Odkrywanie wiedzy

Marcin Szeląg

Zakład ISWD, Instytut Informatyki, Politechnika Poznańska

09.10.2019

Plan prezentacji

1. Informacje organizacyjne
2. Zakres tematyczny przedmiotu
3. Wprowadzenie do Odkrywania Wiedzy
4. Podsumowanie

Informacje organizacyjne

- Wykład: środa, 11.45 - 13.15, sala 128 (BT),
  zaliczenie na podstawie **egzaminu**
- Laboratoria: czwartek, 09.45-11.15 oraz 13.30-15.00, lab. 44 (CW),
  zaliczenie na podstawie ocen ze **sprawdzianów**, aktywności na zajęciach i oceny z **projektu**
- Prowadzący: dr inż. Marcin Szeląg, pok. L.1.6.7,
  marcin.szelag@cs.put.poznan.pl,
  `http://www.cs.put.poznan.pl/mszelag`, $\rightarrow$ Teaching (pl)
  konsultacje: wtorek 12.30 - 13.15, czwartek 11:30-12:30

Zakres tematyczny przedmiotu

- Wprowadzenie do odkrywania wiedzy
- Wybrane techniki przetwarzania danych (selekcja cech, redukcja rozmiaru danych, algorytmy dyskretyzacji, itd.) i ich wizualizacji
- Indukcja regułowych modeli wiedzy, algorytm VCDomLEM
- Analiza danych klasyfikacyjnych z brakującymi wartościami atrybutów
- Uczenie się preferencji w problemie rankingu
- Programowanie inteligentnych silników w grach
- Analiza skupień (grupowanie) - metody numeryczne i konceptualne
- Wnioskowanie na podstawie podobieństwa do znanych przypadków (CBR)
- Teoria i zastosowania zbiorów rozmytych (ang. fuzzy sets)
- Wybrane systemy odkrywania wiedzy

Wprowadzenie do Odkrywania Wiedzy

Motywacje

- Rozwój technologii automatycznego gromadzenia i przechowywania informacji – "eksplozja danych"
- Wzrost rozmiarów przechowywanych danych
- Wzrost mocy obliczeniowej i pojemności pamięci
- Możliwości analizowania i rozumienia dużych wolumenów danych sa znacznie mniejsze od możliwości ich gromadzenia i przechowywania
- Posiadanie danych nie jest równoznaczne z posiadaniem (ew.) wiedzy zawartej w tych danych
- Analizy dużych wolumenów danych celem wyciągania użytecznych wniosków, prowadzących do podejmowania lepszych decyzji

"We are data rich, but knowledge poor".

Historia

1989 – Workshop on Knowledge Discovery in Databases (Detroit)

`https://www.kdnuggets.com/meetings/kdd89`

Czym jest odkrywana wiedza?

"Wiedza jest uporządkowanym zbiorem interesujących i użytecznych regularności", G. Piatetsky-Shapiro (1991)

- Regularność (wzorzec, ang. pattern) – zależność między elementami danych
- Interesujący – nowy (poprzednio nieznany i nieoczekiwany), nietrywialny i zrozumiały
- Użyteczny – nadający się do wykorzystania w przyszłych działaniach

Odkrywanie wiedzy w danych

Odkrywanie wiedzy to (interaktywny i iteracyjny) proces poszukiwania nowych (nieoczekiwanych), potencjalnie użytecznych i zrozumiałych regularności w danych. Jego celem jest przejście od "surowych" danych do zbioru wzorców, które mogą być następnie wykorzystane w procesie wspomagania podejmowania decyzji.

Odkrywanie wiedzy jako proces

Odkrywanie wiedzy (ang. knowledge discovery) postrzegane jest jako proces, na który składają się następujące etapy:

1. analiza i poznanie dziedziny zastosowania, identyfikacja dostępnej wiedzy i celów użytkownika,
2. integracja danych z różnych źródeł,
3. selekcja/tworzenie danych,
4. czyszczenie i wstępne przetwarzanie danych,
5. transformacja danych (przekształcenie i redukcja danych),
6. wybór zadania/zadań (metody/metod) eksploracji danych,
7. wybór algorytmu/algorytmów eksploracji danych,
8. eksploracja danych (ang. Data Mining – DM),
9. interpretacja, analiza i ocena odkrytej wiedzy, wizualizacja odkrytych wzorców,
10. przygotowanie wiedzy do użycia.

