STUDIA INFORMATICA 2011
Volume 32 Number 2A (96)

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

GRUPOWANIE DANYCH ZŁOŻONYCH

Streszczenie. Artykuł stanowi wprowadzenie do tematyki grupowania danych złożonych i przeszukiwania takiej struktury. Przedstawia problemy z tym związane, skupiając się przede wszystkim na aspekcie tworzenia reprezentantów skupień. Przeprowadzone eksperymenty opierające się na wykorzystaniu algorytmu DBSCAN, pozwalają na porównanie efektywności wyszukiwania, relewantnych do zadanego pytania skupień, w zależności od sposobu tworzenia reprezentantów grup.

Słowa kluczowe: DBSCAN, dane złożone, grupowanie, eksploracja danych

CLUSTERING COMPLEX DATA

Summary. This work provides an introduction to the matter of clustering complex data and searching through such a structure. It presents related problems, focusing primarily on the aspect of creating cluster representatives. Carried out experiments based on using the DBSCAN algorithm allow to compare the efficiency of finding relevant to the given question clusters, depending on the way of cluster representatives were created.

Keywords: DBSCAN, complex data, clustering, data

1. Wprowadzenie

Niewątpliwie opracowane dotąd metody grupowania nie nadają sie do zastosowania do danych złożonych. Jedne są efektywne tylko i wyłącznie w zastosowaniu do danych ilościowych, inne osiągają dobre rezultaty, gdy operują na danych jakościowych [3]. Nie ma natomiast opracowanych solidnie metod grupowania danych różnych typów: zarówno opisanych atrybutami jakościowymi, jak i ilościowymi [2]. Mało tego, gdy tych danych jest wiele, zarówno cech opisujących obiekty, jak i samych obiektów, dane te możemy nazwać złożonymi i takie zbiory w niniejszym artykule będą podstawą analizy. Poprzez dane złożone rozumie

się zatem duże ilości danych różnego typu: ciągi znaków, daty, liczby całkowite i rzeczywiste, gromadzone w hurtowniach i bazach danych.

Prócz opracowania efektywnych algorytmów grupowania danych złożonych powstaje problem poprawnego opisu tak powstałych grup obiektów. Grupy te mogą mieć bardzo nieregularne kształty (przykładowo mogą w pewnym stopniu nachodzić na siebie) lub mogą być podobne do siebie tylko w pewnym, niewielkim zakresie [4]. Aspekt tworzenia poprawnych i zrozumiałych reprezentantów grup jest istotny, nie tylko w kontekście właściwej analizy i interpretacji znalezionych powiązań, ale ma on również znaczący wpływ na proces przeszukiwania utworzonej struktury skupień.

Celem niniejszego artykułu jest przybliżenie problematyki grupowania danych złożonych oraz przeszukiwania takiej struktury, w zależności od przyjętej konwencji tworzenia reprezentantów grup. Wszystkie kwestie poruszane w artykule zostaną omówione na przykładzie rzeczywistego problemu eksploracji danych, dotyczącego telefonii komórkowej.

1.1. Opis rzeczywistego problemu eksploracji danych

W wyniku współpracy z jedną z firm świadczącą usługi telekomunikacyjne, autorzy weszli w posiadanie rzeczywistej bazy danych, gromadzącej informacje na temat działania i dostępności urządzeń nadawczo-odbiorczych, rozlokowanych w regionie katowickim. Celem tejże współpracy jest próba rozwiązania następującego problemu eksploracji danych: *co wpływa na wysoką niedostępność urządzeń nadawczo-odbiorczych w regionie katowickim*?
Aby dobrze zrozumieć specyfikę i złożoność tego problemu, należy zdefiniować kilka pojęć związanych z prawidłowym funkcjonowaniem telefonii komórkowej.

Systemy telefonii komórkowej od innych bezprzewodowych systemów łączności radiowej odróżniają dwie najistotniejsze cechy:

- komórkowa struktura sieci. Sieć składa się z wielu urządzeń nadawczo-odbiorczych (tzw. komórek), z których każda jest obsługiwana przez określoną stację bazową. Komórki te są różnych rozmiarów, w zależności od stopnia skomplikowania terenu i skupienia abonentów,
- ciągła aktualizacja stanu aktywnych telefonów komórkowych, znajdujących się w zasięgu określonej stacji bazowej. Aktualizacja ta ma na celu lokalizację przemieszczających się abonentów. Może być dokonywana automatycznie, na bieżąco – podczas inicjowania każdego połączenia lub okresowo – podczas przemieszczania się abonenta z zasięgu jednej komórki do drugiej.

