STUDIA INFORMATICA 2013
Volume 34 Number 2A (111)

Agnieszka NOWAK-BRZEZIŃSKA, Tomasz XIĘSKI
Uniwersytet Śląski, Instytut Informatyki

METODY REPREZENTACJI DANYCH ZŁOŻONYCH

Streszczenie. Artykuł dokonuje przeglądu metod reprezentacji i wizualizacji danych, ze szczególnym uwzględnieniem technik graficznego przedstawienia skupień. Ponadto omawia algorytm wizualizacji struktur hierarchicznych w przestrzeni dwuwymiarowej (*Squarified Treemaps*) oraz prezentuje koncepcję jego zastosowania do rzeczywistego zbioru skupień danych złożonych, wygenerowanych przez gęstościowy algorytm grupowania OPTICS.

Słowa kluczowe: wizualizacja skupień, grupowanie, metody reprezentacji danych, OPTICS

METHODS OF COMPLEX DATA REPRESENTATION

Summary. This work reviews data representation and visualization methods, with emphasis on techniques for clusters' representation. Furthermore it describes an algorithm for hierarchical structures' visualization in a two-dimensional space (*Squarified Treemaps*) and presents the concept of its application to a real-world complex dataset composed of clusters generated by the OPTICS algorithm.

Keywords: cluster visualization, clustering, data representation methods, OPTICS

1. Wprowadzenie

Nieustannie rosnąca liczba gromadzonych danych doprowadziła do powstania wielu metod odkrywania wiedzy przez generowanie reguł asocjacyjnych, grupowanie, tworzenie klasyfikatorów itp. Analiza danych przy użyciu wymienionych technik jest jednak tym trudniejsza, im bardziej złożona jest ich struktura zarówno pod względem dużej liczby atrybutów opisujących każdy obiekt danych, jak i użytych typów danych – informacje zakodowane w bazie często opisane są atrybutami różnych typów, wliczając w to wartości binarne, dys-

kretne, ciągłe, kategoryczne, tekstowe czy daty. Tego typu dane można nazwać złożonymi i sposób ich reprezentacji będzie podstawą analizy w niniejszym artykule.

W literaturze przedmiotu pojęcie reprezentacji (modelu) danych definiuje się jako spójny zbiór zasad opisujących strukturę danych w zbiorze, który swym zakresem obejmuje: sposób reprezentacji dozwolonych w modelu danych oraz ich związków oraz zasady ich gromadzenia, przetwarzania czy modyfikacji [9]. Dla analityka danych istotniejsza od modelu reprezentacji danych jest reprezentacja występującej w nich wiedzy[1] w formie korelacji, trendów i powiązań. Dlatego też dalsze rozważania odnośnie do reprezentacji danych będą dotyczyć ukrytych w nich zależności i wzorców.

Metody reprezentacji wiedzy można podzielić na dwie główne grupy: opisowe oraz graficzne. Najpopularniejsze techniki opisowe wykorzystują rachunek zdań I lub II stopnia czy zapis w formie tablicy decyzyjnej. Równie często spotykany (ze względu na swoją prostotę i intuicyjność) jest zapis regułowy wykorzystywany np. w popularnym systemie Rough Set Exploration System[2]. Wymienione metody reprezentacji zostały już wielokrotnie szczegółowo zaprezentowane i przeanalizowane, dlatego autorzy artykułu zdecydowali się skupić na technikach graficznych coraz częściej stosowanych w obliczu narastającej liczby przechowywanych i przetwarzanych danych. Wśród technik tego typu można dokonać dalszej dekompozycji na pikselowe, geometryczne, symboliczne oraz inne [7].

1.1. Cele reprezentacji i wizualizacji danych

Graficzna reprezentacja danych może być wykorzystywana w realizacji jednego z trzech głównych celów: prezentacji informacji, weryfikacji sformułowanych hipotez lub analizy i eksploracji związków ukrytych w danych. Wizualizacja danych najczęściej kojarzy się z prezentacją informacji w formie graficznej odpowiednio skondensowanej i łatwo przyswajalnej dla potencjalnego odbiorcy. Pozwala to wyróżnić i szybko dostrzec najważniejsze elementy dużego zbioru danych. Równie istotnym celem wizualizacji jest weryfikacja i potwierdzenie sformułowanych hipotez – analitycy danych wykorzystują graficzną prezentację do poparcia słuszności wyciągniętych w procesie badawczym wniosków przed ich przedstawieniem. Pozwala to uniknąć błędów czy fałszywych teorii jeszcze na etapie badań bądź też pomaga w podjęciu decyzji o zmianie ich kierunku. Najbardziej istotna z punktu widzenia wydobywania wiedzy z danych wydaje się wizualizacja, która za cel stawia sobie analizę i eksplorację posiadanych danych, umożliwiając zilustrowanie niejawnie zawartej w nich wiedzy. Stanowi ona zatem nie tylko narzędzie pomocne przy zrozumieniu efektów uzyskanych przez zastosowanie odpowiedniej metody eksploracji danych, lecz także może być postrze-