Odkrywanie wiedzy jako proces – c.d.

ad 2. *Integracja danych z różnych źródeł* – celem etapu jest integracja danych z różnych heterogenicznych i rozproszonych źródeł danych w jeden zintegrowany zbiór danych.
ad 3. *Selekcja danych* – celem etapu jest selekcja danych istotnych z punktu widzenia procesu analizy danych.
ad 4. *Czyszczenie danych* – celem etapu jest usunięcie niepełnych, niepoprawnych lub nieistotnych danych ze zbioru eksplorowanych danych.
ad 5. *Transformacja danych* – celem etapu jest transformacja wyselekcjonowanych danych do postaci wymaganej przez metody eksploracji danych.
ad 9. *Ocena odkrytych wzorców* – celem etapu jest identyfikacja interesujących wzorców oraz ich wizualizacja w taki sposób, aby umożliwić użytkownikowi ich interpretację i zrozumienie.

Interaktywność procesu odkrywania wiedzy

J. Stefanowski, "Algorytmy indukcji reguł decyzyjnych w odkrywaniu wiedzy", rozprawa habilitacyjna, 2001:

"... użytkownik systemu odkrywania wiedzy powinien posiadać **dobre zrozumienie dziedziny zastosowania**, tak aby wybrać właściwy podzbiór danych, określić jakie są zadania analizy, jaka powinna być reprezentacja poszukiwanej wiedzy, których algorytmów należy użyć. Dlatego system odkrywania wiedzy powinien być oprogramowaniem **interaktywnym**, a nie narzędziem w pełni automatycznym. Sam proces odkrywania wiedzy jest procesem intensywnego współdziałania człowieka z oprogramowaniem i składa się najczęściej z **wielu iteracji** obejmujących modyfikacje wstępnych specyfikacji i powtarzanie niektórych kroków".

Eksploracja danych

Eksploracja danych – etap w procesie odkrywania wiedzy – automatyczne odkrywanie nietrywialnych, dotychczas nieznanych, potencjalnie użytecznych reguł, zależności, wzorców schematów, podobieństw lub trendów w dużych repozytoriach danych (bazach danych, hurtowniach danych, itp.).

Eksploracja danych wykonywana może być w perspektywie predykcji i/lub opisu danych. W trakcie eksploracji danych możliwe jest wykorzystanie wiedzy dziedzinowej, np. w postaci hierarchii kategorii.

Typowe reprezentacje poszukiwanej wiedzy

Poszukiwana wiedza może mieć różną postać:

- wartości miar statystycznych, opisy charakterystyczne/dyskryminujące,
- reguły asocjacyjne,
- drzewa i reguły klasyfikacyjne,
- funkcje i równania,
- klauzule logiczne,
- skupienia i ich opis,
- taksonomia (hierarchia),
- trendy i zależności czasowe,
- ...

Zadania eksploracji danych

- Opis danych (podsumowywanie / charakterystyki danych)
- Odkrywanie reguł asocjacyjnych
- Klasyfikacja (zmienna wyjściowa jest jakościowa)
- Regresja (zmienna wyjściowa jest liczbowa)
- Predykcja (wynik dotyczy przyszłości)
- Grupowanie danych (analiza skupień)
- Odkrywanie wzorców sekwencji
- Odkrywanie zależności funkcyjnych
- Analiza przebiegów czasowych
- Poszukiwanie obserwacji osobliwych, odchyleń, anomalii
- Eksploracja dokumentów tekstowych i WWW (ang. text and web mining) – wyszukiwanie według zawartości
- Analiza danych strumieniowych, napływających z sensorów
- ...

Przegląd zadań eksploracji danych

- **Opis danych** – odkrywanie charakterystyk. Przykładem może być opis pacjentów chorujących na anginę: "Pacjenci chorujący na anginę cechują się temperaturą ciała wiekszą niż 37.5C, bólem gardła i osłabieniem organizmu". W celu opisania danych wykorzystuje się często Eksploracyjną Analizę Danych (ang. Exploratory Data Analysis – EDA), której techniki są zwykle interaktywne i wizualne.
- **Odkrywanie reguł asocjacyjnych** – poszukiwanie w danych zbioru reguł asocjacyjnych opisujących zależności lub korelacje między danymi. Przykładem reguły asocjacyjnej jest reguła wygenerowana w odniesieniu do bazy danych supermarketu: "Klienci, którzy kupują pieluszki, kupują również piwo".

