# Metody Inżynierii Wiedzy
# Eksploracja danych, reguły asocjacyjne - wykład 14

Adam Szmigielski
aszmigie@pjwstk.edu.pl
materiały: $ftp(public) : //aszmigie/MIW$

## Źródła i rodzaje danych

- **Dane eksperymentalne**: fizyka, astronomia, biologia,
- **Dane komercyjne**: banki, ubezpieczalnie, firmy, sieci handlowe, firmy przewozowe,
- **Web**: tekst, e-handel portale społecznościowe etc.

## Przykłady danych

- eksperymentalne - Very Long Baseline Interferometry (VLBI) posiada 16 teleskopów, z których każdy produkuje 1 Gigabit/second danych astronomicznych w czasie 25-dniowej sesji obserwacyjnej.
- komercyjne - Firma telekomunikacyjna AT&T obsługuje miliardy połączeń dziennie.
- web - miliardów stron internetowych, rozwojem e-handlu i rozprzestrzenianiem się olbrzymich ilości informacji w postaci tekstowej. Alexa internet archiwum: 7-letnie dane, 500 TB, Google - 8 miliardów stron, Yahoo - 20 miliardów stron , IBM WebFountain, 160 TB (2003), Internet archiwum (www.archive.org), 300 TB;

# Rozmiar informacji tworzonej na świecie

1TB (tera) $10^{12} = 1.000.000.000.000$

1PB (peta) $10^{15} = 1.000.000.000.000.000$

**1EB (egza)** $10^{18} = 1.000.000.000.000.000.000$

1ZB (zetta)$10^{21} = 1.000.000.000.000.000.000.000$

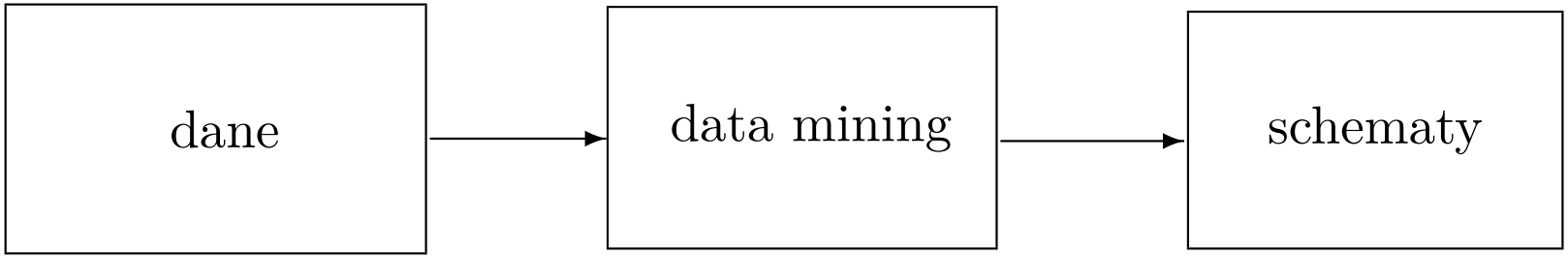
1YB (yotta)$10^{24} = 1.000.000.000.000.000.000.000.000$

- W 1999, suma informacji stworzonych przez człowieka (w tym wszystkich nagrań dźwiękowych, wideo, tekstów i książek) wynosiła około 12EB exabytów danych.
- W 2002 rozmowy telefoniczne na całym świecie zarówno telefonii stacjonarnej, jak i komórkowej zawierały 17,3 EB
- W 2006 60 EB były tworzone, przechwycone na całym świecie,
- W 2007 wynosił 295 EB exabytów

## Eksploracja danych

- Proces automatycznego odkrywania nietrywialnych, dotychczas nieznanych, potencjalnie użytecznych reguł, zależności, wzorców schematów, podobieństw lub trendów w dużych repozytoriach danych (bazach danych, hurtowniach danych, itp.)
- Odkrywanie wiedzy w bazach danych KDD (Knowledge Discovery in Databases),
- Ekstrakcja wiedzy, inteligencja biznesowa, pozyskiwanie wiedzy.

## Eksploracja danych




- Eksploracja danych (ang. Data Mining): zbiór technik automatycznego odkrywania nietrywialnych zależności, schematów, wzorców, reguł (ang.patterns) w dużych zbiorach danych (bazach danych, hurtowniach danych).

