajomych	Funkcja kryterialna	Miara Daviesa-Bouldine'a	Miara Dunna	Miara sylwetki
10	99,7691	0,7133	0,1227	0,4357
15	5,3490	0,7093	0,1101	0,4366
20	1,7847	0,6439	0,2612	0,4709
25	0,6143	0,6712	0,2910	0,4359
30	0,0685	0,5652	0,3981	0,4288
35	0,0432	0,5872	0,4545	0,4061
40	0,0151	0,5682	0,5297	0,4385

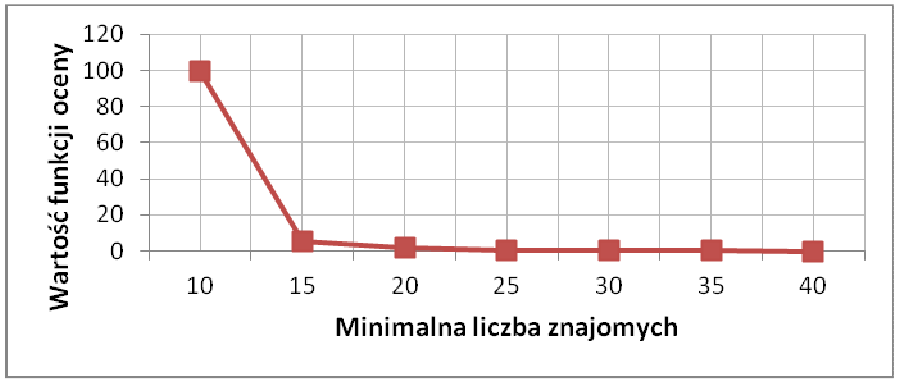

Fig. 6. Wykres funkcji kryterialnej dla metody liderów

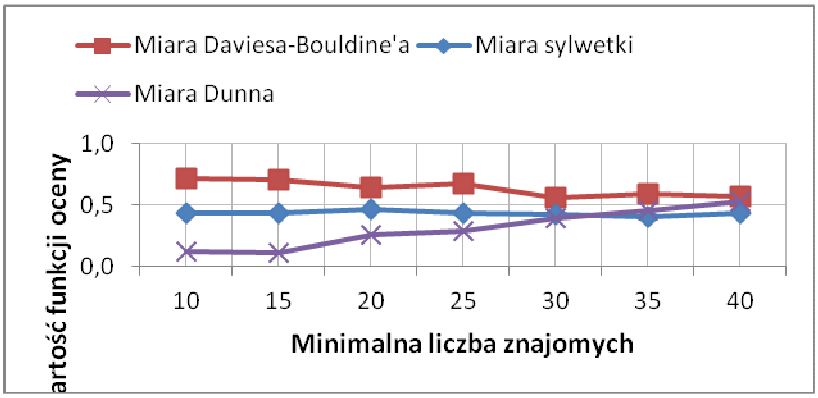

Fig. 7. Wykres trzech funkcji oceny dla metody liderów

13 Podsumowanie

Jak łatwo zauważyć, przebieg miar Daviesa-Bouldine'a oraz sylwetki zmienia się w niewielkim stopniu wraz ze zmianą liczby klastrów (dla algorytmu k-średnich) i minimalnej liczby znajomych (dla algorytmu liderów). Z kolei wartość funkcji kryterialnej zmienia się wykładniczo. Wartość miary Dunna posiada w przybliżeniu liniowy trend wzrostu w funkcji liczby klastrów i minimalnej liczby znajomych.

Z powyższych obserwacji można wyciągnąć wniosek, że pierwsze dwie miary są praktycznie niewrażliwe na przyjęte kryteria tworzenia klastrów. Ponadto należy zauważyć, że trzy funkcje oceny (miary Daviesa-Bouldine'a, Dunna i sylwetki) danych referencyjnych (klastrów wczytanych z bazy danych) są nieco gorsze od najgorszych ocen uzyskanych w pomiarach dla dwóch kolejnych serii klasteryzacji. Istotny wpływ na otrzymane rezultaty ma fakt, że wszystkie użyte algorytmy klasteryzacji operują na tej samej grupie obiektów (osób), których położenie nie ulega zmianie, a stosowano jedynie miarę odległości euklidesowej.