Dwa najważniejsze elementy – biorąc pod uwagę topologię sieci telefonii komórkowej – zapewniające jej działanie to wspomniana już stacja bazowa oraz kontroler takiej stacji. W skład każdej stacji bazowej (ang. base transceiver station) wchodzą następujące elementy:

- komórki (ang. *cells*) – urządzenia nadawczo-odbiorcze,
- wzmacniacz sygnału (ang. *power amplifier*),
- przełącznik antenowy (ang. *duplexer*),
- łącznik sygnału (ang. *combiner*),
- system kontrolno-alarmowy (ang. *alarm and control system*).

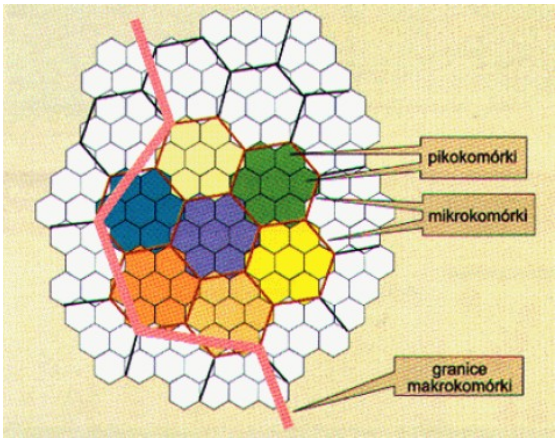
Należy tutaj również nadmienić, że stacja bazowa obsługuje wiele komórek oraz zapewnia bezprzewodową łączność między terminalem abonenta (telefonem) a infrastrukturą operatora telefonii komórkowej.

Kolejnym istotnym urządzeniem, ściśle powiązanym ze stacjami bazowymi, jest kontroler stacji bazowych (ang. *Base Station Controller*), który odpowiada za logikę działania tychże stacji. Do jego zadań należy m.in.:

- wybór i przydzielanie odpowiednich kanałów radiowych,
- kontrola przekazywania obsługi telefonu komórkowego z jednej stacji bazowej do drugiej,
- odbiór i przetwarzanie parametrów pomiarowo-identyfikacyjnych telefonów komórkowych.

Zwykle pod jednym kontrolerem pracuje od kilku do kilkuset stacji bazowych.

Aby zapewnić ciągłość zasięgu minimalnym kosztem na danym obszarze geograficznym, stacje bazowe powinny być ułożone na kształt plastra miodu. Cały obszar, który należy pokryć jest więc dzielony na heksagonalne regiony, w których centrum znajduje się stacja bazowa [9]. Ilustruje to rys. 1.



Rys. 1. Tworzenie obszarów komórkowych
Fig. 1. Creation of cellular regions

Niestety w rzeczywistych warunkach kształt ten jest daleki od idealnego. Charakter zabudowy i konfiguracji ulic, wysokość budynków czy nieregularne ukształtowanie terenu mają znaczący (często negatywny) wpływ na zasięg i jakość połączenia. Dlatego też stacje bazowe rozmieszone są dużo bliżej siebie niż pozwala na to teoria.

1.2. Struktura zestawu danych

Zestaw danych, stanowiący przedmiot analizy, agregował dane dotyczące urządzeń nadawczo-odbiorczych w regionie katowickim, pochodzące z okresu od kwietnia do listopada 2010 roku. Struktura każdego rekordu danych, wraz z przykładowymi wartościami, została przedstawiona w tabeli 1.

