Eksploracja danych
- Odkrywanie wiedzy w danych

Marek Wojciechowski

Instytut Informatyki
Politechnika Poznańska

Wielkie bazy danych

- Wielkie bazy danych (Very Large Databases) i hurtownie danych (Data Warehouses)
- Rozmiary współczesnych baz danych
  - sieć sprzedaży Wal-Mart gromadzi dziennie dane dotyczące ponad 20 milionów transakcji
  - koncern Mobil Oil rozwija magazyn danych pozwalający na przechowywanie ponad 100 terabajtów danych o wydobyciu ropy naftowej
  - system satelitarnej obserwacji EOS zbudowany przez NASA generuje w każdej godzinie dziesiątki gigabajtów danych
  - niewielkie supermarkety rejestrują codziennie sprzedaż tysięcy artykułów
- Wielkie wolumeny danych są trudne w analizowaniu
- Informacje o dotychczasowej działalności przedsiębiorstwa, poziomie i strukturze sprzedaży oraz cechach klientów mogą posłużyć do wspomagania podejmowania decyzji

Zależności w bazach danych – Przykład 1

wiek kierowcy	lat prawo jazdy	kolor pojazdu	poj. silnika	moc	razem szkody
42	24	biały	1610	100	0
19	1	czerwony	650	24	2500
28	4	czerwony	1100	40	0
41	20	czarny	1800	130	0
21	3	czerwony	650	24	1300
20	1	niebieski	650	24	0

- Kierowcy, którzy jeżdżą czerwonymi samochodami o pojemności 650 ccm, powodują wypadki drogowe
- Kierowcy w wieku powyżej 40 lat jeżdżą samochodami o pojemności większej niż 1600 ccm
- Kierowcy, którzy posiadają prawo jazdy dłużej niż 3 lata, nie powodują wypadków
- Kierowcy w wieku poniżej 30 lat jeżdżą samochodami koloru czerwonego

Zależności w bazach danych – Przykład 2

transakcja	produkt	dzień	cena
1	pizza	sobota	48,40
1	mleko	sobota	2,80
1	chleb	sobota	1,50
2	piwo	wtorek	16,20
2	orzeszki	wtorek	8,50
3	chleb	sobota	1,50
3	orzeszki	sobota	25,50
3	piwo	sobota	32,40

- piwo i orzeszki są zawsze kupowane wspólnie
- chleb uczestniczy w transakcjach na kwotę większą niż 50 złotych

Data Mining - Eksploracja danych

- Eksploracja danych: zbiór technik automatycznego odkrywania nietrywialnych zależności i schematów (patterns) w dużych zbiorach danych (bazach i hurtowniach danych)
- Eksploracja danych często nazywana jest również odkrywaniem wiedzy w bazach danych (Knowledge Discovery in Databases) lub eksploracją baz danych (Database Mining)
- Eksploracja danych leży na przecięciu trzech dziedzin naukowych: baz danych, uczenia maszynowego i statystyki

Dziedziny zastosowań eksploracji danych

- Handel i marketing
  - identyfikacja „profilu klienta" na potrzeby marketingu kierunkowego
  - wykrywanie schematów zakupów i planowanie lokalizacji artykułów
- Finanse i bankowość
  - schematy wykorzystywania kradzionych kart kredytowych
  - przewidywanie dochodowości portfela akcji, znajdowanie korelacji wśród wskaźników finansowych
- Nauka i technologia
  - analiza strumieni wyników pomiarów
  - wykrywanie alarmów w sieciach telekomunikacyjnych
- Internet (Web Mining)
  - handel i marketing internetowy
  - analiza zachowań użytkowników WWW
  - personalizacja serwisów WWW

Metody eksploracji danych

- Odkrywanie asocjacji (zbiorów częstych i reguł)
- Odkrywanie wzorców sekwencyjnych
- Klasyfikacja
- Odkrywanie charakterystyk
- Analiza skupień (klastrowanie)
- Dyskryminacja
- Regresja
- Wykrywanie zmian i odchyleń

Metody eksploracji: odkrywanie asocjacji

- Odkrywanie asocjacji: znajdowanie związków pomiędzy występowaniem grup elementów w zbiorach danych
- Przykłady asocjacji:
  - klienci, którzy kupują piwo, kupują również orzeszki
  - klienci, którzy kupują chleb, masło i ser, kupują również wodę mineralną i ketchup
- Zastosowania odkrytych asocjacji:
  - planowanie kampanii promocyjnych
  - planowanie rozmieszczenia stoisk sprzedaży w supermarketach