## Cel eksploracji danych

- Odkrywane w procesie eksploracji danych wzorce mają, najczęściej, postać reguł logicznych, klasyfikatorów (np. drzew decyzyjnych), zbiorów skupień, wykresów, itp.
- Celem eksploracji najogólniej mówiąc jest analiza danych i procesów w celu lepszego ich poznania i zrozumienia.
- Automatyczna eksploracja danych otwiera nowe możliwości formułowanie zapytań na znacznie wyższym poziomie abstrakcji aniżeli pozwala na to standard SQL.

## Zapytania do bazy danych

- Zapytania operacyjne - użytkownik zapytuje o dane w bazie danych (np. SQL),
- Zapytania analityczne oparte o model - użytkownik zapytuje i dokonuje analizy w oparciu o model,
- Zapytania eksploracyjne - mają charakter znacznie bardziej ogólny i znacznie bardziej abstrakcyjny.

## Zapytania eksploracyjne - przykład

Dany jest zbiór danych opisujących pacjentów szpitala. Czy potrafimy w oparciu o ten zbiór danych:

- Poprawnie zdiagnozować pacjenta (określić chorobę)?
- Przewidzieć poprawnie wynik terapii?
- Zaproponować najlepszą terapię?

## Proces odkrywania wiedzy

- Czyszczenie danych,
- Konsolidacja i transformacja danych,
- Wybór metody (metod) eksploracji danych,
- Wybór algorytmów eksploracji danych
- Eksploracja danych
- Interpretacja, analiza i ocena wyników wizualizacja,
- Transformacja, usuwanie redundantnych wzorców, etc.
- Wykorzystanie pozyskanej wiedzy

## Interdyscyplinarność procesu odkrywania wiedzy

- Systemy baz danych, hurtownie danych,
- Statystyka,
- Uczenie maszynowe i odkrywanie wiedzy,
- Techniki wizualizacji danych,
- Teoria informacji,
- Wyszukiwanie informacji,
- Inne dyscypliny – sieci neuronowe, modelowanie matematyczne, rozpoznawanie obrazów, etc.

## Metody eksploracji danych

- klasyfikacja (regresja)
- grupowanie
- odkrywanie sekwencji
- odkrywanie charakterystyk
- analiza przebiegów czasowych
- odkrywanie asocjacji
- wykrywanie zmian i odchyleń
- eksploracja WWW
- eksploracja tekstów

## Klasyfikacja

Metoda analizy danych, której celem jest predykcja wartości określonego atrybutu w oparciu o pewien zbiór danych treningowych

- Klasyfikacja jest metodą analizy danych, której celem jest predykcja wartości określonego atrybutu w oparciu o pewien zbiór danych treningowych.
- Wiele technik:
  - statystyka,
  - drzewa decyzyjne,
  - sieci neuronowe,

## Grupowanie

Metoda pogrupowania obiektów w oparciu o ich wartości.

- Metody te grupują obiekty w klasy w taki sposób, aby maksymalizować podobieństwo obiektów wewnątrz klas i minimalizować podobieństwo pomiędzy klasami obiektów.
- zastosowania grupowania:
  - grupowanie dokumentów,
  - grupowanie klientów,
  - segmentacja rynku

## Odkrywanie asocjacji

Znajdowanie związków pomiędzy występowaniem grup elementów w zbiorach danych

- Celem procesu odkrywania asocjacji jest znalezienie interesujących zależności lub korelacji, nazywanych ogólnie asocjacjami, pomiędzy danymi w dużych zbiorach danych,
- Wynikiem procesu odkrywania asocjacji jest zbiór reguł asocjacyjnych opisujących znalezione zależności lub korelacje między danymi,
- Celem tej analizy jest znalezienie naturalnych wzorców zachowań konsumenckich klientów poprzez analizę produktów.

## Odkrywanie wzorców sekwencji

- Na podstawie danych opisujących zakupy danego klienta, uporządkowanych zgodnie z wartościami etykiet czasowych można uzyskać profil klienta i próbować przewidzieć jego zachowanie w czasie.
- Zastosowania odkrytych wzorców sekwencji:
  - Planowanie inwestycji giełdowych,
  - przewidywanie sprzedaży,
  - Znajdowanie skutecznej terapii.

## Odkrywanie charakterystyk

Metoda ta polega na znajdowaniu zwięzłych opisów (charakterystyk) podanego zbioru danych, czy też znajdowaniu zależności funkcyjnych pomiędzy zmiennymi opisującymi zbiór danych.