Przegląd zadań eksploracji danych – c.d.

- **Klasyfikacja** – predykcja przynależności do klas decyzyjnych (kategorii) w oparciu o zbiór danych treningowych; techniki: statystyczne, drzewa decyzyjne, reguły decyzyjne, sieci neuronowe, k-NN, ...
- **Regresja** (aproksymacja, szacowanie, estymacja) – predykcja wartości funkcji rzeczywistej w oparciu o zbiór danych treningowych; techniki: sieci neuronowe, statystyczne metody regresji
- **Grupowanie** – znajdowanie skończonych zbiorów klas obiektów posiadających podobne cechy. W przeciwieństwie do metod klasyfikacji i predykcji, klasyfikacja obiektów (podział na klasy) nie jest znana a-priori, lecz jest celem metod grupowania. Metody te grupują obiekty w klasy w taki sposób, aby maksymalizować podobieństwo wewnątrzklasowe obiektów i minimalizować podobieństwo pomiędzy klasami obiektów.

Przegląd zadań eksploracji danych – c.d.

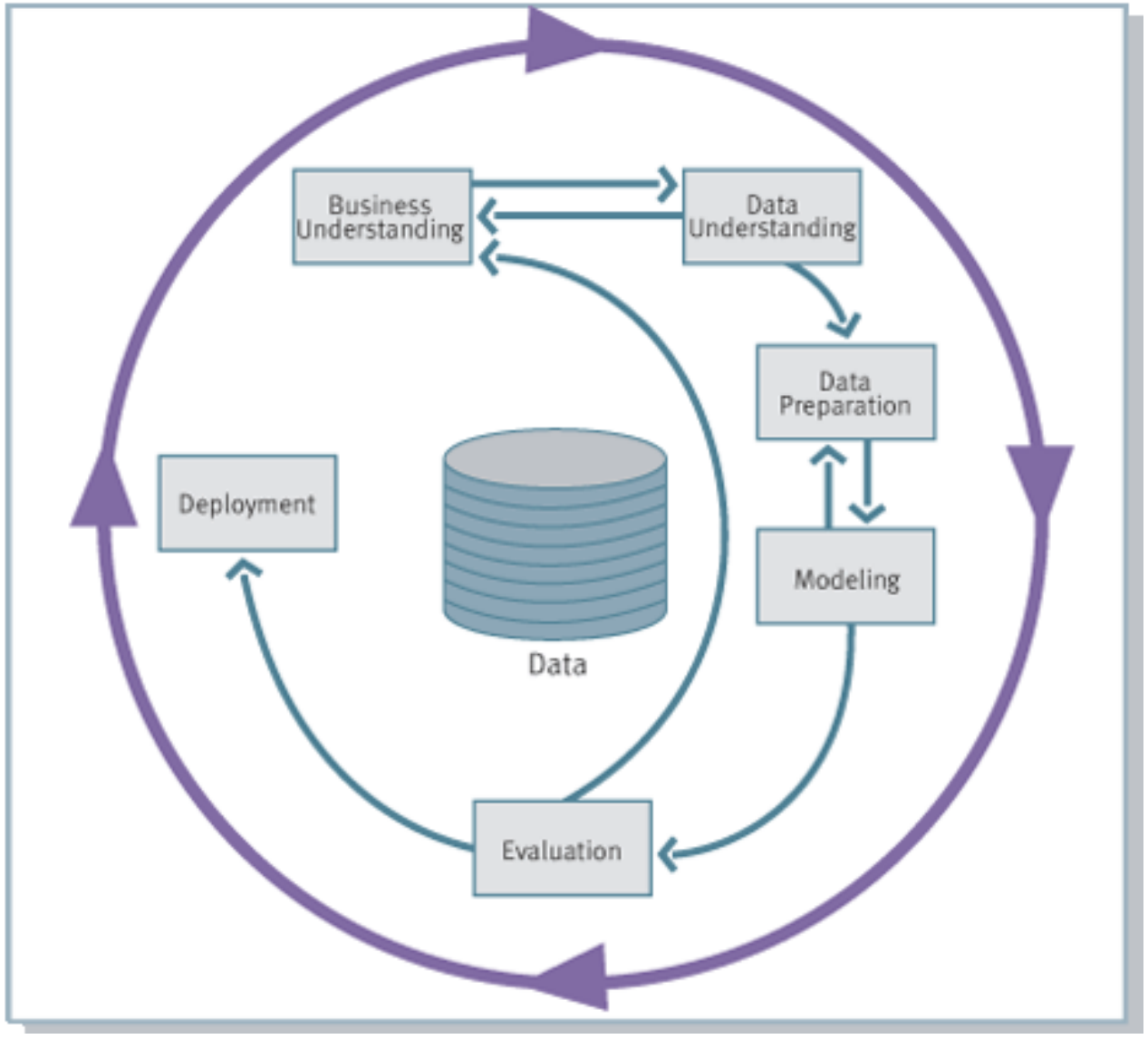
- **Odkrywanie wzorców sekwencji** – poszukiwanie zależnosci pomiędzy występowaniem określonych zdarzeń w czasie. Przykładem wzorca sekwencji dla bazy danych wypożyczalni filmów wideo może być: "Klient, który wypożyczył Gwiezdne Wojny, w ciagu tygodnia wypożyczy film Imperium Kontratakuje, a następnie, w ciagu kolejnego tygodnia, wypożyczy film Powrót Jedi". Zdarzenia wchodzące w skład wzorca sekwencji nie muszą występować bezpośrednio po sobie – mogą być przedzielone wystąpieniem innych zdarzeń.
- **Odkrywanie zależności funkcyjnych** – poszukiwanie wzorów najlepiej wyrażających zależności występujące pomiędzy atrybytami o wartościach liczbowych. Jest to w pewnym sensie uogólnienie regresji na dowolną liczbę atrybutów zależnych z dodatkowym wymogiem, aby zależność była wyrażona za pomocą formuły algebraicznej. Do znajdowania takich zależności wykorzystuje się metody odkrywania równań.

