Streszczenie rozprawy doktorskiej pt. *Wydobywanie wiedzy z danych złożonych*.
mgr Tomasz Xięski

Nieustanny rozwój techniki oraz rosnące możliwości sprzętu komputerowego umożliwiają przechowywanie bardzo dużych ilości danych we wszelkiego rodzaju bazach i repozytoriach. Dane te są gromadzone, ponieważ zakłada się, że mogą być źródłem nieznanych, potencjalnie użytecznych wzorców, korelacji i trendów. Odkryte wzorce, wyrażone w postaci modelu analitycznego, mogą posiadać skomplikowaną strukturę, przez co są trudne do dalszej analizy. Jednakże to nie tylko nadmierna ilość danych wpływa na trudności badawcze. Istotniejszym czynnikiem jest ich złożona struktura, zarówno pod względem dużej liczby atrybutów opisujących każdy obiekt danych, jak również użytych typów danych. Informacje zakodowane w bazie często opisane są atrybutami różnych typów, wliczając w to wartości binarne, dyskretne, ciągłe, kategoryczne, tekstowe czy reprezentujące daty. Tego typu dane można nazwać złożonymi i będą one podstawą analizy w niniejszej rozprawie.

Celem rozprawy jest zatem opracowanie metody wydobywania wiedzy ze złożonych zbiorów danych rzeczywistych o dużej liczebności, uwzględniającej ich specyfikę i dziedzinowy charakter oraz efektywne środki wizualizacji wydobytej wiedzy. Badania zostaną oparte na dwóch rzeczywistych zbiorach: pierwszy z nich zawiera dane dotyczące funkcjonowania urządzeń nadawczo-odbiorczych operatora telefonii komórkowej (rozmieszczonych na terenie aglomeracji śląskiej), drugi agreguje statystyki gromadzone w oprogramowaniu do zarządzania sieciami bezprzewodowymi. Mimo, że oba zestawy danych wydają się być ze sobą mocno powiązane pod względem tematyki telekomunikacyjnej, to jednak posiadają zupełnie odmienną strukturę i charakterystykę.

Spośród wielu technik eksploracji danych zdecydowano się wybrać analizę skupień i to właśnie wszystkie aspekty realizacji tej techniki w odniesieniu do danych złożonych są podstawą niniejszej rozprawy. Wydobywanie wiedzy z rzeczywistych baz wiedzy jest procesem wieloetapowym i stawia szereg wymogów wobec algorytmów grupowania jak: możliwość odkrywania skupień o różnej strukturze, odporność na występowanie wartości izolowanych, posiadanie relatywnie niskiej złożoności obliczeniowej i zajętości pamięci, jasno określone kryteria stopu algorytmu oraz wysoka jakość tworzonych skupień. Niestety klasyczne metody analizy skupień (jak niehierarchiczny algorytm k-średnich) nie spełniają podanych wymagań. Dodatkowo takie rzeczywiste bazy danych najczęściej charakteryzują się występowaniem wartości pustych (brakujących), czy zduplikowanych, co znacząco utrudnia ich przetwarzanie oraz analizę. Zatem w procesie badawczym wykorzystywane są bardziej złożone algorytmy, ale te dla osiągnięcia optymalnego rezultatu wymagają zdefiniowania różnej liczby parametrów. Dlatego też istotnym problemem, omawianym w pracy, jest określenie metod: optymalnego doboru parametrów dla procesu grupowania oraz opisu utworzonej struktury złożonych grup.

Niniejsza praca odnosi się również do problemu, w jaki sposób wizualizacja danych może funkcjonować jako efektywne i autonomiczne narzędzie analizy, jak również służyć jako technika łącząca wiedzę dziedzinową i zdolności kognitywne człowieka w procesie odkrywania wiedzy. Omawia proces graficznej analizy eksploracyjnej (ang. *visual data mining*) oraz dokonuje porównania najpopularniejszych technik reprezentacji skupień, spotykanych w literaturze przedmiotu, z autorską koncepcją opartą na algorytmie generowania tzw. map prostokątów (ang. *squarified treemaps*). Kolejnym istotnym aspektem omawianym w rozprawie jest przegląd i porównanie możliwości obecnie dostępnych systemów analizy danych, wraz z wykazaniem ich wad i zalet, szczególnie w kontekście efektywności zaimplementowanych technik grupowania. Stanowi to jednocześnie motywację do stworzenia autorskiego systemu wydobywania wiedzy DensGroup.