Procesy i systemy Business Intelligence

Wprowadzenie do eksploracji danych

Geneza (1)

- Dostępność danych
  - Rozwój nowoczesnych technologii przechowywania i przetwarzania danych (systemy baz danych, hurtownie danych, repozytoria danych)
  - Upowszechnienie systemów informatycznych we wszystkich praktycznie dziedzinach życia (bankowość, ubezpieczenia, administracja, medycyna, nauka, sport, handel, produkcja, marketing itd.
  - Spadek cen sprzętu komputerowego

Geneza (2)

- Jaka jest wartość nagromadzonych danych z punktu widzenia przedsiębiorstwa?
  - służą one do obsługi i wspomagania bieżącej działalności przedsiębiorstw
  - zawierają bardzo często istotną wiedzę o otaczającym nas świecie
  - nagromadzone mogą zawierać istotną wiedzę o prawidłowościach i regułach procesów biznesowych, zachowaniach klientów, o zależnościach występujących pomiędzy danymi generowanymi przez różne procesy
- Dylemat przedsiębiorstw: w jaki sposób efektywnie i racjonalnie wykorzystać nagromadzoną w danych wiedzę dla celów wspomagania swojej działalności?

Czym jest eksploracja danych

- ***Eksploracja danych***: zbiór metod automatycznego odkrywania nietrywialnych, dotychczas nieznanych, potencjalnie użytecznych reguł, zależności, wzorców schematów, podobieństw lub trendów (ang. patterns) w dużych repozytoriach danych (bazach danych, hurtowniach danych, itp.)
- Celem eksploracji danych jest analiza danych i procesów w celu lepszego ich rozumienia

Metody eksploracji danych

- odkrywanie asocjacji
- klasyfikacja/regresja
- grupowanie
- odkrywanie sekwencji
- odkrywanie charakterystyk
- analiza przebiegów czasowych
- wykrywanie zmian i odchyleń
- eksploracja WWW
- eksploracja dokumentów tekstowych
- itd.

Metody eksploracji: odkrywanie asocjacji

- odkrywanie asocjacji: znajdowanie związków pomiędzy występowaniem grup elementów w zbiorach danych
- przykłady asocjacji:
  - klienci, którzy kupują pieluszki, kupują również piwo
  - klienci, którzy kupują chleb, masło i ser, kupują również wodę mineralną i ketchup
  - klienci (ubezpieczaln)i, którzy mają poniżej 25 lat często powodują wypadki drogowe
- zastosowania odkrytych asocjacji:
  - planowanie kampanii promocyjnych
  - planowanie rozmieszczenia stoisk sprzedaży w supermarketach

Metody eksploracji: odkrywanie wzorców sekwencji

- odkrywanie wzorców sekwencji: znajdowanie najczęściej występujących sekwencji zdarzeń lub elementów
- przykłady wzorców sekwencji:
  - klienci, którzy kupili farbę emulsyjną, kupią w najbliższym czasie pędzel płaski
  - klienci, którzy realizowali dostęp do strony A, w kolejnym kroku przejdą na stronę C, a następnie, na stronę D
- zastosowania odkrytych wzorców sekwencji:
  - planowanie inwestycji giełdowych
  - przewidywanie sprzedaży
  - znajdowanie skutecznej terapii
  - znajdowanie profili klientów serwisu web-owego

Metody eksploracji: klasyfikacja

- klasyfikacja: predykcja wartości określonego atrybutu w oparciu o pewien zbiór danych treningowych
- przykład klasyfikacji: automatyczny podział kierowców na powodujących i nie powodujących wypadki drogowe:
  - kierowcy prowadzący czerwone pojazdy o pojemności 650 ccm powodują wypadki drogowe
  - kierowcy, którzy posiadają prawo jazdy ponad 7 lat lub jeżdżą niebieskimi samochodami nie powodują wypadków drogowych
- zastosowania klasyfikacji:
  - diagnostyka medyczna
  - rozpoznawanie trendów na rynkach finansowych
  - przydział kredytów bankowych