Przegląd zadań eksploracji danych – c.d.

- **Wyszukiwanie według zawartości** – poszukiwanie wzorców podobnych do podanego wzorca. Zadanie to jest wykonywane najczęściej na zbiorach danych zawierających teksty i obrazy. W przypadku tekstu wzorzec może być np. zbiorem słów kluczowych, a zbiorem danych np. strony WWW (wyszukiwanie dokumentów w Internecie). W przypadku obrazów, użytkownik może posiadać przykładowy obraz, szkic lub opis obrazu i chce znaleźć podobne obrazy w dostępnym zbiorze obrazów $\rightarrow$ QBIC (Query by Image Content) IBM'a (1995), umożliwiający interaktywne przeszukiwanie bazy obrazów i filmów video z wykorzystaniem deskryptorów zawartości (kolor, kształt, tekstura, szkic).

CRISP-DM – Cross-Industry Standard Process for DM

http://crisp-dm.eu



CRISP-DM – Cross-Industry Standard Process for DM – c.d.

- Metodologia stworzona w 1996 roku przez analityków z DaimlerChrysler, SPSS i NCR
- Ogólnie dostępny standardowy proces dopasowania eksploracji danych do ogólnej strategii rozwiązywania problemów
- CRISP-DM jest procesem iteracyjnym (zewn. strzałki) i adaptacyjnym (wewn. strzałki)

CRISP-DM – Cross-Industry Standard Process for DM – c.d.

Cykl życia projektu odkrywania wiedzy składa się z 6 etapów:

1. **Zrozumienie uwarunkowań biznesowych/badawczych**:
   - Sformułowanie celów projektu
   - Wykorzystanie celów i ograniczeń do opracowania definicji problemu eksploracji danych
   - Stworzenie wstępnego planu działań, zmierzających do osiągnięcia zamierzonych celów
2. **Zrozumienie danych**:
   - Zebranie danych
   - Wstępna analiza danych, mająca na celu zaznajomienie się z danymi
   - Ocena jakości danych
   - Ewentualny wybór interesującego podzbioru danych

CRISP-DM – Cross-Industry Standard Process for DM – c.d.

