

Databázové systémy a metody zpracování dat

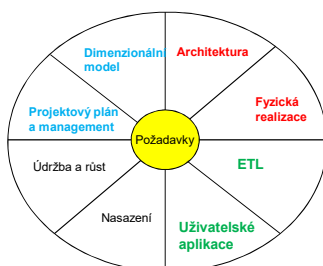
Data Mining

Úvod do problematiky, základní metody

Uživatelské aplikace

8.přednáška

Architektura-aplikace



Proč data mining?

- Stále větší množství dat uložených v databázích
 - Neustále generujeme data
 - Obchodní a bankovní transakce
 - Biologická, astronomická data atd...
 - Ukládáme stále více dat
 - Databázové technologie jsou stále rychlejší a levnější
 - Databázové systémy jsou schopny pracovat se stále rozsáhlejšími daty

Proč data mining?

- Data jsou stále rozsáhlejší, ale vyvodit z nich užitečné závěry je stále složitější
 - Velké množství nákupů v supermarketech
 - Miliony různých spojení (denně) u telekomunikačních operátorů
 - ...

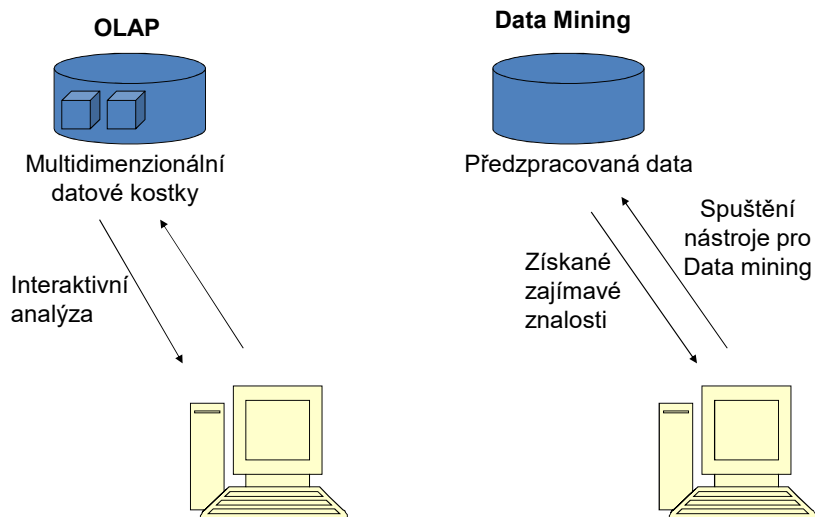
Co je to data mining?

- Zavedení pojmu: 1991 – Frawley
- Definice
 - Netriviální proces identifikace
 - nových,
 - platných,
 - potenciálně použitelných
 - a snadno pochopitelných vzorů v datech
- Zahrnuje poznatky z několika oborů matematiky a informatiky

OLAP vs. Data Mining

- Data Mining
 - Hledání zcela nových vzorů, znalostí, které v datech nejsou explicitně uvedeny
 - Znalostí je dosahováno pomocí sofistikovaných algoritmů
- OLAP
 - Soubor operací (drill-down, roll-up...) poskytující různé pohledy na data
 - Výsledků je dosahováno pomocí sumačních a předdefinovaných operací