Metody eksploracji: odkrywanie wzorców sekwencyjnych

- Odkrywanie wzorców sekwencyjnych: znajdowanie najczęściej występujących sekwencji elementów
- Przykłady wzorców sekwencyjnych:
  - 10% klientów, kupiło wędkę, a następnie kalosze
  - 5% użytkowników serwisu WWW odwiedziło w ciągu jednej sesji najpierw stronę wakacje.html, później promocje.html, a następnie dojazd_wlasny.html
- Zastosowania odkrytych sekwencji:
  - przewidywanie sprzedaży
  - marketing kierunkowy
  - wykrywane symptomów wskazujących na możliwość awarii
  - analiza zachowań użytkowników WWW

Metody eksploracji: klasyfikacja

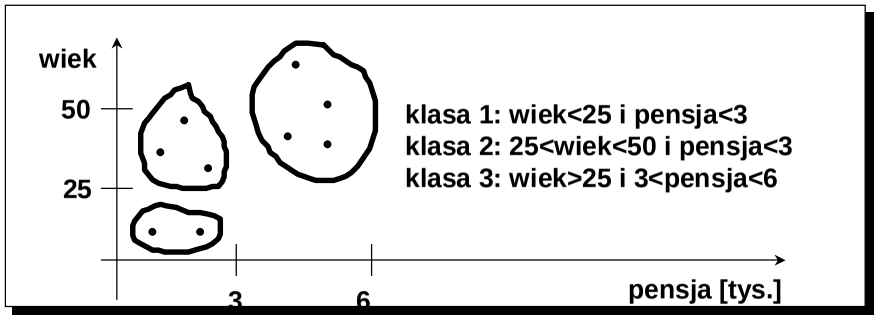
- Klasyfikacja: znajdowanie sposobu odwzorowywania danych w zbiór predefiniowanych klas (podzbiorów)
- Przykład klasyfikacji: automatyczny podział kierowców na powodujących i niepowodujących wypadków drogowych:
  - kierowcy prowadzący czerwone pojazdy o pojemności 650 ccm powodują wypadki drogowe
  - kierowcy, którzy posiadają prawo jazdy ponad 3 lata lub jeżdżą niebieskimi samochodami nie powodują wypadków drogowych
- Zastosowania klasyfikacji:
  - diagnostyka medyczna
  - rozpoznawanie trendów na rynkach finansowych
  - automatyczne rozpoznawanie obrazów
  - przydział kredytów bankowych

Metody eksploracji: odkrywanie charakterystyk

- Odkrywanie charakterystyk: znajdowanie zwięzłych opisów (charakterystyk) podanego zbioru danych
- Przykład odkrywania charakterystyk: opis pacjentów chorujących na anginę
  - pacjenci chorujący na anginę cechują się temperaturą ciała wyższą niż 37.5 C, bólem gardła, osłabieniem organizmu
- Zastosowania odkrywania charakterystyk:
  - znajdowanie zależności funkcyjnych pomiędzy zmiennymi
  - określanie profilu klienta - zbioru cech charakterystycznych

Metody eksploracji: analiza skupień

- Analiza skupień (klastrowanie): znajdowanie skończonego zbioru klas (podzbiorów) w bazie danych



- Zastosowania analizy skupień:
  - określanie segmentów rynku na podstawie cech klientów
  - odkrywanie grup podobnie zachowujących się użytkowników WWW na potrzeby personalizacji

Formy reprezentacji odkrytych schematów

- Znane w dziedzinach uczenia maszynowego i sztucznej inteligencji:
  - sieci neuronowe
  - drzewa decyzyjne
  - listy decyzyjne
  - sieci semantyczne
  - proste i złożone reguły logiczne
- Założenie: wiedza powinna być reprezentowana w prostej i czytelnej dla człowieka postaci
- Eksploracja danych najczęściej wykorzystuje:
  - wzorce częste (zbiory, sekwencje)
  - reguły logiczne
  - drzewa decyzyjne

Reguły logiczne (1/2)

- Przykład prostej reguły logicznej:

```
kolor_poj=czerwony AND pojemnosc=650 -> szkoda=TAK
```

- Definicja reguły logicznej:

```
r1(a1, v1)  AND r2(a2, v2) ...  rj(aj, vj) ->
-> rk(ak, vk) AND rl(al, vl)  ...  rn(an, vn)
```

  - ai jest atrybutem,
  - vi jest wartością prostą (np. liczba, ciąg znaków) lub złożoną (np. zbiór),
  - ri jest predykatem (np. równość, zawieranie)
- Lewa strona reguły nazywa się ciałem reguły (body), prawa strona nazywa się głową reguły (head)
- Reguła może być potwierdzana lub naruszana przez wybraną krotkę relacji