[1] Reprezentacja wiedzy rozumiana jest jako niezależny od znaczenia rozpatrywanej informacji ogólny formalizm zapisywania i gromadzenia dowolnego fragmentu wiedzy.

gana jako zupełnie niezależna technika analiz, która opiera się na wrodzonych zdolnościach kognitywnych człowieka. W literaturze przedmiotu [7, 8] coraz częściej spotyka się pojęcie graficznej analizy eksploracyjnej (ang. *Visual Data Mining*), co dodatkowo potwierdza ważność i rosnącą popularność takiej formy wizualizacji i reprezentacji danych. Autor pracy [9] wyszczególnia wiele możliwych do realizacji celów badawczych wizualizacji eksploracyjnej, takich jak:

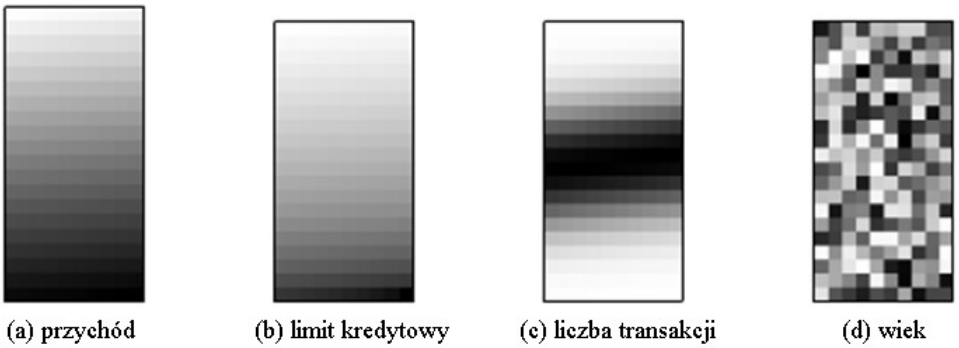
- ocena kompletności i prawidłowości zebranych danych,
- detekcja wyjątków, anomalii i obserwacji odstających,
- rozpoznanie i zrozumienie trendów, tendencji, podobieństw i symetrii,
- wyodrębnienie skupień,
- porównanie różnic między grupami,
- weryfikacja założeń dotyczących rozkładu danych,
- pomiar postępów i obserwacja procesu w czasie,
- przewidzenie oraz ocena potencjalnych przyszłych trendów.

Istotnym czynnikiem, który znacząco wpływa na sukces i realizację zdefiniowanych wcześniej zamiarów badawczych, jest wybór właściwej techniki reprezentacji oraz wizualizacji danych.

2. Graficzne metody reprezentacji danych

Pierwszą grupą metod graficznej reprezentacji danych, które zostaną omówione, są tzw. techniki pikselowe (ang. *pixel-oriented techniques*). Idea ich działania polega na wizualizacji wartości przyjmowanych przez dany atrybut przez przypisanie im określonych pikseli na ekranie oraz ustalonych wcześniej kolorów. Przykładowo, zakładając wizualizację w skali szarości, ciemny kolor piksela reprezentuje wysoką wartość danego atrybutu, natomiast im jest on jaśniejszy, tym wartość przyjmowana przez atrybut jest niższa. Dla zestawu *m*-wymiarowych danych techniki pikselowe tworzą *m* obszarów na ekranie (zwanych oknami), po jednym dla każdego wymiaru. Kolor piksela w każdym oknie na tej samej pozycji odzwierciedla wartości przypisane poszczególnym atrybutom określonego obiektu ze zbioru danych. Piksele wewnątrz okna są zorganizowane zgodnie ze ściśle określonym porządkiem globalnym, wspólnym dla każdego okna. Rysunek 1 ilustruje przykładowy zestaw danych czterowymiarowych, zawierających informacje o klientach hipermarketu i dokonywanych przez nich zamówieniach.

[2] Szczegółowe omówienie systemu RSES znajduje się na stronie http://logic.mimuw.edu.pl/~rses/.