Pomiar dostępności danej komórki dokonywany był w godzinnych interwałach czasu. Znaczenie poszczególnych atrybutów jest następujące:

- **cellname** – identyfikator określonej komórki,
- **startTime** – godzina i data startu pomiaru,
- **data** – data dokonania pomiaru (pole pozostawione celowo w strukturze),
- **strata** – bezwzględny procent niedostępności danej komórki w danej godzinie,
- **strataWzglRegionu** – niedostępność komórki określana względem danego regionu,
- **strataWzglMikroRegionu** – niedostępność komórki określana względem danego mikroregionu,
- **czyProblem** – określenie czy występuje jakiś problem z daną komórką (została wyłączona celowo, ze względu na zaplanowane prace bądź też z innych powodów),
- **czyWoINN** – dział utrzymania sieci ma zlecenie na wykonywanie prac przy danej komórce,
- **czyWoTeren** – inny dział niż utrzymania sieci ma zlecenie na wykonywanie prac przy danej komórce,
- **czyWorkflow** – zostało wystawione zlecenie na dokonanie prac przy komórce,
- **czestosc** – liczba zdarzeń jakie odnotowano przez cały dzień, związanych z pracą określonej komórki,
- **czasTrwaniaH** – czas trwania określonego zdarzenia, wyrażony w godzinach,
- **typPrbId** – określa typ problemu, jaki wystąpił z daną komórką związanego z zasilaniem, transmisją, sprzętowy lub inny,
- **czyPlanowane** – określa czy dane zdarzenie było zaplanowane,
- **zdarzenieId** – identyfikator zdarzenia,
- **technologiaId** – technologia nadawcza, w której pracuje komórka,
- **kontrolerId** – identyfikator kontrolera, który steruje pracą danej komórki,
- **dostawcaId** – identyfikator producenta danej komórki,
- **obszarId** – identyfikator obszaru, na którym pracuje dana komórka.

W bazie danych występuje pięć atrybutów dychotomicznych (*czyProblem*, *czyWoINN*, *czyWoTeren*, *czyWorkflow*, *czyPlanowane*), które ze względu na swoją specyfikę znacząco utrudniają grupowanie (a konkretnie prawidłowe wyliczenie podobieństwa dwóch rekordów między sobą). Ponadto, występuje również pięć atrybutów niezmiennych dla określonej ko-

mórki, takich jak: jej identyfikator, identyfikator kontrolera, identyfikator dostawcy i obszaru, na którym pracuje dana komórka czy używana technologia nadawcza. Już na tym etapie widać, że pierwszym problemem do rozwiązania będzie właściwe wyznaczenie stopnia podobieństwa między dwoma wpisami z zestawu danych. Szczególnie kłopotliwe wydają się być atrybuty jakościowe (wszelkie identyfikatory), dla których w tym przypadku niemożliwe jest określenie jakiejkolwiek hierarchii porządku i podobieństwa. Przykładowo kontrolery o identyfikatorach 101 i 102, mimo że numerycznie różnią się tylko o jedną wartość (co by wskazywało na ich relatywnie wysokie podobieństwo) mogą tyczyć się dwóch, zupełnie niepowiązanych ze sobą urządzeń. Biorąc ten problem pod uwagę, podejściem zastosowanym w tym artykule jest miara stopnia podobieństwa jako, liczba cech mających dokładnie te same wartości.

Tabela 1

Rekord danych poddawanych analizie

NAZWA ATRYBUTU	WARTOŚĆ
cellname	50010A1
startTime	2010-10-14 14:00
data	2010-10-14
strata	0,000263158
strataWzglRegionu	1,64372E-09
strataWzglMikroRegio-	1,30448E-08
czyProblem	0
czyWoINN	0
czyWoTeren	0
czyWorkflow	0
czestosc	1
czasTrwaniaH	1
typPrbId	0
czyPlanowane	0
zdarzenieId	1027271
technologiaId	1
kontrolerId	108
dostawcaId	4
obszarId	20

2. Opis algorytmu grupującego

Kolejnym problemem dotyczącym grupowania przedstawionego we wcześniejszym punkcie zbioru danych jest, co zostało również zaznaczone we wprowadzeniu, wybór właściwego algorytmu analizy skupień. Przy dużych i złożonych wolumenach danych pierwszorzędne znaczenie ma oczywiście złożoność obliczeniowa. Eliminuje to wykorzystanie algorytmów z grupy hierarchicznych, dla których z reguły ten parametr nie jest ich mocną stroną.