- Znajdowanie zwięzłych opisów (charakterystyk) podanego zbioru danych,
- Przykład odkrywania charakterystyk: Pacjenci chorujący na anginę cechują się
  - Temperaturą ciała większą niż 37,5 C,
  - bólem gardła,
  - osłabieniem organizmu

## Problemy odkrywania wiedzy

- W dużych bazach danych mogą zostać odkryte tysiące reguł,
- Człowiek nie potrafi rozumieć i przeanalizować bardzo dużych zbiorów informacji,
- Różni użytkownicy systemu bazy danych są zainteresowani różnymi typami reguł z różnych relacji,
- Odkrywanie reguł jest procesem bardzo złożonym obliczeniowo.

## Odkrywanie reguł asocjacyjnych

- **Celem procesu odkrywania asocjacji** jest znalezienie interesujących zależności lub korelacji, nazwanych ogólnie asocjacjami, pomiędzy danymi w dużych zbiorach danych.
- **Wynikiem procesu odkrywania asocjacji** jest zbiór reguł asocjacyjnych opisujących znalezione zależności lub korelacje pomiędzy danymi.
- **Geneza problemu odkrywania reguł asocjacyjnych** sięga problemu odkrywania asocjacji rozważanego w kontekście tak zwanej analizy koszyka zakupów (ang. MBA - Market Basket Analysis).

## Analiza koszyka zakupów

- **Cel analizy koszyka zakupów:**
  Znalezienie naturalnych wzorców zachowań konsumenckich klientów
- **Wykorzystanie wzorców zachowań:**
  - organizacji półek w supermarkecie,
  - opracowania akcji promocyjnych,
  - opracowania katalogu oferowanych produktów.

## Analiza koszyka zakupów - zastosowanie

- **Znaleziony wzorzec:**
  ktoś kto kupuje paluszki, najczęściej kupuje również piwo
- **Akcja promocyjna:** (typowy trick)
  Ogłoś obniżkę cen paluszków, jednocześnie podnieś piwa
- **Organizacja sklepu:**
  Staraj się umieszczać produkty kupowane wspólnie w przeciwległych końcach sklepu, zmuszając klientów do przejścia przez cały sklep

# Tablica obserwacji - tablica informacyjna

- Dany jest zbiór atrybutów $A = \{A_1, A_2, ..., A_n\}$, oraz zbiór obserwacji $T = \{T_1, T_2, ..., T_m\}$

| $TR_{id}$ | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ |
|---|---|---|---|---|---|
| $T_1$ | 1 | 0 | 1 | 1 | 1 |
| $T_2$ | 1 | 1 | 1 | 0 | 1 |
| $T_3$ | 0 | 1 | 1 | 1 | 1 |
| $T_4$ | 0 | 0 | 1 | 0 | 1 |
| $T_5$ | 1 | 0 | 1 | 0 | 0 |
| $T_6$ | 1 | 1 | 1 | 1 | 1 |
| $T_7$ | 1 | 0 | 0 | 0 | 1 |
| $T_8$ | 1 | 1 | 1 | 0 | 0 |

## Tablica obserwacji - tablica informacyjna

Tablicę obserwacji można wykorzystać również do analizy koszyka zakupów. Zbiór atrybutów tablicy obserwacji odpowiada liście produktów oferowanych przez supermarket, natomiast wiersze tablicy reprezentują klientów i ich koszyki zakupów.

- Atrybuty tablicy reprezentują wystąpienia encji "produkty",
- Wiersze tablicy reprezentują wystąpienia encji "koszyki",
- Dodatkowy atrybut $TR_{id}$ – wartościami atrybutu są identyfikatory poszczególnych obserwacji
- Pozycja $T_i[A_j] = 1$ tablicy wskazuje, że $i$-ta obserwacja zawiera wystąpienie $j$-tego atrybutu

## Reguły asocjacyjne

- Wynikiem analizy koszyka jest zbiór reguł asocjacyjnych postaci następującej relacji:

$$\{(A_{i1} = 1) \wedge ... \wedge (A_{ik} = 1)\} \Rightarrow \{(A_{ik+1} = 1) \wedge ... \wedge (A_{ik+l} = 1)\}$$

- interpretacja reguły:
  "Jeżeli klient kupił produkty:
  $A_{i1}$, $A_{i2}$, ..., $A_{ik}$,
  to prawdopodobnie kupił również produkty:
  $A_{ik+1}$, $A_{ik+2}$, ..., $A_{ik+l}$"

## Reguły asocjacyjne

- Regułę asocjacyjną można przedstawić jednoznacznie w równoważnej postaci:

$$\theta \rightarrow \phi$$

$$(A_{i1}, A_{i2}, ..., A_{ik}) \rightarrow (A_{ik+1}, A_{ik+2}, ..., A_{ik+l})$$