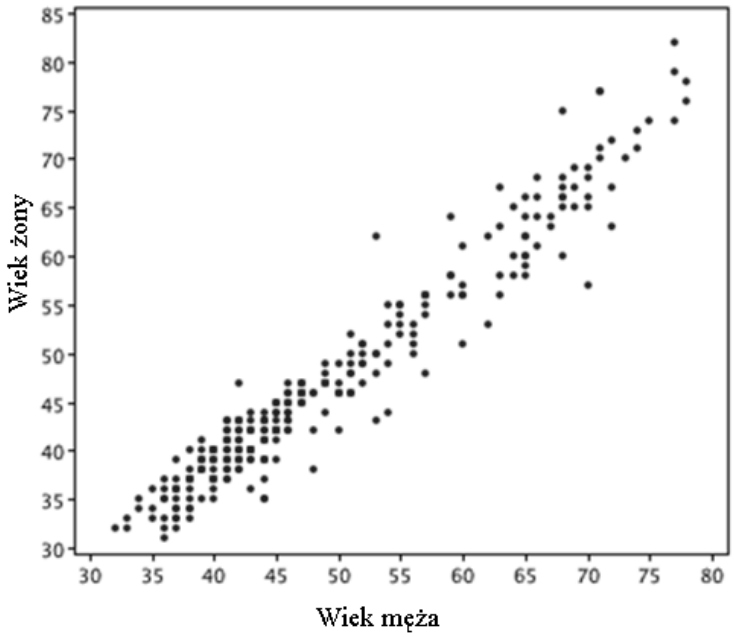


Rys. 1. Przykład użycia technik pikselowych [7]
Fig. 1. Usage example of pixel-oriented techniques [7]

Każdy klient jest charakteryzowany za pomocą swojego przychodu, limitu kredytowego, średniej liczby wykonywanych transakcji oraz wieku. Wizualizacja została wykonana przy założeniu, że ciemne kolory określają wysoką wartość danej cechy, natomiast jasne kolory sytuację odwrotną, a wszystkie obiekty w bazie zostały posortowane rosnąco według ich przychodu, co narzuca globalny porządek pikseli wewnątrz wszystkich czterech okien. Można zauważyć, że wraz ze wzrostem przychodu powiększa się również dostępny limit kredytowy. Ponadto najwięcej transakcji generują klienci o średnim przychodzie. Nie zaobserwowano jednak żadnej korelacji między wiekiem klienta a jego przychodem.

Niestety wizualizacja wielowymiarowych danych złożonych za pomocą prostokątnych okien jest dość kłopotliwa i negatywnie wpływa na czytelność tej techniki. Dlatego też w [2] poczyniono próbę ulepszenia tej techniki, reprezentując poszczególne okna jako wycinki koła. Pozwoliło to na zaprezentowanie danych 50-wymiarowych [1], jednakże w dalszym ciągu dla dużej liczby obiektów taka forma graficznej reprezentacji jest trudna w analizie.

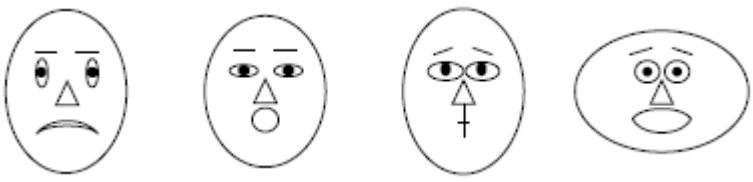
Istotną wadą technik pikselowych jest fakt, że nie pozwalają one na przedstawienie rozkładu danych w przestrzeni. Przykładowo nie mogą one pokazać podobieństwa obiektów względem siebie w przestrzeni celem zbadania, czy w zestawie danych występują naturalne skupienia obiektów. Techniki geometryczne próbują wyeliminować tę niedoskonałość, osadzając analizowane dane w przestrzeni dwu- bądź trójwymiarowej. Najpopularniejszym przedstawicielem tej techniki jest tzw. wykres rozrzutu (ang. *scatter plot*). Służy on do wizualizacji zwykle danych dwuwymiarowych (gdzie zbiory wartości atrybutów są zaznaczone na osiach odciętych i rzędnych), jednakże kolejny wymiar może zostać łatwo dodany przez użycie różnych kolorów lub kształtów do reprezentowania obiektów danych. Można również stworzyć wykres trójwymiarowy, dodając kolejną oś współrzędnych, co w połączeniu z kolorowaniem wykresu pozwala na przedstawienie danych czterowymiarowych.



Rys. 2. Przykładowy wykres rozrzutu dla danych dwuwymiarowych
Fig. 2. A sample scatter plot for two-dimensional data
Źródło: http://onlinestatbook.com/chapter4/pearson.html

Rysunek 2 prezentuje prosty przykład wykorzystania wykresu rozrzutu w analizie eksploracyjnej. Niech celem analityka danych będzie zbadanie zależności między wiekiem mężczyzn a wiekiem ich żon przy wykorzystaniu danych hipotetycznego serwisu społecznościowego. Na podstawie zaprezentowanego wykresu można stwierdzić, że występuje niemalże liniowa zależność między wiekiem męża i żony – wraz ze wzrostem wieku męża wzrasta proporcjonalnie również wiek jego żony. Można również powiedzieć, że mężczyźni wybierają kobiety w tym samym wieku bądź nieznacznie młodsze od siebie.