Cykl życia projektu odkrywania wiedzy składa się z 6 etapów:

3. **Przygotowanie danych**:
   - Przygotowanie ostatecznego zbioru danych z danych "surowych" (duża pracochłonność!)
   - Wybór przypadków i atrybutów do analizy
   - Ewentualna transformacja atrybutów
   - Czyszczenie danych pod kątem narzędzi modelujących
4. **Modelowanie**:
   - Wybór metody (metod) modelujących
   - Kalibracja parametrów w celu optymalizacji wyników
   - Wykorzystanie metody (metod) modelujących
   - W przypadku stosowania wielu metod, często zachodzi konieczność powrotu do poprzedniego etapu w celu dostosowania danych do konkretnej metody

CRISP-DM – Cross-Industry Standard Process for DM – c.d.

Cykl życia projektu odkrywania wiedzy składa się z 6 etapów:

5 **Ewaluacja**:
- Ocena modelu (modeli) z etapu modelowania pod kątem jakości i efektywności
- Ustalenie czy model spełnia założone cele
- Ustalenie czy są jakieś cele biznesowe lub badawcze, które nie zostały w należy sposób uwzględnione
- Podjęcie decyzji co do wykorzystania wyników modelowania

6 **Wdrożenie**:
- Umożliwienie wykorzystania stworzonego modelu. Stworzenie modelu zasadniczo nie stanowi zakończenia projektu – konieczne jest zaprezentowanie wiedzy w formie przystępnej dla klienta
- Przykład prostego wdrożenia: sporządzenie raportu
- Przykład złożonego wdrożenia: implementacja powtarzalnego procesu eksploracji danych

Eksploracja danych a inne systemy informatyczne

Bazy danych

- Zapytania SQL i raporty (zapytania operacyjne):
  - Który klient dokonał największego zakupu?
  - Jak wygląda lista klientów, którzy zakupili produkt A w ostatnim roku?

Hurtownie danych

- Wielowymiarowa agregacja danych i podsumowania (zapytania analityczne oparte o model OLAP):
  - Jakie są średnie zakupy klientów, którzy kupili produkt A w ostatnim roku, z podziałem na regiony?

Zaawansowane systemy eksploracji danych

- Opis lub predykcja (zapytania eksploracyjne):
  - Jakie są cechy charakterystyczne klientów, którzy mogą kupić produkt A?
  - Do kogo skierować ofertę reklamową?

Źródła danych

Dane generowane są przez:

- banki, ubezpieczalnie, firmy, sieci handlowe, szpitale, etc.,
- eksperymenty naukowe: fizyka, astronomia, biologia, etc.,
- web, tekst, e-handel,
- ...

Przykłady

- Very Long Baseline Interferometry (VLBI) - sieć radioteleskopów, z których każdy produkuje > 1 Gb/s danych.
- AT&T obsługuje miliardy połaczeń dziennie. Danych jest tyle, że nie można ich wszystkich zapamiętać – analiza tych danych jest wykonywana "w locie" (ang. "on the fly") – tzw. strumienie danych.
- Sieć sprzedaży Wal-Mart gromadzi dziennie dane dotyczace ponad 20 milionów transakcji.

Przykłady – c.d.

- Koncern Mobil Oil posiada magazyn danych pozwalający na przechowywanie setek terabajtów danych o wydobyciu ropy naftowej.
- System NASA satelitarnej obserwacji ziemi (EOS) generuje w ciągu każdej godziny gigabajty danych obrazowych.
- Niewielkie supermarkety rejestrują codziennie sprzedaż tysięcy artykułów.

5 milionów TB wygenerowanych w 2002

- UC Berkeley 2003 szacuje:
  - 5 exabytes (5 million terabytes) nowych danych wygenerowanych w 2002 roku – `http://www2.sims.berkeley.edu/research/projects/how-much-info-2003`
  - Najwiekszy "producent danych" – USA – produkuje ok. 40% danych światowych
- `https://www.forbes.pl/technologie/jak-wiele-danych-produkujemy-kazdego-dnia/4mn4w69` "- W 1992 r. na świecie powstawało 100 GB danych dziennie, w 1997 r. – 100 GB na godzinę, w 2002 r. – 100 GB na sekundę, w 2018 r. przewiduje się przyrost 50 000 GB danych na sekundę."