Reguły logiczne (2/2)

- Każda reguła posiada wskaźniki statystycznej ważności i siły: wsparcie (support) i zaufanie (confidence)
- Wsparcie reguły odpowiada liczbie krotek potwierdzających daną regułę
- Zaufanie reguły odpowiada jej wiarygodności, tj. poprawności reguły w zbiorze krotek

Temperatura	Ból_głowy	Ból_gardła	Diagnoza
wysoka	tak	nie	zatrucie
wysoka	tak	nie	zdrowy
wysoka	tak	tak	angina
wysoka	nie	tak	angina

Ból_gardła = „tak” → Diagnoza = „angina” (S=50% C=100%)

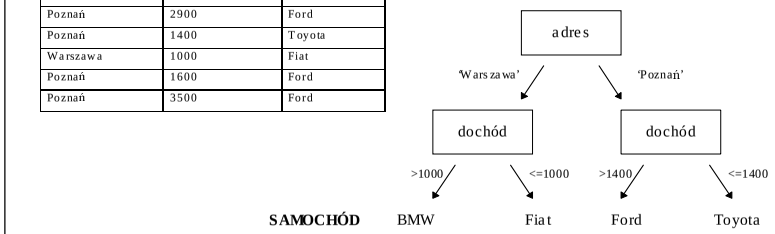
Temperatura=”wysoka” ∧ Ból_głowy=”tak” ∧ Ból_gardła=”nie” → Diagnoza=”zatrucie” (S=25% C=100%)

Temperatura=”wysoka” ∧ Ból_głowy=”tak” ∧ Ból_gardła=”nie” → Diagnoza=”zdrowy” (S=25% C=100%)

Drzewa decyzyjne

- Drzewo decyzyjne jest formą opisu wiedzy klasyfikującej
- Węzłom drzewa odpowiadają atrybuty eksplorowanej relacji
- Krawędzie opisują wartości atrybutów
- Liśćmi drzewa są wartości atrybutu klasyfikującego

Adres	Dochód	Samochód
Warszawa	4000	BMW
Poznań	2900	Ford
Poznań	1400	Toyota
Warszawa	1000	Fiat
Poznań	1600	Ford
Poznań	3500	Ford



Algorytmy eksploracji danych – Przykład (1/2)

- Problem odkrywania zbiorów częstych (ang. frequent itemsets)

database D

{a,b,d,f}
{a,d,e,f,g}
{a,b,c,d,e,f}
{f,h,i,j}

minumum support = 3

frequent itemsets L

{a}	(3)
{d}	(3)
{f}	(4)
{a,d}	(3)
{a,f}	(3)
{d,f}	(3)
{a,d,f}	(3)

itemset = a set of items
D = a set of itemsets
support(T) = number of itemsets in D that contain T
frequent itemset = an itemset with support above user defined minumum

Algorytmy eksploracji danych – Przykład (2/2)

- Algorytm Apriori (Agrawal, Srikant, 1994)

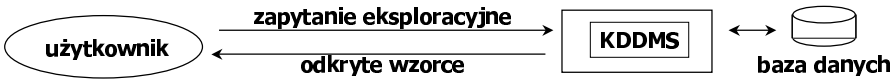
{a,b,d,f}
{a,d,e,f,g}
{a,b,c,d,e,f}
{f,h,i,j}

C_1 = {all 1-itemsets from D}
for (k=1; $C_k \neq \emptyset$; k++)
count(C_k, D);
$L_k = \{c \in C_k \mid c.count \geq minsup\}$;
C_{k+1} = generate_candidates(L_k);
Answer = $\cup_k L_k$;

minsup=3
C_1={{a},{b},{c},{d},{e},{f},{g},{h},{i},{j}}
C_1={{a}(3),{b}(2),{c}(1),{d}(3),{e}(2),{f}(4),{g}(1),{h}(1),{i}(1),{j}(1)}
L_1={{a},{d},{f}}
C_2={{ad},{af},{df}}
C_2={{ad}(3),{af}(3),{df}(3)}
L_2={{ad},{af},{df}}
C_3={{adf}}
C_3={{adf}(3)}
L_3={{adf}}
Answer={{a},{d},{f},{ad},{af},{df},{adf}}

Eksploracja danych jako zaawansowane zapytania do bazy danych (1/2)