- Z każdą regułą asocjacyjną $\theta \rightarrow \phi$ związane są dwie podstawowe miary określające statystyczną ważność i siłę reguły:
  - **wsparcie** $supp(\theta \rightarrow \phi)$
  - **ufność** $conf(\theta \rightarrow \phi)$

## Statystyczna ważność i siła reguły

- **Wsparciem reguły asocjacyjnej** sup $\theta \rightarrow \phi$ nazywać będziemy stosunek liczby obserwacji, które spełniają warunek $\theta \wedge \phi$, do liczby wszystkich obserwacji:
  *(wsparcie reguły = prawdopodobieństwu zajścia zdarzenia $\theta \wedge \phi$)*
- **Ufnością reguły asocjacyjnej** conf $\theta \rightarrow \phi$ nazywać będziemy stosunek liczby obserwacji, które spełniają warunek $\theta \wedge \phi$, do liczby obserwacji, które spełniają warunek $\theta$
  *(ufność reguły = warunkowemu prawdopodobieństwu $p(\phi|\theta)$*

## Reguły asocjacyjne – miary

- **Wsparcie** $(X \to Y)$ oznacza liczbę transakcji w bazie danych, które potwierdzają daną regułę
  – miara wsparcia jest symetryczna względem zbiorów stanowiących poprzednik i następnik reguły,
- **Ufność**$(X \to Y)$ oznacza stosunek liczby transakcji zawierających $X \cup Y$ do liczby transakcji zawierających $Y$
  – miara ta jest asymetryczna względem zbiorów stanowiących poprzednik i następnik reguły

# Wsparcie i ufność reguł - przykład

| $Trans_{Id}$ | $Produkty$ |
|---|---|
| 100 | A, B, C |
| 200 | A, C |
| 300 | A, D |
| 400 | B, E, F |

- w przedstawionej bazie danych można znaleźć przykładowe reguły asocjacyjne:
  - $A \rightarrow C$ $sup = \frac{1}{2}$, $conf = \frac{2}{3}$,
  - $C \rightarrow A$ $sup = \frac{1}{2}$, $conf = 1$.

## Algorytm apriori

Założenia:

- Zakładamy, że wszystkie transakcje są wewnętrznie uporządkowane,
- Kolekcję zbiorów częstych o rozmiarze $k$, nazywanych częstymi zbiorami $k$-elementowymi,
- Kolekcję zbiorów kandydujących o rozmiarze $k$, nazywanych kandydującymi zbiorami $k$-elementowymi,
- Dowolny niepusty podzbiór zbioru częstego jest zbiorem częstym.

# Algorytm apriori

- Algorytm rozpoczynamy od znalezienia zbiorów częstych jednoelementowych będących kandydatami do zbiorów częstych,
- posługując się kryterium liczności (supp.) odrzucamy te zbiory, które nie są odpowiednio liczne. Utrzymane zbiory są zbiorami częstymi,
- Poszukujemy kolejnych (o 1 większych) zbiorów częstych w oparciu o uzyskane już zbiory częste
- Zbiory uznajemy za częste o ile:
  - mają wymaganą liczność,
  - wszystkie jego podzbiory poprzednio uznane zostały za zbiory częste
- Zwiększamy liczność kandydatów do zbiorów częstych, aż do momentu gdy nie będzie można utworzyć zbioru częstego

# Algorytm apriori - przykład

Rozważmy zbiór transakcji:

| transakcja | koszyk zakupów |
|---|---|
| $t_1$ | a,b |
| $t_2$ | a,c,d,e |
| $t_3$ | a,b,c,e |
| $t_4$ | c,d |
| $t_5$ | b,c,d,e |
| $t_6$ | a,d,e |
| $t_7$ | c,d,e |

Chcemy znaleźć wszystkie zbiory częste o wsparciu 3. Rozpoczynamy od zbiorów jednoelementowych:

kandydaci:

| zbiory | wsparcie |
|---|---|
| $\{a\}$ | 4 |
| $\{b\}$ | 3 |
| $\{c\}$ | 5 |
| $\{d\}$ | 5 |
| $\{e\}$ | 5 |

zbiory częste:

| zbiory | wsparcie |
|---|---|
| $\{a\}$ | 4 |
| $\{b\}$ | 3 |
| $\{c\}$ | 5 |
| $\{d\}$ | 5 |
| $\{e\}$ | 5 |