Proste algorytmy partycjonujące (k-means) i ich pochodne (fuzzy c-means), mimo znacznie niższej złożoności obliczeniowej, mają wiele innych wad, które dyskwalifikują je w możliwości rozwiązania przedstawionego problemu (jak: duża zależność od warunków początkowych, sztywny podział na z góry określoną liczbę grup, wrażliwość na występowanie wartości izolowanych). Dlatego też ostatecznie zdecydowano się na wykorzystanie gęstościowego algorytmu DBSCAN (ang. *Density-Based Spatial Clustering of Applications with Noise*). Ze względu na to, że został on szczegółowo omówiony przez autorów już we wcześniejszych publikacjach (m.in. w [1, 2, 3]), pozwolono sobie przytoczyć tutaj jedynie ogólną zasadę jego działania.

Algorytm DBSCAN można przedstawić w kilku punktach:

1) dowolny wybór jednego z obiektów p podlegającego grupowaniu,
2) znalezienie wszystkich obiektów „osiągalnych gęstościowo" opierając się na parametrach *Eps* oraz *MinPts*:
   a) jeśli ***p*** jest obiektem „centrum", to formowany jest klaster,
   b) jeśli ***p*** jest obiektem „granicznym" i żaden z punktów nie jest „osiągalny gęstościowo" z obiektu ***p***, wtedy algorytm DBSCAN przechodzi do następnego obiektu w zbiorze danych,
3) proces jest kontynuowany do momentu przeanalizowania wszystkich elementów.

Pierwszym krokiem algorytmu jest wylosowanie obiektu ***p*** oraz wyznaczenie wszystkich obiektów, które są gęstościowo osiągalne z obiektu ***p*** (przy zadanych wartościach *Eps* – maksymalnego promienia sąsiedztwa i *MinPts* – minimalnej liczbie obiektów, wchodzących w skład grupy). Jeżeli ***p*** jest obiektem wewnętrznym, to krok ten skutkuje powstaniem pierwszej grupy. Jeżeli ***p*** jest obiektem krańcowym, to żaden obiekt nie jest „gęstościowo osiągalny" z *p*, więc algorytm wybiera kolejny obiekt ze zbioru danych. Proces ten jest powtarzany, aż nie zostaną przeanalizowane wszystkie obiekty ze zbioru danych wejściowych. Obiekty niezaklasyfikowane do żadnego skupienia są oznaczane jako szum informacyjny [4].

Widać wyraźnie, że obiekty łączone są w grupy głównie na podstawie podobieństwa między sobą, a skupienia są to (intuicyjnie rzecz ujmując) gęsto ułożone obszary obiektów (co odpowiada naturalnie pojmowanej definicji grupowania). Pewnym problemem może być prawidłowe ustawienie parametrów startowych algorytmu – *Eps* i *MinPts*. W niniejszym artykule przyjęto następującą metodę: parametry te zostały eksperymentalnie tak dobrane, by liczba grup stanowiła około 10% całego zbioru danych (jest to wartość akceptowalna, by nie utrudniała dalszej analizy wykrytych zależności) oraz by grupa obiektów izolowanych była jak najmniejsza (ponieważ zbyt duża liczba takich wpisów również mogłaby być stosunkowo trudna w dalszej analizie).

2.1. Koncepcje tworzenia reprezentantów grup

Następną istotną kwestią przy grupowaniu danych złożonych jest (jak to zaznaczono na wstępie) sposób opisu utworzonych grup. Naturalnie każda grupa ma swojego reprezentanta i to zazwyczaj jego opis stanowi informacje o zawartości określonego skupienia. W literaturze przedmiotu ([5], [6]) wyróżnia się między innymi trzy koncepcje tworzenia reprezentantów grup:

- reprezentant jako uśrednienie wartości cech obiektów należących do grupy (centroid),
- reprezentant jako wybrany obiekt ze zbioru danych (medoid),
- reprezentant jako zestaw najczęściej występujących deskryptorów w obrębie grupy.

W przypadku zastosowania centroidu jako reprezentanta grup, nie zawsze możliwe jest jego trywialne określenie, a także może on nie odpowiadać dobrze faktycznej zawartości danego skupienia (chociażby w przypadku, gdyby występował duży rozrzut między wartościami danej cechy obiektów, należących do tej samej grupy).