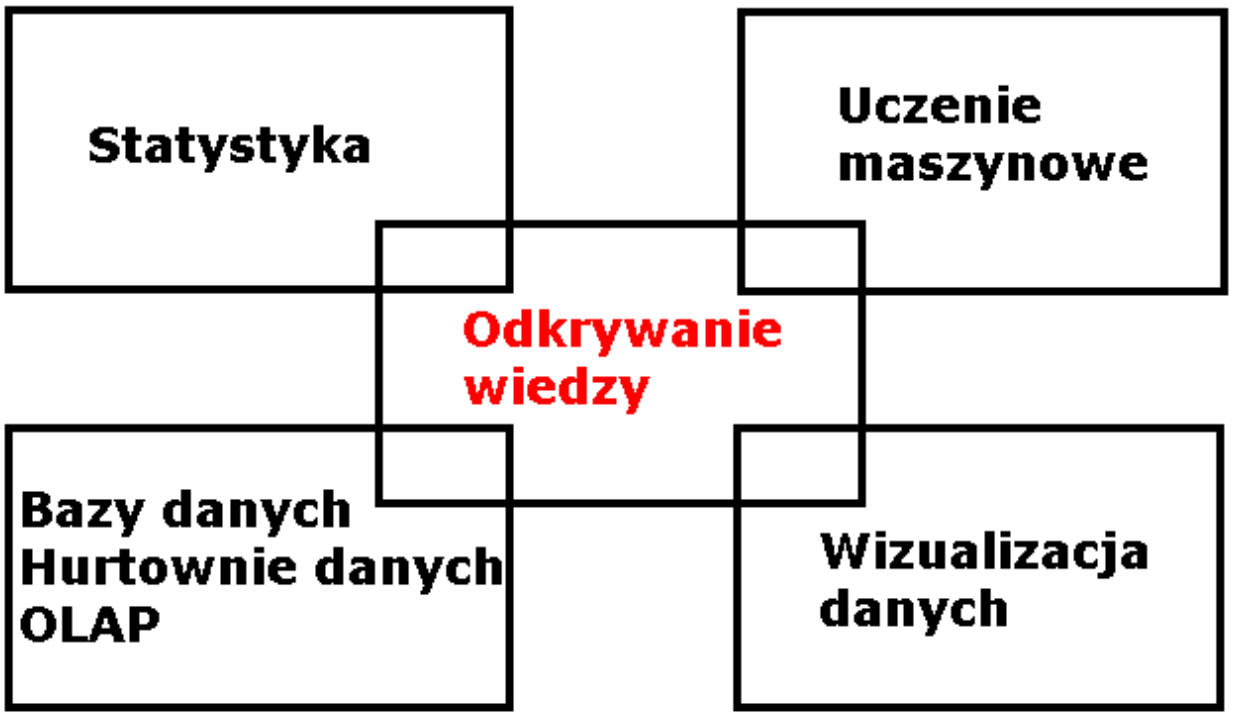
Co można eksplorować?

- Relacyjne bazy danych
- Hurtownie danych
- Repozytoria danych
- Zaawansowane systemy informatyczne
  - Obiektowe i obiektowo-relacyjne bazy danych
  - Przestrzenne bazy danych
  - Przebiegi czasowe i temporalne bazy danych
  - Tekstowe i multimedialne bazy danych
  - WWW
  - ...

Przykłady zastosowań eksploracji danych

- Marketing
  - Identyfikacja profilu klientów, ocena lojalności klientów, problem koszyka zakupów – asocjacje produktów w sieciach sprzedaży, segmentacja rynków, klientów, ...
- Analizy finansowe
  - Analiza ryzyka kredytowego, rekomendacje produktów, przewidywanie trendów i przebiegów czasowych, ...
- Wykrywanie nieprawidłowości i anomalii
  - Analiza defraudacji i nieprawidłowości kart kredytowych, systemy telekomunikacyjne, towarzystwa ubezpieczeniowe, systemy opieki medycznej, ...
- Text mining oraz Web mining (zachowania użytkowników w e-serwisach, wspomaganie wyszukiwania informacji), ...
- Wiele innych (przemysł, nauka, administracja, projektowanie hurtowni danych, optymalizacja działań związanych z systemami CRM, reklamy skierowane), ...