Dla danych o większej liczbie wymiarów stosuje się macierz wykresów rozrzutu (ang. *scatter plot matrix*). Uwzględniane są wówczas wszystkie możliwe kombinacje par atrybutów. Przy wizualizacji danych *n*-wymiarowych powstaje zatem macierz kwadratowa *n* x *n*, której każdy element to wykres rozrzutu. Niestety efektywność i czytelność tej metody reprezentacji maleje wraz ze wzrostem liczby wymiarów, przez co jej bezpośrednie zastosowanie do danych złożonych jest niezalecane lub w pewnych przypadkach niemożliwe.



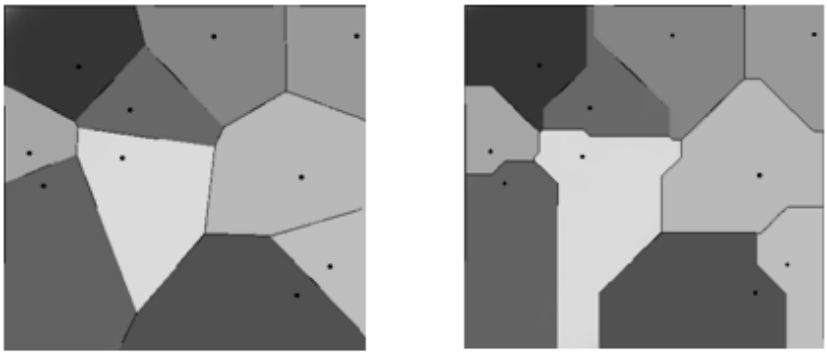
Rys. 3. Wizualizacja danych za pomocą twarzy Chernoffa [7]
Fig. 3. Data visualization using Chernoff faces [7]

Rozwiązanie problemu wizualizacji danych wielowymiarowych doprowadzało czasami do bardzo oryginalnych pomysłów. Przykładem są tzw. twarze Chernoffa, zaprezentowane na

rysunku 3. Podobieństwo lub niepodobieństwo danych wielowymiarowych odzwierciedlono tutaj w mimice twarzy (poszczególne cechy obiektów przedstawione są jako odpowiednio ułożone elementy głowy człowieka, takie jak oczy, uszy, usta, brwi itp.). Przykładowo duży rozstaw oczu będzie oznaczał dużą wartość danej cechy, natomiast mały rozstaw – odwrotnie. Metoda ta należy do grupy metod wykorzystujących naturalne predyspozycje człowieka do zwracania uwagi na różnice w wyglądzie poszczególnych twarzy. Niezaprzeczalną wadą tej techniki jest ograniczenie polegające na wizualizacji danych maksymalnie 36-wymiarowych – w przypadku założenia asymetrii twarzy [7]. Podobnie analiza dużej liczby obiektów przez człowieka jest czasochłonna i znacząco utrudniona, dlatego wspomniana metoda nie powinna być stosowana w odniesieniu do danych złożonych.

2.1. Wizualizacja skupień

Zarówno omówione w poprzedniej sekcji techniki, jak i klasyczne metody wizualizacji w postaci wykresów liniowych, słupkowych, kołowych czy grafów, mimo dużej popularności i prostoty, są również nieczytelne w przypadku wizualizacji dużych zbiorów danych złożonych. Dlatego też autorzy niniejszego artykułu postanowili wykorzystać gęstościowy algorytm analizy skupień OPTICS[3] zarówno jako metodę ekstrakcji wiedzy (przez analizę utworzonych grup), jak i sposób na zmniejszenie liczby danych poddawanych procesowi wizualizacji (graficznie przedstawione będą wyłącznie wygenerowane grupy, a nie poszczególnie obiekty). Dokonano zatem przeglądu metod graficznej reprezentacji skupień celem wyboru najlepszej w odniesieniu do grup danych złożonych.



Rys. 4. Porównanie diagramów Woronoja dla różnych miar odległości
Fig. 4. Comparison of Voronoi diagrams for different distance measures
Źródło: http://en.wikipedia.org/wiki/Voronoi_diagram

Dwie najpopularniejsze techniki wizualizacji skupień to niewątpliwie reprezentacja w formie dendrogramu lub diagramu Woronoja. Pierwsza wymieniona technika opiera się na

[3] Motywacja do wyboru algorytmu gęstościowego, a także struktura rzeczywistego zbioru danych złożonych oraz informacje o utworzonym na jego podstawie zestawie grup zostały podane w [10], [12] i [14].

strukturze drzewiastej i jest stosowana w przypadku występowania hierarchii skupień. Największą jej wadą jest mała czytelność w przypadku dużej liczby skupień. W przypadku diagramów Woronoja uzyskuje się wizualizację płaskiego podziału, bez możliwości zamodelowania wprost struktury hierarchicznej (chyba że diagram ten będzie generowany dla każdego poziomu hierarchii z osobna). Ponadto rozkład i kształt skupień na diagramie Woronoja są bardzo zależne od przyjętej miary podobieństwa obiektów. Taka sytuacja (dla sztucznego zbioru danych podzielonego na 10 grup) została zaprezentowana na rysunku 4. Ze względu na fakt, że żadna z wymienionych metod wizualizacji struktury grup nie spełnia wymogów odnośnie do przedstawienia tej struktury w sposób jasny i czytelny, autorzy zdecydowali się wykorzystać algorytm generowania tzw. map prostokątów (ang. *treemaps*).