Jeżeli zdecydowano by wykorzystać koncepcję medoidów do tworzenia reprezentantów, to analityk danych nadal miałby problem (analizując tylko samego reprezentanta) ustalić jakie faktycznie obiekty (o jakich wartościach cech) wchodzą w skład danej grupy.

Reprezentant traktowany jako zestaw najczęściej występujących deskryptorów, z logicznego i intuicyjnego podejścia, dość dobrze reprezentuje konkretnie skupienie, natomiast problem wystąpiłby w przypadku grup mało spójnych (ponieważ tak naprawdę nie wiadomo dokładnie ile obiektów posiada w swoim opisie dany deskryptor).

Biorąc pod uwagę przedstawione aspekty, postanowiono zaproponować dwie inne koncepcje tworzenia reprezentantów, oparte na operatorach sumy i iloczynie logicznym klasycznej logiki (wykorzystywane bardzo często przy formułowaniu pytań do wyszukiwarek internetowych). Wyróżniono zatem dwie koncepcje tworzenia reprezentantów grup:

- reprezentant jako przecięcie deskryptorów opisujących obiekty wchodzące w skład danej grupy; przykład:
  (obszarId, 20)AND(dostawcaId, 2)AND(kontrolerId, 171)AND(technologiaId, 2)
- reprezentant jako zestaw unikalnych deskryptorów wchodzących w skład opisów obiektów danej grupy; przykład:
  [(cellname, 50028B1)OR(cellname, 50028B2)OR(cellname, 50028B5)]AND
  (dostawcaId, 4)

Zaletą pierwszej koncepcji jest fakt, że na pierwszy rzut oka widać dlaczego dane obiekty zostały połączone w jedną grupę oraz jakimi wartościami poszczególnych parametrów wyróżniają się na tle innych skupień. Reprezentant tego typu jest zatem relatywnie łatwy i prosty w analizie.

Druga z przedstawionych koncepcji daje pełniejszy obraz zawartości danej grupy, jednakże może być dużo trudniejsza w interpretacji, jeżeli w ramach danej grupy będzie stosunkowo dużo unikalnych deskryptorów.

3. Przeprowadzone eksperymenty

Celem przeprowadzonych eksperymentów było zbadanie efektywności przeszukiwania złożonych grup obiektów pod kątem różnych pytań (różnych kryteriów), kierowanych do systemu, w zależności od metody tworzenia reprezentantów skupień. Pytanie kierowane do systemu porównywane było wyłącznie z reprezentantami skupień, natomiast stopień relewancji reprezentanta i pytania określony był jako liczba wspólnych deskryptorów. We wszystkich przeprowadzonych eksperymentach liczba utworzonych grup była stała i równa 242, a do grupowania brano pod uwagę wszystkie 19 atrybutów. Ze względu na ograniczony czas wykonania wstępnych eksperymentów ograniczono zestaw danych do 3000 rekordów.

Analiza efektywności przeszukiwania utworzonej struktury złożonych grup została przeprowadzona bazując na trzech parametrach: kompletności i dokładności odpowiedzi oraz zysku czasowego. W klasycznym ujęciu (proponowanym m.in. w pracach [7, 8]) kompletnością nazywa się zdolność systemu do wyszukiwania obiektów relewantnych, a dokładnością zdolność do niewyszukiwania obiektów nierelewantnych. Obiekt uważany jest za relewantny, jeśli w swoim opisie ma co najmniej jeden deskryptor wchodzący w skład pytania. Kompletność odpowiedzi zatem rozumiana jest jako stosunek liczby relewantnych, wyszukanych obiektów, do wszystkich relewantnych do zadanego pytania obiektów. Dokładność odpowiedzi natomiast jest to stosunek liczby wyszukanych, relewantnych obiektów do wszystkich wyszukanych obiektów. Komentarza wymaga również pojęcie zysku czasowego, które zostało określone jako krotność przyspieszenia generowania przez system odpowiedzi w stosunku do zastosowania metody przeglądu zupełnego wszystkich obiektów w bazie.

Wszystkie eksperymenty zostały przeprowadzone wykorzystując platformę bazodanową Microsoft SQL 2008 Enterprise Server, na podstawie której zaimplementowano wybrany algorytm grupowania.