OLAP vs. Data Mining



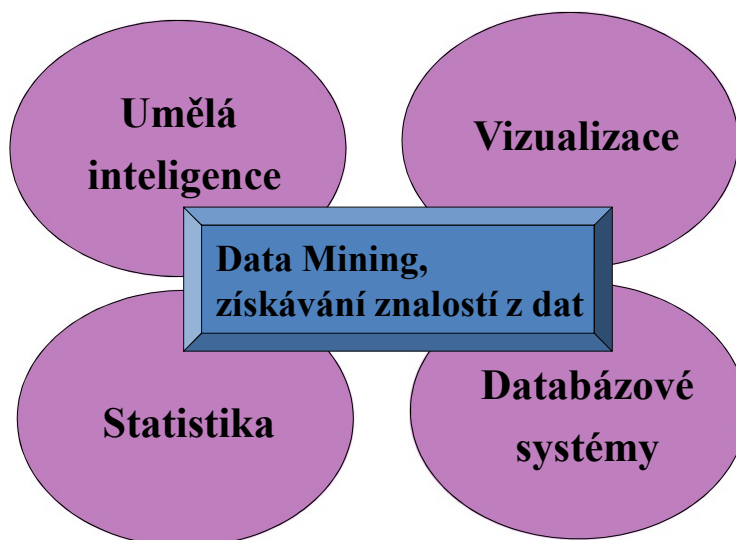
OLAP vs. Data Mining

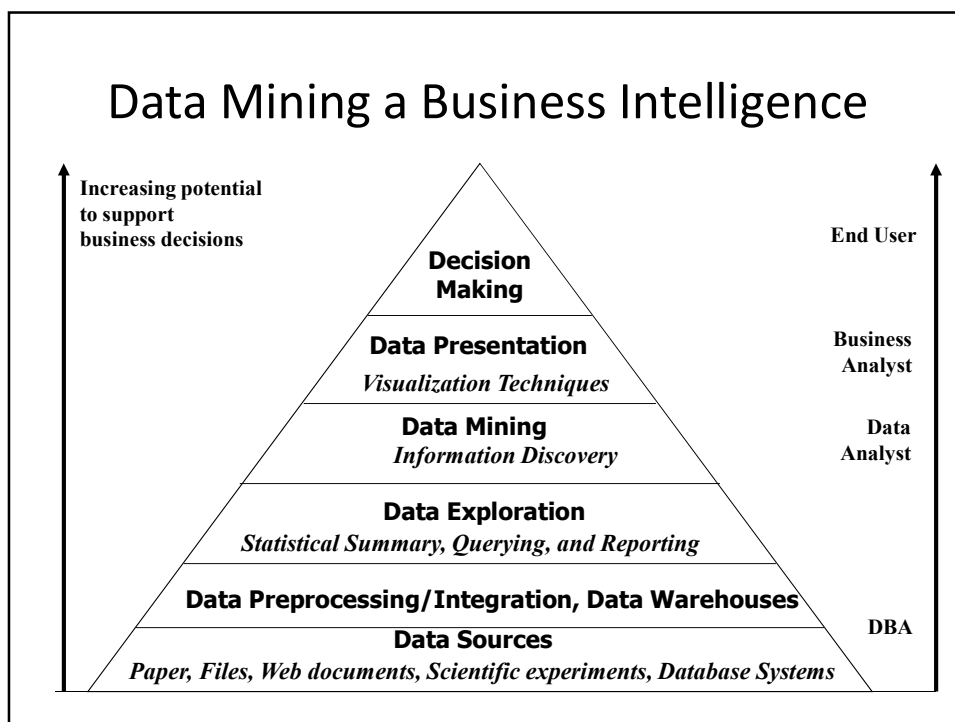
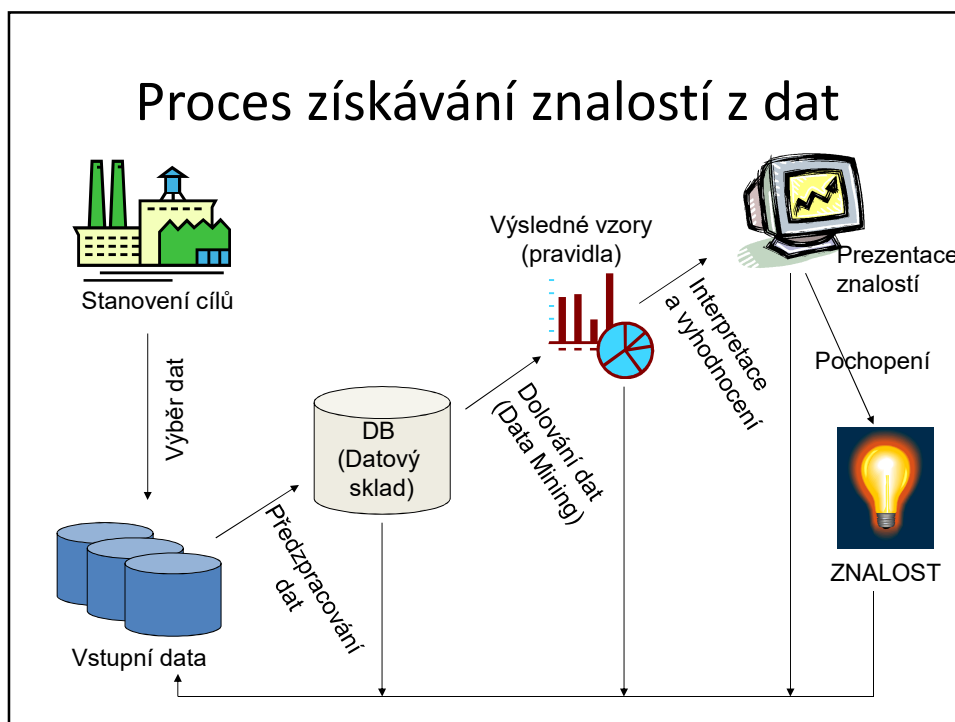
Vlastnost	OLAP	Data Mining
Motivace použití	Co se děje v podniku?	Predikce budoucnosti, skryté znalosti
Granularita dat	Sumační data	Data na úrovni záznamu
Počet obchodních dimenzí	Omezený počet dimenzí	Velký (až nekonečný) počet dimenzí
Počet vstupních atributů	Spíše velmi nízký počet atributů	Mnoho atributů
Velikost dat pro jednu dimenzi	Ne velká pro každou dimenzi	Obvykle velmi rozsáhlá pro každou dimenzi