Elementy zbiorów częstych 1-elementowych: $\{a, b, c, d, e\}$

# Algorytm apriori - przykład cd.

Z elementów zbiorów częstych 1-elementowych: $\{a, b, c, d, e\}$ konstruujemy kandydatów na zbiory częste dwuelementowe:

kandydaci:

| zbiory | wsparcie |
|---|---|
| $\{a, b\}$ | 2 |
| $\{a, c\}$ | 2 |
| $\{a, d\}$ | 2 |
| $\{a, e\}$ | 3 |
| $\{b, c\}$ | 2 |
| $\{b, d\}$ | 1 |
| $\{b, e\}$ | 2 |
| $\{c, d\}$ | 3 |
| $\{c, e\}$ | 4 |
| $\{d, e\}$ | 4 |

zbiory częste:

| zbiory | wsparcie |
|---|---|
| $\{a, e\}$ | 3 |
| $\{c, d\}$ | 3 |
| $\{c, e\}$ | 4 |
| $\{d, e\}$ | 4 |

Elementy zbiorów częstych 2-elementowych: $\{a, c, d, e\}$

# Algorytm apriori - przykład cd.

Z elementów zbiorów częstych 2-elementowych: $\{a, c, d, e\}$ konstruujemy kandydatów na zbiory częste trójelementowe:

kandydaci:

| zbiory | wsparcie |
|---|---|
| $\{a, c, d\}$ | 1 |
| $\{a, c, e\}$ | 2 |
| $\{a, d, e\}$ | 2 |
| $\{c, d, e\}$ | 3 |

zbiory częste:

| zbiory | wsparcie |
|---|---|
| $\{c, d, e\}$ | 3 |

- Jest tylko jeden zbiór częsty 3-elementowy: $\{c, d, e\}$
  - Wsparcie tego zbioru spełnia wymogi $supp(\{c, d, e\}) = 3$,
  - Każdy podzbiór zbioru $\{c, d, e\}$ jest zbiorem częstym tj. zbiory $\{\{c, d\}, \{c, e\}, \{d, e\}, \{c\}, \{d\}, \{e\}\}$ są zbiorami częstymi.
- Nie można skonstruować większych zbiorów częstych.

# Reguły asocjacyjne

Dla zbiorów częstych dwu i więcej elementowych możemy napisać reguły asocjacyjne i ufność tych reguł

- Dla zbioru $\{c, d, e\}$:
  $c \to \{d, e\}\ conf = \frac{supp(\{c,d,e\})}{supp(c)} = \frac{3}{5}$
  $d \to \{c, e\}\ conf = \frac{supp(\{c,d,e\})}{supp(d)} = \frac{3}{5}$
  $e \to \{c, d\}\ conf = \frac{supp(\{c,d,e\})}{supp(e)} = \frac{3}{5}$
  $\{c, d\} \to e\ conf = \frac{supp(\{c,d,e\})}{supp(\{c,d\})} = \frac{3}{3}$
  $\{d, e\} \to c\ conf = \frac{supp(\{c,d,e\})}{supp(\{d,e\})} = \frac{3}{4}$
  $\{c, e\} \to d\ conf = \frac{supp(\{c,d,e\})}{supp(\{c,e\})} = \frac{3}{4}$
- Dla zbiorów dwuelementowych:
  dla zbioru $\{a, e\}$:
  $a \to e\ conf = \frac{supp(\{a,e\})}{supp(a)} = \frac{3}{4}$

$e \rightarrow a\ conf = \frac{supp(\{a,e\})}{supp(e)} = \frac{3}{5}$

dla zbioru $\{c, d\}$:

$c \rightarrow d\ conf = \frac{supp(\{c,d\})}{supp(c)} = \frac{3}{5}$

$d \rightarrow c\ conf = \frac{supp(\{c,d\})}{supp(d)} = \frac{3}{5}$

dla zbioru $\{c, e\}$:

$c \rightarrow e\ conf = \frac{supp(\{c,e\})}{supp(c)} = \frac{4}{5}$

$e \rightarrow c\ conf = \frac{supp(\{c,e\})}{supp(e)} = \frac{4}{5}$

dla zbioru $\{d, e\}$:

$d \rightarrow e\ conf = \frac{supp(\{d,e\})}{supp(d)} = \frac{4}{5}$

$e \rightarrow d\ conf = \frac{supp(\{d,e\})}{supp(e)} = \frac{4}{5}$

Wybieramy tylko te reguły co spełniają kryterium ufności tj conf=0.7