- Podejście szczególnie uzasadnione dla asocjacji i wzorców sekwencyjnych
- Użytkownik specyfikuje:
  - klasę szukanych wzorców
  - zbiór danych wejściowych
  - kryteria selekcji (ograniczenia) dla wzorców
- System eksploracji danych (KDDMS):
  - dobiera odpowiedni algorytm
  - zwraca odkryte wzorce jako wynik zapytania
  - kryteria selekcji (ograniczenia) dla wzorców
- Eksploracja danych ma charakter interaktywny i iteracyjny



Eksploracja danych jako zaawansowane zapytania do bazy danych (2/2)

- Wiele prototypowych rozszerzeń SQL zaproponowanych w literaturze
- *MineSQL* (Politechnika Poznańska):

mysets

i	s
1	{a,b,d,e,f}
2	{a,c,d,h}
...	...

```
mine itemset
from (select s
      from   mysets
      where  i<=100)
where support(itemset) > 10
```

- Rozszerzenie standardu SQL o funkcje eksploracji danych mało prawdopodobne
- Eksploracji danych poświęcono oddzielne standardy
  - część z nich "współpracuje" z językiem SQL

Eksploracja danych – dotychczasowe kierunki badań

- Najpopularniejsze dotychczasowe kierunki badań:
  - coraz efektywniejsze algorytmy eksploracji danych
  - skalowalne algorytmy eksploracji danych
  - języki zapytań eksploracyjnych
  - przetwarzanie zapytań eksploracyjnych
  - algorytmy odkrywania wzorców częstych i reguł z ograniczeniami
  - inkrementalne algorytmy eksploracji danych
  - narzędzia graficzne dla eksploracji danych (wizualizacja)
  - integracja eksploracji danych z systemami zarządzania bazami danych
  - eksploracja rozproszonych baz danych
  - architektury równoległe w eksploracji danych

Eksploracja danych – nowe trendy

- Eksploracja strumieni danych
  - jedno „spojrzenie" na dane
- Eksploracja danych w biologii molekularnej
  - analiza sekwencji DNA, protein
- Eksploracja danych semi-strukturalnych
  - kolekcje dokumentów XML
- Eksploracja danych multimedialnych
  - np. wykrywanie podobieństw, plagiatów
- Kwestie prywatności w eksploracji danych

Standardy dla eksploracji danych

- SQL/MM Part 6
  - specyfikacja standardowej biblioteki typów obiektowych SQL
- Java Data Mining API
  - interfejs do eksploracji danych z poziomu języka Java
- PMML
  - język na bazie XML do opisu zadań (procesów) eksploracji danych
  - Umożliwia współdzielenie (wymianę) modeli między aplikacjami
- Microsoft OLE DB for Data Mining
  - protokół umożliwiający wykorzystywanie funkcji eksploracji danych z poziomu SQL
  - uwzględnia PMML

Oprogramowanie komercyjne dla eksploracji danych

- IBM Intelligent Miner, współpracuje z DB2, Oracle, Sybase, przeznaczony na platformy AIX, AS, OS
- Oracle9i Database Server with Data Mining Option

- Integral Solutions Clementine, współpracuje z Oracle, Sybase, Informix, Ingres,
- SAS Enterprise Miner
- ...

Data Mining - success stories

- Database Marketing w American Express
  - analiza danych o klientach w celu znajdowania schematów ich preferencji
  - wykorzystanie schematów dla precyzyjnej selekcji kolejnych klientów
  - Efekt: ok. 10% wzrost zakupów z wykorzystaniem kart kredytowych
- Weryfikacja poprawności danych w Reuters
  - wykrywanie prawdopodobnych przekłamań w wysokości publikowanych kursów wymiany walut
- Profil słuchacza w BBC
  - odkrywanie profili widowni programów telewizyjnych w celu wyboru optymalnych pór ich nadawania
- Skład zespołu w Orlando Magic
  - odkrywanie optymalnego składu i ustawienia zespołu
  - rezultat: likwidacja trendu spadkowego

Przyszłość eksploracji danych

- Eksploracja danych z nowej dziedziny naukowej staje się dziedziną dojrzałą
- Przyszłość dziedziny zależy od jej upowszechnienia się i praktycznej przydatności
- Problemy upowszechniania się eksploracji danych:
  - ciągle wysoki koszt narzędzi eksploracji danych
  - złożoność problemów eksploracji danych
  - wiele instytucji dopiero wdraża hurtownie danych i „jest na etapie" analiz OLAP
- Sygnały pozytywne:
  - powstawanie standardów regulujących sposoby wykorzystywania eksploracji danych
  - dostęp wielu narzędzi komercyjnych, w tym przede wszystkim tych związanych z serwerami baz danych
  - pozytywne doświadczenia wielu przedsiębiorstw i instytucji