Metody eksploracji: grupowanie

- grupowanie: znajdowanie „naturalnego” pogrupowania (podziału) obiektów w oparciu o ich wartości
- przykłady grupowania:
  - automatyczne grupowanie dokumentów tekstowych (np. maili)
  - grupowanie klientów serwisu
  - grupowanie konsumentów energii elektrycznej
- zastosowania grupowania:
  - systemy rekomendacyjne (grupowanie klientów)
  - wyszukiwanie informacji w sieci web (np. grupowanie stron www)
  - astronomia
  - handel elektroniczny

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 większą niż 37.5 C, bólem gardła, osłabieniem organizmu
  - automatyczne tworzenie streszczeń dokumentów
  - automatyczne tworzenie charakterystyk produktów na podstawie informacji z blogów i forów internetowych
- zastosowania odkrywania charakterystyk:
  - znajdowanie zależności funkcyjnych pomiędzy zmiennymi
  - określanie profilu klienta - zbioru cech charakterystycznych

Metody eksploracji: odkrywanie punktów osobliwych

- odkrywanie punktów osobliwych: znajdowanie obiektów (zdarzeń) odbiegających znacząco od modelu pozostałych obiektów (zdarzeń) analizowanego zbioru danych
- przykład odkrywania punktów osobliwych:
  - znajdowanie klientów, których konsumpcja energii odbiega znacząco od innych klientów o podobnej charakterystyce
  - znajdowanie pacjentów, których wyniki odbiegają znacząco od wyników analiz innych pacjentów chorujących na ta samą chorobę
- zastosowania odkrywania punktów osobliwych:
  - wykrywanie oszustw podatkowych, kradzieży prądu, itp..
  - astronomia, fizyka – odkrywanie obiektów o nieznanej dotychczas charakterystyce

Metody eksploracji: eksploracja sieci www

- eksploracja sieci www: metody analizy korzystania z sieci web w celu :
  - znajdowania typowych wzorców zachowań użytkowników sieci
  - znajdowania powiazań stron w sieci web w celu określenia ważności i koncentratywności stron (w celu poprawy efektywności procesu wyszukiwania stron)
  - grupowania i klasyfikacji stron WWW na podstawie ich zawartości i schematu zewnętrznego
  - znajdowania ukrytych „stron lustrzanych” i wewnętrznych „środowisk” (ang. *communities*) oraz analiza ich ewolucji w czasie
  - analizy reklam internetowych (ich efektywności, rozliczania i propagacji).

Metody eksploracji: eksploracja danych multimedialnych i przestrzennych

- metody analizy i eksploracji baz danych przechowujących obrazy, mapy, dźwięki, wideo itp.
- celem jest wspomaganie procesów wyszukiwania danych (wyszukiwanie na podstawie zawartości, wideo na zadanie itd.)
- metody służące do grupowania i klasyfikacji danych multimedialnych są najczęściej silnie powiązane z mechanizmami systemu zarządzania bazą danych (indeksowanie i buforowanie danych)

Metody eksploracji: eksploracja struktur grafowych

- struktury grafowe są szeroko stosowane do modelowania złożonych obiektów, takich jak: obwody elektroniczne, związki chemiczne, struktury białkowe, sieci biologiczne, sieci społecznościowe, procedury obiegu dokumentów, dokumenty XML
- metody analizy struktur grafowych: grupowanie i klasyfikacja struktur grafowych, odkrywanie częstych podstruktur (podgrafów) w bazie danych struktur grafowych, klasyfikacja struktur grafowych umożliwiająca znajdowanie zależności pomiędzy pewną charakterystyką struktury grafowej a jej budową (np. analiza i klasyfikacja sekwencji DNA)

Metody eksploracji: eksploracja sieci społecznościowych

- algorytmy analizy sieci społecznościowe wspomagające:
  - procesy wykrywania oszustów uczestniczących w aukcjach internetowych,
  - wykrywanie przestępstw w kryminalistyce,
  - analizę dużych sieci elektrycznych i telekomunikacyjnych itp.
  - powiazania pomiędzy uczestnikami gier i aukcji internetowych
  - wykrywanie środowisk w sieciach społecznościowych
  - rozpowszechnianie się epidemii, itp.