3. Wizualizacja danych za pomocą map prostokątów

Mapy prostokątów to technika wizualizacji struktur hierarchicznych na płaszczyźnie dwuwymiarowej, dzieląca rekurencyjnie dostępną przestrzeń roboczą ekranu na wiele prostokątów, których wielkość jest zależna od wartości przyjętego parametru ilościowego. Motywacją do jej powstania była potrzeba graficznego przedstawienia zajętości dysku twardego przez poszczególne pliki i katalogi w celu szybkiej identyfikacji zasobów w znaczny sposób wpływających na jego zapełnienie – umożliwiłoby to przeniesienie lub usunięcie niepotrzebnych danych i zwiększenie puli dostępnego miejsca. Do dnia dzisiejszego pozostaje ona bardzo wygodną i czytelną formą wizualizacji, zdolną do przedstawienia dużej ilości danych, czego dowodem jest jej wykorzystanie w wielu programach przedstawiających strukturę danych na dysku, jak np. w aplikacji SpaceSniffer[4].

Idea działania tego typu aplikacji jest bardzo prosta: użytkownik wskazuje, którą część dysku chce poddać analizie, po czym jest mu prezentowana mapa prostokątów (odzwierciedlająca katalogi bądź pliki na dysku twardym), których wielkość jest zależna od rozmiaru danego zasobu[5] – im większy jest rozmiar pliku, tym większy jest narysowany prostokąt. Elementy reprezentujące katalogi są również dzielone na mniejsze prostokąty celem zaprezentowania hierarchii zgodnej ze strukturą zasobów na dysku. Dzięki wykorzystaniu całego obszaru ekranu użytkownik może z łatwością zidentyfikować największe pliki i katalogi oraz podjąć na tej podstawie decyzję o ich ewentualnym przeniesieniu lub usunięciu w celu zwiększenia wolnego miejsca, nawet w przypadku dużej liczby danych, często niemożliwych do przedstawienia czy analizy przy użyciu innych technik. Wspomniana metoda graficznej

[4] Opis oraz zrzuty ekranu aplikacji SpaceSniffer dostępne są na stronie producenta: http://www.uderzo.it/main_products/space_sniffer/.
[5] Pojęcie zasobu jest tutaj utożsamiane z konkretnym plikiem lub katalogiem na dysku twardym.

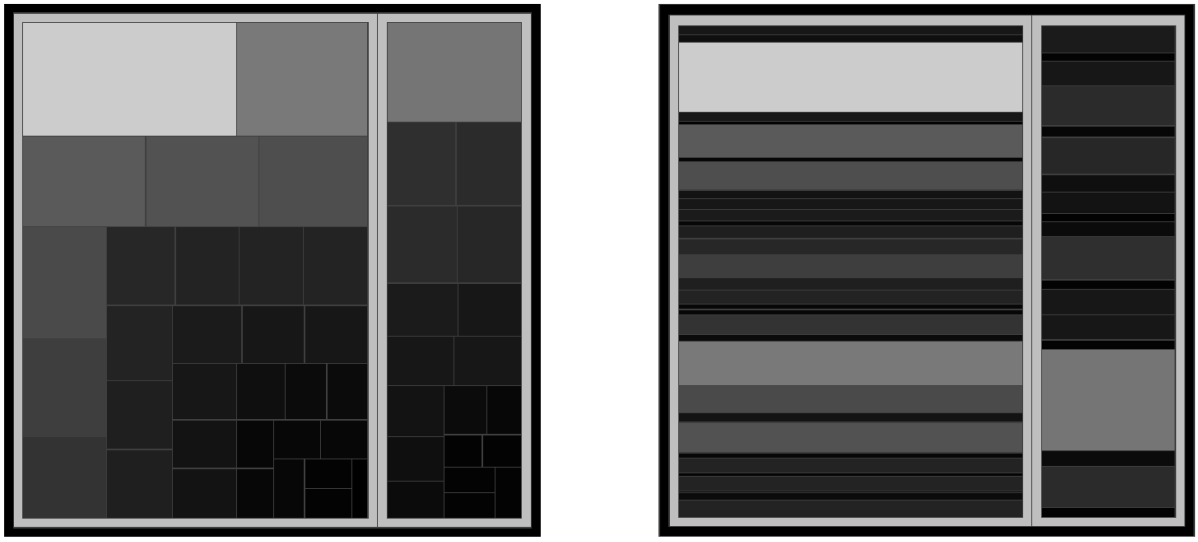
reprezentacji informacji wykorzystywana jest jednak z powodzeniem w wielu innych zastosowaniach, jak segmentacja rynku, monitorowanie sprzedaży, katalogowanie oferowanych produktów [3], [5]. Autorzy niniejszego artykułu, zainspirowani możliwościami wizualizacji dużych zbiorów danych wielowymiarowych na płaszczyźnie dwuwymiarowej, postanowili dostosować i wykorzystać mapy prostokątów jako środek do przedstawienia struktury grup dla rzeczywistego zbioru danych złożonych.