OLAP vs. Data Mining

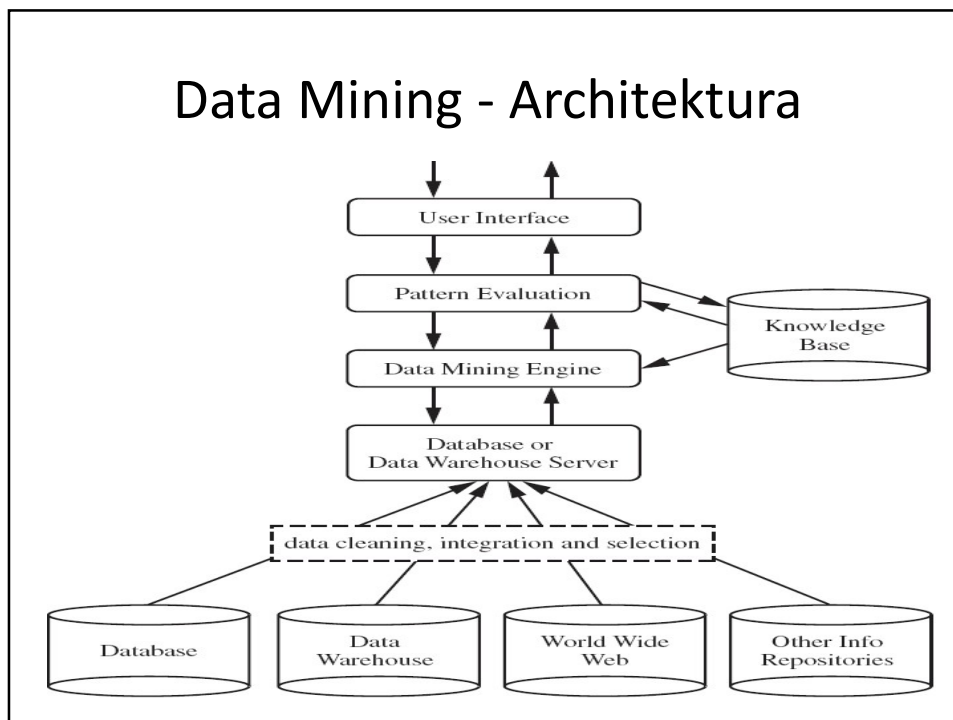
Vlastnost	OLAP	Data Mining
Přístup k analýze	Řízený uživatelem, interaktivní analýza	Automatický, řízený daty
Techniky analýzy	Multidimenzionální, drill-down, slice-and-dice	Příprava dat, použití nástrojů pro získávání znalostí
Stav technologie	Známý a rozsáhle využívaný	Stále se vyvíjející, některé metody jsou již využívány v praxi

Související obory





Data Mining - Architektura



Proces získávání znalostí z dat

- Stanovení cílů
 - Jaký typ znalosti chceme nalézt?
 - Nad jakými daty budeme proces získávání znalostí provádět?
 - Je problém řešitelný?
 - Budou získané výsledky užitečné v praxi?
 - V jakém tvaru a formě chceme výsledky získávání znalostí zobrazit?
 - Jsou naše data vhodná pro danou dolovací metodu vhodná?

Proces získávání znalostí z dat

- Výběr zdrojů dat
 - Typy dat pro data mining z hlediska zaměření
 - Demografická data (charakteristika osob - pohlaví, věk, vzdělání) – jsou levná, ale často neúplná
 - Behaviorální data (nákupy, prodeje atd.) – jsou dražší, ale z hlediska data miningu nejcennější
 - Psychografická data (typicky získaná průzkumem veřejného mínění) pomáhají při analýze chování zákazníka

Proces získávání znalostí z dat

- Výběr zdrojů dat
 - Typy databází z hlediska obsahu
 - Zákaznické databáze – údaje o zákazníka, případně o jeho aktivitách
 - Transakční databáze – údaje o aktivitách zákazníků (většinou anonymních)
 - Databáze historie nabídek – databáze o oslovování zákazníků kampaněmi
 - Datový sklad
 - Externí data

Proces získávání znalostí z dat

- Výběr zdrojů dat
 - Typy dat z hlediska formátu
 - Relační a transakční databáze
 - Objektově-orientované databáze
 - Multimediální databáze
 - WWW
 - Textové dokumenty
 - Prostorová, časová data...

Předzpracování dat

- Proč předzpracování?
 - Objemné databáze – je potřeba vybrat relevantní data
 - Nesprávná, nekonzistentní data, chybějící hodnoty
 - Zvýší efektivitu a usnadní proces získávání znalostí

Předzpracování dat – čištění dat

- Položky obsahující neúplné hodnoty
 - Zanedbání záznamu, doplnění průměrnou hodnotou nebo konstantou „unknown“, ruční zadání, predikce
- Položky obsahující chybné hodnoty
 - Binding – vyhlazení na základě sousedních hodnot
 - Shlukování – podobné hodnoty jsou organizovány do skupin, ostatní jsou chybné
 - Regresní metody
 - Kombinace lidské a počítačové kontroly

Předzpracování dat – čištění dat

- Nekonzistentní data
 - Vznikají při vkládání dat do databáze
 - Při integraci dat (např. různé názvy atributů)
- Řešení
 - Ruční opravení
 - Opravné rutiny

Předzpracování dat – integrace dat

- Integrace více zdrojů do jedné databáze
 - Redundance
 - Jak určit ekvivalentní entity z více zdrojů?
 - Detekce a řešení konfliktů hodnot atributů
 - např. různé kódování, měrné jednotky nebo různé vyjádření hodnoty

Předzpracování dat – transformace dat

- Transformace dat do formátu vhodného pro
dolování dat
 - Slučující techniky
 - Sumační operace atd... (z více hodnot jedna hodnota)
 - Generalizace
 - Data nižší úrovně nahrazena úrovní vyšší (např. ulice – město)
 - Normalizace
 - Přepočítání hodnot do daného intervalu

Předzpracování dat – transformace dat

- Přidávání nových atributů (odvozených)
- Diskretizace hodnot numerických atributů
 - Rozdělení numerických hodnot na intervaly
 - Ekvidistantní
 - Do hloubky
 - Pokročilé metody