3.1. Wyniki eksperymentów

Pierwszy przeprowadzony eksperyment polegał na zadaniu do systemu opartego na algorytmie DBSCAN następującego pytania: *Znajdź komórki sterowane przez kontroler 106, dla których nie były zaplanowane żadne prace, a które były niedostępne przez godzinę w ciągu całego dnia*. W tym przypadku reprezentant tworzony był jako przecięcie wszystkich de-

skryptorów (opisów obiektów wchodzących do grupy). Podsumowanie wyników zwróconych w odpowiedzi prezentuje tabela 1.

Przy bardzo niewielkiej liczbie obiektów relewantnych (w tym przypadku równiej cztery) osiągnięto pełną dokładność oraz dość wysoki poziom kompletności. Odpowiedź na zadane pytanie osiągnięto ponad 12-krotnie szybciej, aniżeli przy zastosowaniu metody przeglądu zupełnego.

Drugi przeprowadzony eksperyment polegał na zadaniu do systemu tego samego pytania, jednakże tym razem reprezentanci grup tworzeni byli jako zestaw unikalnych deskryptorów (wchodzących w skład opisu obiektów danej grupy). Wyniki zostały zaprezentowane w tabeli 2.

Mimo zmiany metody tworzenia reprezentantów, wyniki przeprowadzonego eksperymentu (w porównaniu z poprzednim) są identyczne. Nie bez znaczenia jest jednak fakt, że ogólna liczba obiektów wchodzących w skład zwróconej w odpowiedzi grupy jest bardzo niska – istnieje duże prawdopodobieństwo, że wszystkie obiekty zawarte w tej grupie mają niemalże identyczne opisy, przez co zmiana sposobu tworzenia reprezentanta na bardziej restrykcyjną (reprezentant tworzony z wykorzystaniem spójnika AND w ogólnym przypadku jest dużo krótszy niż ten stworzony używając spójnika OR) nie wpłynęło negatywnie na wyniki wyszukiwania.

Tabela 2

Wyniki dla I przypadku testowego

	DBSCAN
Liczba obiektów relewantnych	4
Liczba obiektów znalezionych przez system	3
Kompletność	0,750000
Dokładność	1,000000
Zysk czasowy w stosunku do MPZ	12,40

Tabela 3

Wyniki dla II przypadku testowego

	DBSCAN
Liczba obiektów relewantnych	4
Liczba obiektów znalezionych przez system	3
Kompletność	0,750000
Dokładność	1,000000
Zysk czasowy w stosunku do MPZ	12,40

Kolejny przeprowadzony eksperyment polegał na zmianie pytania do systemu i obserwacji jego zachowania. Tym razem pytanie brzmiało: *Znajdź komórki sterowane przez kontroler 107, które były niedostępne 2 godziny w ciągu całego dnia*. Reprezentant znowuż tworzony był jako przecięcie wszystkich deskryptorów. Tym razem jednak wyniki są już znacznie bardziej interesujące, co ilustruje tabela 3.

Mimo iż nadal tylko mały procent ogółem obiektów w bazie był relewantny do zadanego pytania oraz system w odpowiedzi zwrócił grupę liczącą niewiele obiektów (bo tylko cztery), to żaden obiekt zwrócony przez system nie stanowił prawidłowej odpowiedzi na zadane pytanie (stąd zerowe kompletność i dokładność). Zysk czasowy w stosunku do zastosowania metody przeglądu zupełnego jest ujemny, ponieważ czas przeznaczony na przeszukiwanie struktury grup jest czasem straconym – system nie zwrócił pożądanej odpowiedzi i należałoby i tak wykonać przegląd zupełny zestawu danych.

Odmienną sytuacje prezentują wyniki zamieszczone w tabeli 4. Pytanie kierowane do systemu pozostało niezmienione, natomiast zmianie uległ sposób tworzenia reprezentantów – jako zestaw unikalnych deskryptorów. W tym przypadku system zwrócił więcej obiektów niż oczekiwano (bo aż 28 podczas, gdy relewantnych do pytania ogółem było dziesięć), stąd niska wartość parametru dokładności. Kompletność na poziomie 0,6 jest również wartością niższą niż oczekiwano, aczkolwiek jest to (mimo wszystko) sytuacja dużo lepsza, niż zademonstrowana w poprzednim przypadku testowym.