3.1. Mapy prostokątów jako narzędzie analizy eksploracyjnej zbioru danych złożonych

Poprzednie prace badawcze autorów (opisane m.in. w [13], [14]) dotyczyły wykrywania zależności w zbiorze danych złożonych zawierającym dane dotyczące funkcjonowania konkretnych urządzeń nadawczo-odbiorczych ustawionych w różnych regionach Śląska. Proponowane podejście do problemu uwzględniało zastosowanie algorytmu gęstościowego OPTICS celem osiągnięcia struktury grup oraz analizę uzyskanych powiązań przez wygenerowanie zrozumiałych reprezentantów skupień, a także zastosowanie graficznej techniki wizualizacji. Niestety wykorzystanie tzw. wykresu osiągalności w celach graficznego przedstawienia powiązań i struktury danych okazało się nieefektywne (i niemożliwe do bezpośredniej aplikacji bez ograniczenia danych) w kontekście analizowanego zbioru [14]. Dlatego też dalszym kierunkiem badań była próba dostosowania koncepcji map prostokątów do wizualizacji stworzonej struktury grup. Mimo że technika ta pierwotnie służy do wizualizacji hierarchii, to nic nie stoi na przeszkodzie, aby zastosować ją do płaskiej struktury skupień.

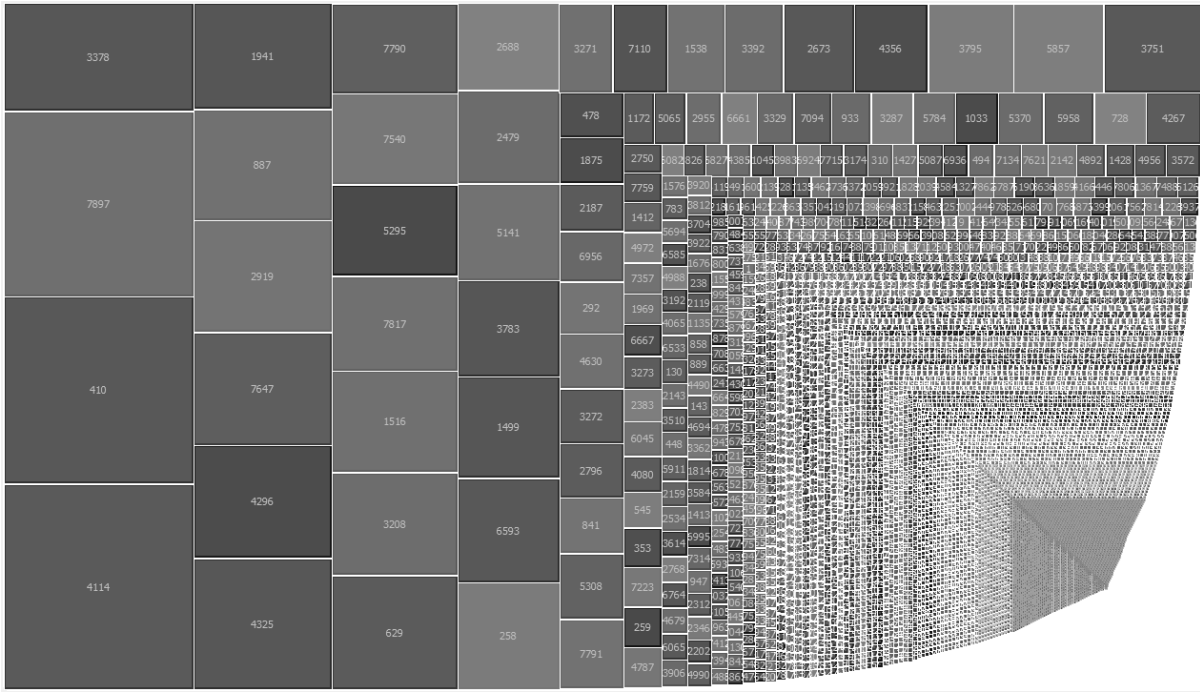
Istnieje kilka metod generowania podziału obszaru roboczego na prostokąty, gdyż dla ustalonej liczby prostokątów można zaproponować wiele różnych ułożeń, pozostawiając ich pole powierzchni bez zmian. Istotnym parametrem branym pod uwagę przy ocenie jakości danej metody generowania takiego podziału jest zdolność do zachowania proporcji między długością a szerokością prostokąta, która powinna być bliska jedynce. Dzięki temu porównanie powstałych kształtów jest łatwiejsze dla analityka posługującego się tą metodą. Rysunek 5 przedstawia różnicę między dwoma popularnymi sposobami tworzenia podziału: Slice and Dice [11] oraz Squarified [4]. Zadaniem obu algorytmów jest podział identycznego obszaru roboczego na dokładnie taką samą liczbę prostokątów. W przypadku pierwszej techniki tworzone są często bardzo cienkie prostokąty, co negatywnie wpływa na czytelność wizualizacji i może być przyczyną wysunięcia błędnych wniosków na temat prezentowanych danych. Dlatego też autorzy po zapoznaniu się z różnymi technikami generowania podziału postanowili wybrać metodę określaną jako Squarified[6] do graficznej reprezentacji wygenerowanej w poprzednich etapach badań struktury grup (omówionej w [10]).