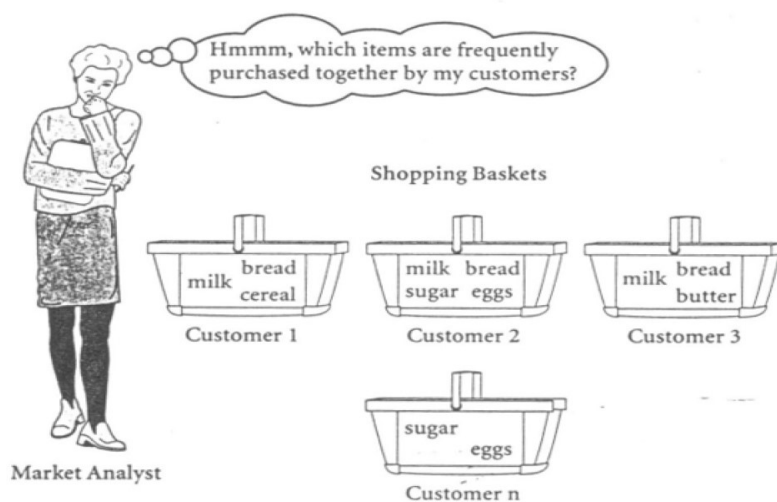
Předzpracování dat – redukce dat

- Agregace v kostce
 - Redukce dat sumačními operacemi
- Redukce rozměrů
 - Nadbytečné a nepoužívané atributy jsou detekovány a odstraněny
- Komprese dat
 - Zmenšení objemu dat
- Numerosity
 - Data jsou nahrazena alternativní menší reprezentací

Dolování dat

- Aplikace zvoleného algoritmu na předzpracovaná data, dle typu znalosti a dat
- Typy znalostí
 - Asociační pravidla
 - Shlukování
 - Klasifikace
 - Predikce

Asociační pravidla – analýza nákupního košíku



Asociační pravidla

- Původně pro transakční data
- Pravidlo ve tvaru $A \Rightarrow B$
 - A, B ... množiny položek
 - s ... podpora
 - c ... spolehlivost
- Interpretace asociačního pravidla:
 - „Jestliže transakce obsahuje položky z množiny A, pak také pravděpodobně obsahuje položky z B“

Asociační pravidla – základní pojmy

- Zajímavost pravidla $A \Rightarrow B$ určují tyto ukazatele:
 - **podpora (support)** - pravděpodobnost, že se vyskytnou v databázi položky z obou stran asociačního pravidla
 - **spolehlivost (confidence)** - podmíněná pravděpodobnost, že se vyskytne množina položek B, za předpokladu, že se tam vyskytnou položky z A

Asociační pravidla – základní pojmy

- Pravidlo, které má podporu a spolehlivost vyšší než je uživatelem zadaná hodnota, nazveme *silné asociační pravidlo*.
- Množina položek, která má podporu vyšší než minimální hodnota, se nazývá *frekventovaná množina*.

Asociační pravidla – základní postup

- Výpočet frekventovaných množin
 - na základě minimální podpory
 - časově náročnější krok
- Generování silných asociačních pravidel z frekventovaných množin
 - na základě minimální spolehlivosti

Základní algoritmus - Apriori

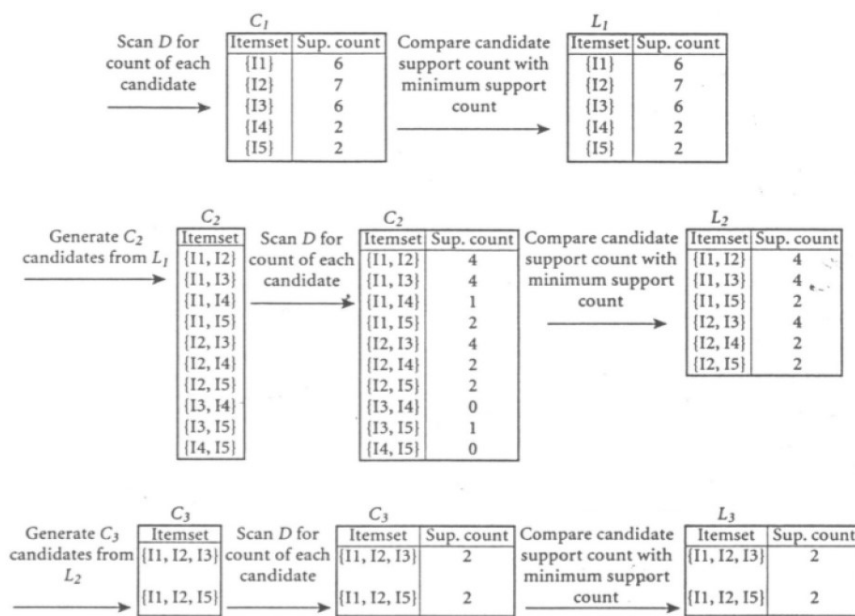
- Založeno na postupném generování kandidátů na frekventované položky
- Začíná se u množin o velikosti 1, postupně se generují množiny větší.
 - **spojovací fáze:** spojují se dvě stejně velké množiny, které se liší pouze v jednom prvku
 - **vylučovací fáze:** vylučují se ty množiny, jejichž některá podmnožina není frekventovaná

Základní algoritmus - Apriori

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

- Vstup: Transakční databáze
- Výstup: Frekventované množiny
- Apriori vlastnost: Podpora k-množiny nemůže být vyšší než podpora její podmnožiny, tj. frekventovaná (k+1)-množina může vzniknout pouze z frekventované k-množiny