Sztuczna, wygenerowana losowo struktura sieci społecznościowej, wydaje się mało podatna na dobrej jakości grupowanie. W niniejszej pracy omówiono różne aspekty sieci społecznościowych, skupiając największą uwagę na ich analizie, ze szczególnym uwzględnieniem oceny klasteryzacji. Ze względu na specyfikę sieci społecznościowych, których elementami są ludzie, we wszystkich rozważaniach oraz algorytmach używana była wyłącznie miara odległości euklidesowej. Każda osoba ma bowiem zdefiniowane miejsce zamieszkaniaw postaci współrzędnych geograficznych.

14 Literatura

[1] P. Pilny, A. Pelikant, *Typy użytkownika CLR – wprowadzenie obiektowości do relacyjnej bazy danych*, Zeszyty Naukowe Wyższej Szkoły Informatyki w Łodzi, Vol. 11, Nr 2, 2012 ss. 51-81

[2] E. Konopka, A. Pelikant, *Funkcje i typy użytkownika CLR w zadaniach statystycznych,* Zeszyty Naukowe Wyższej Szkoły Informatyki w Łodzi, Vol. 11, Nr 2, 2012 ss. 5-30

[3] Robin J. Wilson, *Wprowadzenie do teorii grafów*, Wydawnictwo Naukowe PWN, Warszawa, 2007.

[4] A. Fronczak, P. Fronczak, *Świat sieci złożonych: Od fizyki do Internetu*, Wydawnictwo Naukowe PWN, Warszawa, 2009.

[5] D. J. Watts, S.H. Strogatz, *Collective dynamics of "small-world" networks*, Nature, Vol. 393, 440-442, 1998.

[6] A.-L. Barabási, R Albert, *Emergence of scaling in random networks*, Science, Vol. 286, 509-512, 1999.

[7] Z. Tarapata, *Czy sieci rządzą światem? - Od Eulera do Barabasiego*, WAT, Warszawa, 2012.

[8] Adam Lessnau, *Klasteryzacja*, 31.03.2005, http://subversion.assembla.com/svn/klasteryzacja/materialy/02Klasteryzacja.pdf

[9] http://wazniak.mimuw.edu.pl/index.php?title=ED-4.2-m11-1.0-Slajd7, Studia Informatyczne, BETA, 29.08.2006.

[10] J. Bezdek, N. Pal, *Some new indexes of cluster validity*. IEEE Transactions on Systems, Man, And Cybernetics.Part B: Cybernetics 28, (3), 301-315, 1998.

[11] J. C. Dunn, *A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Cluster*". Journal of Cybernetics 3 (3): 32–57,1973.

[12] Peter J. Rousseeuw, *Silhouettes: a Grafical Aid to the Interpretation and Validation of Cluster Analiysis*, Computational and Applied Mathematics 20: 53-65, 1987.

[13] Stephen C. Perry, *C# i .NET*, Helion, 2006.

[14] User–Defined Type Requiements. http://technet.microsoft.com/pl-pl/library/ms131082.aspx, Microsoft, 2014.

[15] A. Pelikant, *MS SQL Server. Zaawansowane metody programowania*, Helion, 2014,

THE USE OF CLUSTERING METHODS IN THE ANALYSIS OF SOCIAL NETWORKS

Summary – The purpose of work is to discuss the various methods of clustering in social networks. Analyzed data are initially divided into clusters according to the place of residence of the members of the network. Developed algorithm and application based on it evaluates clustering quality and enables redistribution according to various methods, and then comparing the results of their actions. There were implemented many algorithms which gives different results. The application works with database created on Microsoft SQL Server platform. Two user defined data types have been applied in CLR technologies that implement the objects corresponding to the component network-graph: person (SocNetPerson) and the cluster (SocNetCluster).

Keywords: Graph theory, graph analysis, social networks, user defined data types CLR, clustering, clustering quality evaluation, data exploration, object-oriented programming in databases.