[6] Szczegółowe porównanie oraz omówienie algorytmów generowania map prostokątów znajduje się w [5].



Rys. 5. Porównanie metod generowania map prostokątów [5]
Fig. 5. Comparison of treemaps' generation methods [5]

Proponowana przez autorów koncepcja zastosowania techniki generowania map prostokątów do analizowanej struktury grup (powstałej przez zastosowanie algorytmu gęstościowego) jest następująca. Dany prostokąt symbolizuje konkretną grupę, a jego pole powierzchni jest tożsame z liczbą obiektów wchodzących w skład reprezentowanego skupienia. Dodatkowo z każdym prostokątem powiązane są informacje, takie jak opis reprezentanta danego skupienia czy minimalne, maksymalne i średnie wartości dla atrybutów obiektów wchodzących w skład określonej grupy. Należy jednak zaznaczyć, że pole powierzchni danego prostokąta może symbolizować dowolny parametr ilościowy istotny z punktu widzenia zadania eksploracji danych. To analityk danych powinien zdecydować, jaka informacja do zbadania jest dla niego najistotniejsza w danej chwili – może to być średnia wartość np. zarejestrowanej liczby zdarzeń w ciągu dnia dla wszystkich obiektów skupienia, przez co możliwa jest identyfikacja urządzeń najczęściej wadliwych lub poddawanych naprawom. Prostokątom można tak przydzielać kolory, by – podobnie jak dla technik pikselowych – kolor określał średnią wartość analizowanej cechy wśród obiektów danego skupienia. W przypadku badanego zbioru interesującym parametrem może być stopień niedostępności danego urządzenia nadawczo-odbiorczego. I tak, przykładowo, ciemny kolor danego prostokąta oznaczałby sytuację, w której średnia wartość tego atrybutu wśród wszystkich obiektów danej grupy jest bardzo wysoka – byłaby to grupa urządzeń często niedostępnych. Jasny kolor prostokąta oznacza natomiast stan odwrotny. Dzięki takiemu podejściu analityk danych ma bezpośredni wgląd w rozkład struktury grup, co w przypadku ich dużej liczebności może znacząco ułatwić analizę i wyciągnięcie wniosków z procesu grupowania. Należy również nadmienić, że celem omówionej metody wizualizacji nie jest przedstawienie wszystkich cech dla obiektów wielowymiarowych, ale skupienie się na tych najbardziej istotnych w danym momencie.



Rys. 6. Mapa prostokątów prezentująca strukturę grup
Fig. 6. Treemap presenting a cluster structure

Rysunek 6 przedstawia mapę prostokątów wygenerowanych dla badanego zbioru 7933 grup urządzeń nadawczo-odbiorczych. Mimo iż technika wizualizacji pozwoliła na przedstawienie rozkładu wszystkich grup oraz wykorzystuje niemalże w pełni dostępny obszar roboczy ekranu, to jednak w dalszym ciągu analiza takiej struktury jest czasochłonna i dość skomplikowana. Dodatkowo można zauważyć problem z właściwym rozmieszczeniem prostokątów w prawym dolnym rogu ekranu. Wynika to z faktu, że mapy prostokątów zajmują dokładnie cały obszar roboczy tylko w przypadku, gdy suma pól wszystkich prostokątów jest równa polu powierzchni obszaru roboczego (lub jego wielokrotności). Dlatego też celem dalszych eksperymentów i analiz jest znalezienie metody rozwiązania tego problemu, a także zastosowanie hierarchicznego algorytmu analizy skupień, ale w odniesieniu do reprezentantów już stworzonych grup. Pozwoli to na wygenerowanie hierarchicznej struktury grup oraz wizualizację wyłącznie wybranych i interesujących (z punktu widzenia ekstrakcji wiedzy) zależności.