Základní algoritmus - Apriori



Generování asociačních pravidel z frekventovaných množin

- Založeno na výpočtu spolehlivosti

$$confidence(A \Rightarrow B) = P(B|A) = \frac{support(A \cup B)}{support(A)}$$

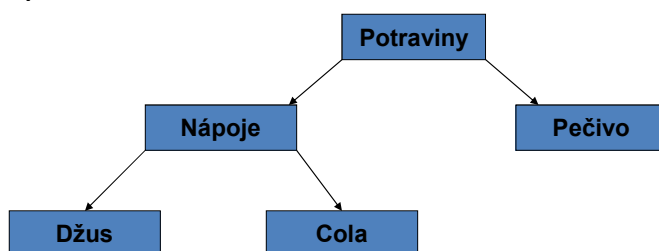
- Pro každou frekventovanou množinu se zjistí všechny její podmnožiny
- Pro každou podmnožinu s frekventované množiny se generuje

$$s \Rightarrow (l - s)$$

- Kontrola minimální spolehlivosti

Víceúrovňová asociační pravidla

- Důvod: málo silných asociačních pravidel
- Položky se sdružují do skupin (konceptů), musí být definována tzv. konceptuální hierarchie položek



Asociační pravidla v relačních databázích

- Kategorické atributy
 - Mají konečný počet hodnot
 - Lze na ně použít známé modifikované metody pro transakční data, např. algoritmus Apriori
- Kvantitativní atributy
 - Nemají konečný počet hodnot
 - Nutnost diskretizace – základní problém asociačních pravidel v relačních datech

Metody diskretizace kvantitativních atributů

- Základní metody
- Pokročilé metody
 - Postupné spojování menších intervalů ve větší
 - Shlukovací metody – jsou hledány shluky hodnot ležící blízko sebe, ty pak vytvoří interval
- Diskretizovaný atribut už lze považovat za kategorický a lze použít některou z metod

Sekvenční vzory

- Podobné jako frekventované množiny, ale hraje zde důležitou roli čas
- Příklad: Koupí-li si zákazník notebook, pak si později koupí také mobil.
 - Odpovídá to sekvenčnímu vzoru („notebook“, „mobil“)
- Důležité je tedy pořadí položek sekvenčního vzoru

Sekvenční vzory - příklad

SALE DATE	NAME OF CUSTOMER	PRODUCTS PURCHASED	Transaction Data File
Nov. 15, 2000	John Brown	Desktop PC, MP3 Player	
Nov. 15, 2000	Cindy Silverman	Desktop PC, MP3 Player, Digital Camera	Sequential Patterns -- Customer Sequence
Nov. 15, 2000	Robert Stone	Laptop PC	
Dec. 19, 2000	Terry Goldsmith	Laptop PC	Sequential Pattern Discovery with Support Factors
Dec. 19, 2000	John Brown	Digital Camera	
Dec. 19, 2000	Terry Goldsmith	Digital Camera	Sequential Pattern Discovery with Support Factors
Dec. 19, 2000	Robert Stone	Digital Camera	
Dec. 20, 2000	Cindy Silverman	Tape Backup Drive	Sequential Pattern Discovery with Support Factors
Dec. 20, 2000	Richard McKeown	Desktop PC, MP3 Player	

NAME OF CUSTOMER	PRODUCT SEQUENCE FOR CUSTOMER	Sequential Pattern Discovery with Support Factors
John Brown	Desktop PC, MP3 Player, Digital Camera	
Cindy Silverman	Desktop PC, MP3 Player, Digital Camera, Tape Backup Drive	Sequential Pattern Discovery with Support Factors
Robert Stone	Laptop PC, Digital Camera	
Terry Goldsmith	Laptop PC, Digital Camera	Sequential Pattern Discovery with Support Factors
Richard McKeown	Desktop PC, MP3 Player	

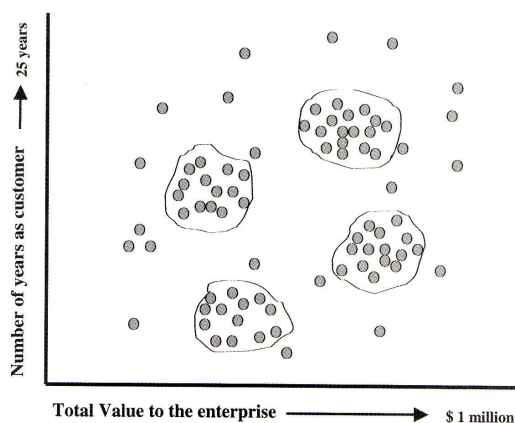
SEQUENTIAL PATTERNS (Support Factor > 60%)	SUPPORTING CUSTOMERS
Desktop PC, MP3 Player	John Brown, Cindy Silverman, Richard McKeown

SEQUENTIAL PATTERNS (Support Factor > 40%)	SUPPORTING CUSTOMERS
Desktop PC, MP3 Player, Digital Camera	John Brown, Cindy Silverman
Laptop PC, Digital Camera	Robert Stone, Terry Goldsmith