Odkrywanie wiedzy a inne dziedziny



Odkrywanie wiedzy a inne dziedziny – c.d.

- Statystyka:
  - oparta na silnych podstawach teoretycznych i mocnych założeniach dotyczących danych,
  - ukierunkowana na testowanie hipotez oraz estymację parametrów.
- Uczenie maszynowe:
  - nacisk na polepszanie działania przez uczącego się agenta,
  - metody heurystyczne i wywodzące się ze sztucznej inteligencji,
  - ustrukturalizowane dane,
  - problemy dobrze zdefiniowane (zadany cel – wiemy czego szukamy),
  - adaptacja do rzeczywistego i zmiennego środowiska, np. w przypadku robotów (nie rozważane w OW).

Odkrywanie wiedzy a inne dziedziny – c.d.

- Odkrywanie wiedzy:
  - większa różnorodność i złożoność analizowanych danych (nietypowych dla statystycznej analizy danych),
  - często ostateczny cel analizy krystalizuje się dopiero podczas pracy z danymi,
  - często brak jasnej definicji pojęć do odkrycia,
  - nacisk na reprezentacje wiedzy,
  - duża rola przygotowania i wstępnego przetwarzania danych.

Problemy odkrywania wiedzy

- Niespójność danych, szum
- W dużych bazach danych mogą zostać odkryte duże ilości wzorców.
- Człowiek nie potrafi przeanalizować i zrozumieć bardzo dużych zbiorów informacji.
- Różnorodność oczekiwań uczestników procesu odkrywania wiedzy.
- Złożoność obliczeniowa procesu odkrywania wiedzy.

Oprogramowanie

Oprogramowanie wspierające proces odkrywania wiedzy w danych:

- IBM SPSS Modeler,
- StatSoft Statistica,
- WEKA,
- RapidMiner Studio,
- R,
- Matlab/GNU Octave/Scilab,
- KNIME Analytics Platform,
- KEEL,
- Orange,
- ruleRank/RUDE.

Podsumowanie

Podsumowanie

- Problemem nie jest elektroniczne gromadzenie danych ale ich właściwa analiza i wyciąganie użytecznych wniosków.
- Metody statystyczne i uczenia maszynowego mogą być podstawą do odkrywania wiedzy z danych.
- Należy zwracać uwagę na wcześniejsze etapy procesu odkrywania wiedzy, np. integracji danych z różnych źródeł, czyszczenia danych, przetwarzania wstępnego oraz redukcji rozmiarów danych.
- Istotność integracji z hurtowniami danych i biznesowymi systemami wspomagania decyzji.
- Celem jest odkrywanie **interesujących, potencjalnie użytecznych** regularności w danych.

Referencje

- Eksploracja danych. D. Hand, H. Mannila, P. Smyth, WNT, Warszawa, 2005.
- Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych. Daniel T. Larose, PWN, Warszawa, 2006.
- Pattern Recognition and Machine Learning. C.M. Bishop, Springer 2006.
- The Elements of Statistical Learning. Data Mining, Inference, and Prediction. T. Hastie, R. Tibshirani, J. Friedman, Springer, New York, 2016, 2nd edition.
- Data Mining: Practical Machine Learning Tools and Techniques. Ian H. Witten, Eibe Frank, Mark A. Hall, Chris Pal, 4th edition, 2016.
- Data Mining: Concepts and Techniques. Jiawei Han, Micheline Kamber, Jian Pei, 3nd edition, 2011.
- WEKA – open source project = źródła + dokumentacja i podręczniki dostępne na WWW.
- ...

Strony www

- https://www.kdnuggets.com,
  https://www.kdnuggets.com/data_mining_course
- http:
  //www.the-data-mine.com?title=Eksploracja_danych
- http://wazniak.mimuw.edu.pl/index.php
- http://www.sixsigma.pl/textbook/stathome.html
- http://archive.ics.uci.edu/ml/index.php
- https://www.chessprogramming.org
- ...