4. Podsumowanie

Celem niniejszego artykułu był przegląd spotykanych w literaturze technik graficznej reprezentacji wielowymiarowych danych złożonych, ze szczególnym uwzględnieniem technik prezentacji skupień. Kolejnym istotnym celem było zbadanie i omówienie techniki generowania map prostokątów w kontekście analizy eksploracyjnej rzeczywistego zbioru danych

złożonych. Zastosowanie metody map prostokątów – mimo jej niezaprzeczalnych zalet, jak możliwość wizualizacji hierarchii czy wykorzystywanie w pełni obszaru roboczego [11] – uwypukliło również wady tej techniki, szczególnie widoczne w przypadku dużej liczby danych. Problematyka prawidłowego i czytelnego rozmieszczenia elementów na wizualizacji przy zachowaniu właściwych proporcji powinna być przedmiotem dalszych badań i analiz.

BIBLIOGRAFIA

1. Ankerst M., Breunig M. M., Kriegel H. P., Sander J.: Optics: Ordering points to identify the clustering structure. SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data, Philadelphia, USA 1999.
2. Ankerst M., Keim D. A., Kriegel H. P.: Circle Segments. A Technique for Visually Exploring Large Multidimensional Data Sets. Proceedings of International Conference on Visualization '96, Hot Topic Session, USA 1996.
3. Shneiderman B.: Discovering Business Intelligence Using Treemap Visualizations. Dostępny w internecie: http://www.b-eye-network.com/view/2673.
4. Bruls M., Huizing K., Van Wijk J.: Squarified Treemaps. Proceedings of the Joint Eurographics and IEEE TCVG Symposium on Visualization, USA 1999.
5. Engdahl B.: Ordered and Unordered Treemap Algorithms and Their Applications on Handheld Devices Master's Degree Project. Stockholm, Sweden 2005.
6. Paradowski M. B.: Wizualizacja danych – dużo więcej niż prezentacja, [w:] Kluza M. (red.): Wizualizacja wiedzy: Od Biblia Pauperum do hipertekstu. Wiedza i Edukacja, Lublin 2011.
7. Han J., Kamber M., Pei J.: Data Mining: Concepts and Techniques, Third Edition. Morgan Kaufmann Publishers, USA 2011.
8. Myatt G. J., Johnson W. P.: Making Sense of Data II. John Wiley & Sons, Inc., USA 2009.
9. Felcenloben D.: Geoinformacja – wprowadzenie do systemów organizacji danych i wiedzy. Gall, Katowice 2011.
10. Nowak-Brzezińska A., Xięski T.: Grupowanie danych złożonych. Studia Informatica, Vol. 32, No. 2A (96), Gliwice 2011, s. 391÷402.
11. Johnson B., Shneiderman B.: Tree-Maps: a space-filling approach to the visualization of hierarchical information structures. Proceedings of the 2nd conference on Visualization '91. IEEE Computer Society Press, USA 1991.

12. Wakulicz-Deja A., Nowak-Brzezińska A., Xięski T.: Efficiency of complex data clustering. Lecture Notes in Artifitial Intelligence. Proceedings of 6th International Conference on Rough Sets and Knowledge Technology, RSKT 2011, Canada 2011.
13. Xięski T.: Grupowanie danych złożonych, [w:] Wakulicz-Deja A. (red.): Systemy wspomagania decyzji. Wydawnictwo Uniwersytetu Śląskiego, Katowice 2011.
14. Xięski T., Nowak-Brzezińska A.: Gęstościowa metoda grupowania i wizualizacji danych złożonych. Studia Informatica, Vol. 33, No. 2A (105), Gliwice 2012, s. 453÷464.

Wpłynęło do Redakcji 9 stycznia 2013 r.

Abstract

In this paper the topic of representation and visualization of a complex data structure is discussed. Authors review currently found in literature algorithmic solutions that are used to graphically present data, focusing on their disadvantages and problems when dealing with high-dimensional data. What is more the authors analyze a method for visualizing hierarchical structures called treemaps. Furthermore a concept of applying this technique to a real-world cluster data set (aggregating information about the functioning of a cellular phone operator's transceivers located in Poland) is proposed.

Adresy

Agnieszka NOWAK-BRZEZIŃSKA: Uniwersytet Śląski, Instytut Informatyki,
ul. Będzińska 39, 41-200 Sosnowiec, Polska, agnieszka.nowak@us.edu.pl.
Tomasz XIĘSKI: Uniwersytet Śląski, Instytut Informatyki,
ul. Będzińska 39, 41-200 Sosnowiec, Polska, tomasz.xieski@us.edu.pl.