Shlukování

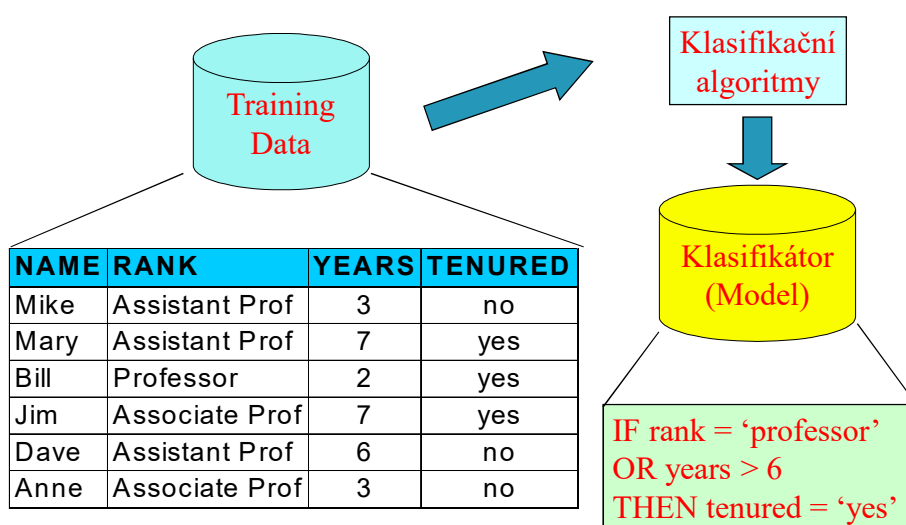
- Nejstarší nástroje data miningu
- Roztřídění skupiny objektů do skupin (shluků), které nejsou předem stanoveny
- Rozdíly objektů uvnitř shluků musí být minimální, rozdíly jednotlivých shluků musí být maximální
- Problém: Jakou metriku použít pro měření rozdílu?

Shlukování - ilustrace



- Např. je možné nyní oslovit kampaní skupinu zákazníků tvořících shluk

Klasifikace



Shlukování - vlastnosti

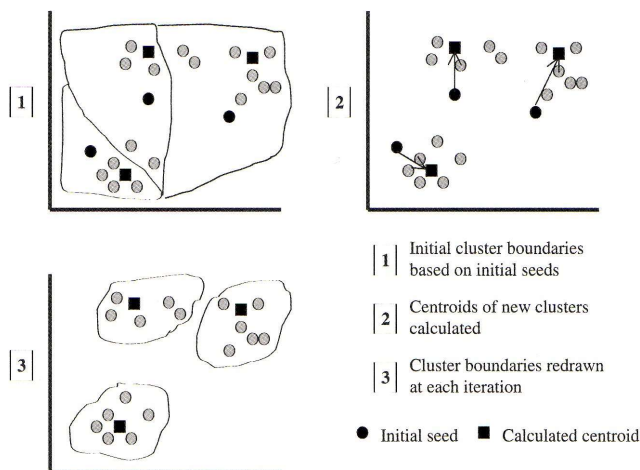
- Shluky nejsou předem dány a nemají tedy význam – ten je potřeba zjistit – ne vždy se to podaří
- Při 2-3 attributech je možné použít jednoduché metody, pro více atributů je potřeba použít pokročilé metody

Shlukování – některé metody

- Rozdělovací metody
 - Rozdělení objektů na předem daný počet shluků
 - Např. algoritmus K-means, který optimalizuje těžiště jednotlivých shluků a dané prvky pak přiřadí k nejbližšímu těžišti
 - V každé iteraci se počítají vzdálenosti prvků od těžiště. Tato hodnota musí pro každý shluk (těžiště) minimální

Shlukování – některé metody

- Ukázka použití algoritmu K-means



Shlukování – některé metody

- Hierarchické metody
 - Postupné rozdělování velkých shluků nebo postupné slučování malých shluků
 - Vzniká tím hierarchická struktura shluků
 - Ukončení procesu rozdělování (slučování) při splnění určité podmínky (např. určitý minimální počet shluků)
- Další metody (neuronové sítě, mřížky apod.)

Shlukování – příklady aplikací

- Marketing – možnost identifikace skupin zákazníků, použití cílených reklam
- Plánování města – identifikace skupin domů na základě typu, ceny a polohy
- Studie zemětřesení – shlukování epicenter zemětřesení dle jejich vlastností
- Pojištění – hledání potenciálních zákazníků s vysokým povinným ručením
- Geografie – hledání shluků pozemků na základě jeho typu

Klasifikace

- Rozdělování objektů do předem známých skupin
- Nejčastěji se využívají rozhodovací stromy
 - 1. krok: konstrukce rozhodovacího stromu na základě vzorku dat
 - 2. krok: klasifikace objektů na základě vytvořeného rozhodovacího stromu
- Úspěšnost se měří procentem úspěšně klasifikovaných objektů

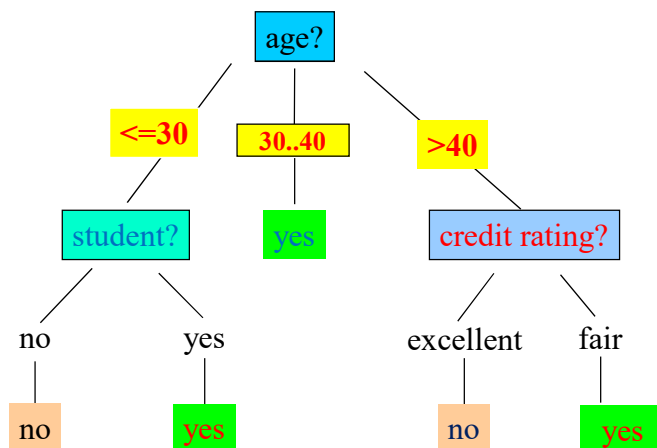
Klasifikátory (modely)