Tabela 4

Wyniki dla III przypadku testowego

	DBSCAN
Liczba obiektów relewantnych	10
Liczba obiektów znalezionych przez system	4
Kompletność	0
Dokładność	0
Zysk czasowy w stosunku do MPZ	-0,81

Tabela 5

Wyniki dla IV przypadku testowego

	DBSCAN
Liczba obiektów relewantnych	10
Liczba obiektów znalezionych przez system	28
Kompletność	0,600000
Dokładność	0,214286
Zysk czasowy w stosunku do MPZ	12,40

4. Podsumowanie

Celem niniejszego artykułu było krótkie przedstawienie problemów narastających przy zagadnieniu grupowania danych złożonych, na rzeczywistym zbiorze danych odnośnie telefonii komórkowej. Szczególnym przedmiotem analizy był aspekt tworzenia reprezentantów skupień oraz jego wpływ zarówno na właściwy opis elementów, jak i na proces wyszukiwania obiektów relewantnych do zadanego pytania. Przeprowadzone eksperymenty miały na celu zbadać poziom efektywności odpowiedzi systemu, mierzonej standardowymi miarami

kompletności oraz dokładności. Wyniki jednoznacznie wskazują na to, że sposób tworzenia reprezentantów ma spore znaczenie dla wyszukania bądź nie obiektów relewantnych.

Grupowanie danych złożonych rodzi wiele problemów implementacyjnych, wśród których należałoby wyróżnić: brak prostej struktury do przechowywania danych różnego typu, duże złożoności obliczeniową i pamięciową, związane z wykonywaniem operacji na tych danych, problematyczne wyliczanie podobieństwa, gdy występują zarówno dane ilościowe, jak i jakościowe.

BIBLIOGRAFIA

1. Xięski T.: Zastosowanie algorytmu DBSCAN dla grupowania danych tekstowych, [w:] Wakulicz-Deja A. (red.): Systemy wspomagania decyzji. Instytut Informatyki Uniwersytetu Śląskiego, Sosnowiec 2010.
2. Nowak-Brzezińska A., Jach T., Xięski T.: Finding a relevant document in the clusters of documents' characteristics. Intelligent Information Systems, 2010, s. 273÷283.
3. Nowak-Brzezińska A., Jach T., Xięski T.: Wybór algorytmu grupowania a efektywność wyszukiwania dokumentów. Studia Informatica, Wyd. Pol. Śląskiej, Vol. 31, No. 2A (89), 2010, s. 147÷162.
4. Ester M., Ester K., Sander H.-P., Sander J., Xu X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceedings of 2nd Conference on Knowledge Discovery and Data Mining, USA 1996.
5. Nowak A., Wakulicz-Deja A., Bachliński S.: Optimization of Speech Recognition by Clustering of Phones. Fundamenta Informaticae, Holandia 2006, s. 283÷293.
6. Tan P.-N., Steinbach M., Vipin K.: Introduction to Data Mining. Addison-Wesley, USA 2006.
7. Rijsbergen C. J.: Information retrieval. Butterworth-Heinemann, UK 1979.
8. Wakulicz-Deja A.: Podstawy systemów wyszukiwania informacji. Analiza metod. Akademicka Oficyna Wydawnicza PLJ, Warszawa 1995.
9. Hustecki J.: Vademecum Teleinformatyka, cz. I, praca zbiorowa. IDG Poland, Warszawa 2004.

Recenzenci: Dr hab. inż. Marcin Gorawski, prof. Pol. Wrocławskiej
Dr hab. inż. Adam Pelikant, prof. Pol. Łódzkiej

Wpłynęło do Redakcji 15 stycznia 2011 r.

Abstract

In this paper the topic of clustering complex data (using the well known density based DBSCAN algorithm) and searching though such a structure is discussed. Authors focus on comparing the efficiency of the search process based on two presented methods of creating the cluster representatives and various types of questions. The data set used in the experiments includes real life information about the functioning of transceivers of a cellular phone operator located in various parts of the Katowice region. The results of performed experiments show that domain knowledge and ways of creating cluster representatives have a huge impact on discovering inner data relationships as well as the search results. In conclusion authors also note that clustering complex data brings up many implementation problems, which may not be so obvious from the beginning.

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.