- Pravidla
 - Ve tvaru: if (podmínka atributu) then result = ...
 - Lze je převést na rozhodovací strom
- Rozhodovací stromy
 - Vnitřní uzel – test hodnoty jistého atributu
 - Koncový uzel – třída, do které je objekt klasifikován
- Neuronové sítě

Klasifikace – příklady

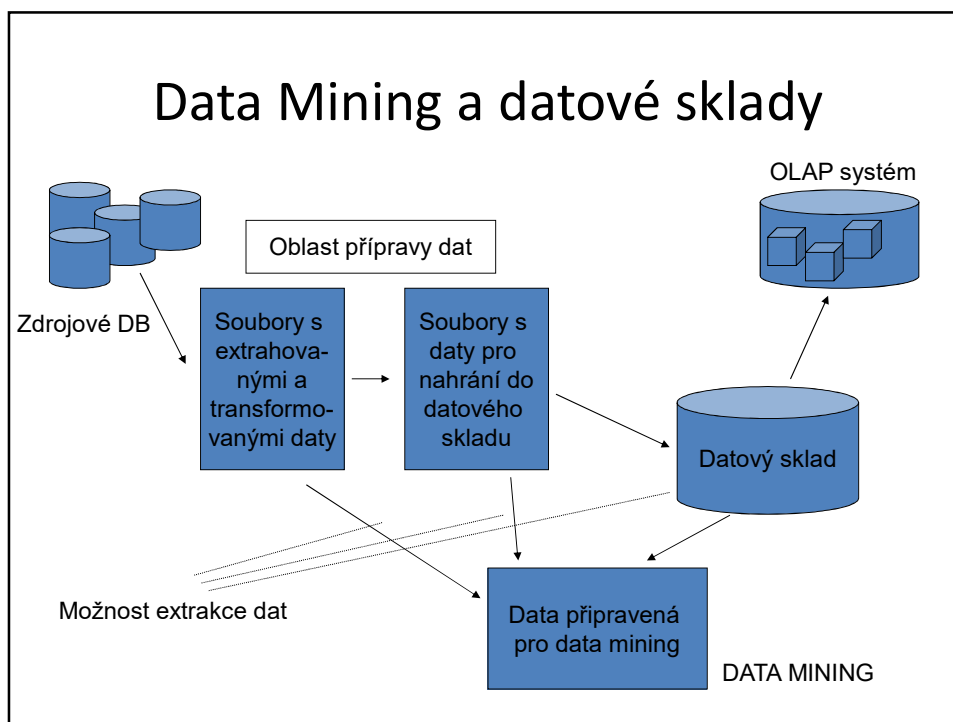
- Určení, zda je možné zákazníkovi možné poskytnout úvěr na základě několika atributů (věk, příjem...)
- Určení pohlaví zákazníka na základě toho, jaký notebook si koupí – to např. umožňuje směřovat kampaň...

Klasifikace – příklad (buys_computer)



Data Mining a datové sklady

- Data Mining hraje důležitou roli v prostředí datového skladu.
- Společné znaky
 - Velké množství dat, většinou na detailní úrovni – ale ne vždy jsou tam všechna data
 - Data Mining nejlépe pracuje s integrovanými a vyčištěnými daty
 - Máme-li datový sklad, není potřeba investovat do HW pro data mining



Využití data miningu

- Členění (segmentace) zákazníků
 - Cíl: porozumět zákazníkovi a jeho chování
- Analýza nákupního košíku
 - Nalezení závislostí mezi různým zbožím, které si zákazník koupí
- Management rizik
 - Odhalení rizikových zákazníků (např. u pojišťoven)
- Detekce podvodů
 - Např. hledání extrémních útrat na kreditní kartě
- Odhalování zločinnosti
 - Odhalení potenciálních neplatičů půjček...
- Predikce požadavků
 - Předpověď zájmu zákazníků o různé zboží...

Dotazovací jazyky pro data mining

- Data mining by měl být interaktivním procesem
- Základ pro uživatelské rozhraní
- Standardizace
- Součásti dotazu pro data mining
 - Relevantní data
 - Typ znalosti
 - Doménová znalost
 - Metriky zajímavosti
 - Vizualizace/prezentace získaných znalostí

Dotazovací jazyky pro data mining – součásti dotazu

- Relevantní data
 - Jméno databáze/datového skladu
 - Databázové tabulky/kostky
 - Podmínky pro selekci dat
 - Relevantní atributy nebo dimenze
 - Kritéria pro seskupování dat
- Typ získávané znalosti
 - Asociační pravidla, shlukování, klasifikace, ...

Dotazovací jazyky pro data mining – součástí dotazu

- Doménová znalost
 - Typické využití: Konceptuální hierarchie
 - Stromová hierarchie: město – kraj – země – světadíl
 - Seskupovací hierarchie: Např.: (15-39) – mladý; (40-59) – střední věk
 - Hierarchie založená na pravidlech: nízký_zisk(X) = cena(X) = p AND náklady(X) = q AND p-q < 50\$
 - Hierarchie odvozená z operace: emailová adresa: hagonzal@cs.uiuc.edu – login – ústav – univerzita – země

Dotazovací jazyky pro data mining – součástí dotazu

- Metriky zajímavosti
 - Jednoduchost – počet prvků pravidla, velikost rozhodovacího stromu
 - Použitelnost – např. podpora a spolehlivost
 - Jedinečnost – odstranění podobných znalostí
- Prezentace/Vizualizace
 - Různé formy reprezentace – grafy, tabulky...
 - Reprezentace konceptuální hierarchie
 - Vizualizace různých typů znalostí

Dotazovací jazyky pro data mining

Příklad dotazu jazyka DMQL

```
use database AllElectronics_db
use hierarchy location_hierarchy for T.branch, age_hierarchy for C.age
mine classification as promising_customers
in relevance to C.age, C.income, I.type, I.place_made, T.branch
from customer C, item I, transaction T
where I.item_ID = T.item_ID and C.cust_ID = T.cust_ID
      and C.income ≥ 40,000 and I.price ≥ 100
group by T.cust_ID
having sum(I.price) ≥ 1,000
display as rules
```

Získávání znalostí z komplexních dat

- Prostorové databáze
 - Nutnost předzpracování...
 - Příklad asociačního pravidla
- Multimediální databáze
 - Konstrukce vektoru rysů
 - Histogramy
 - Identifikace objektů v obrázku

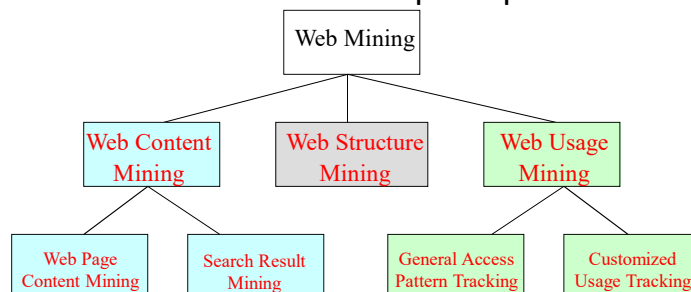
```
is_a(x, large_town) ^ intersect(x, highway) → adjacent_to(x, water)
```

Získávání znalostí z komplexních dat

- Časová a sekvenční data
 - Obsahují sekvence hodnot a událostí závislých na čase
 - Použití v meteorologii, lékařství (krevní tlak), burza (inlace, ceny akcií)
- Textové databáze
 - Velké kolekce dokumentů...
 - Hledání podobných kolekcí dokumentů obsahujících zadaná slova
 - Asociační pravidla založená na klíčových slovech
 - Klasifikace dokumentů

Získávání znalostí z komplexních dat

- World-Wide-Web
 - WWW dokumenty
 - Databáze s informacemi o přístupu...



Získávání znalostí z komplexních dat

- Objektové databáze
 - Lze použít upravené metody pro získávání znalostí z relačních dat
- XML

Dolování dat – Data Mining

SHRNUTÍ

Data Mining (DM)

1. **Deskriptivní model** – popisuje nalezené vzory a vztahy v datech, které mohou ovlivnit rozhodování (Př. Analýza prodeje zboží v supermarketu na jejímž základě je pak umístěno zboží v regálech).
2. **Prediktivní model** – umožňuje předvídat budoucí hodnoty atributů na základě nalezených vzorů v datech (Př. Analýza zákazníků, u kterých je vysoká pravděpodobnost, že budou reagovat na písemnou reklamní nabídku...)

Predikce

- Klasifikace
- Regresní analýza
- Analýza časových řad

Deskriptivní DM

- Zjišťování odchylek
- Segmentace
- Shlukování
- Asociační pravidla
- Sumarizace
- Vizualizace
- Dolování v textu

Kategorie úloh Data Mining

- **Klasifikace** – bude produkt úspěšný?
- **Regrese** – závislost mezi dvěma proměnnými
- **Shlukování** – rozdělení do množin dle společných znaků
- **Sumarizace**
- **Predikce podle časových řad** (autoregresní modely)
- **Modelování závislostí**
- **Asociace** – např. analýza nákupního koše
- **Analýza sekvencí** – např. procházení webu návštěvníkem
- **Analýza odchylek** – bankovní podvody

Metody DM

- **regresní metody** (lineární regresní analýza, nelineární regresní analýza, neuronové sítě)
- **klasifikace** (diskriminační analýza, logistická regresní analýza, rozhodovací stromy, neuronové sítě),
- **segmentace – shlukování** - shluková analýza, genetické algoritmy, neuronové shlukování (Kohonenovy mapy)
- **analýza vztahů** (asociační algoritmus pro odvozování pravidel typu „if X then Y“)
- **predikce v časových řadách** (Boxova-Jenkinsonova metoda, neuronové sítě, autoregresní modely, ARIMA)
- **detekce odchylek**

Lineární regrese

- Závislost mezi dvěma proměnnými
 - Nezávislá
 - Závislá (tu se snažíme predikovat)
- Je hledána přímka procházející mezi hodnotami tak, že součet druhých mocnin odchylek je minimální
- Tato přímka pak definuje vztah závislosti s možností určit nejpravděpodobnější hodnotu závislé proměnné

Lineární regrese

- Někdy je vztah nelineární -> nezávislou proměnnou je pak třeba transformovat
- Použití: modely pro marketing, odhad rizika

Logistická regrese

- Závislá proměnná není spojitá, ale diskrétní (kategorická)
- Užitečné např. v marketingu – předpověď odezvy na nějakou akci

Neuronové sítě

- Nevychází ze statistiky, ale řešení je modelováno na základě modelů funkcí lidského mozku
- Přijímání informací a poučení se z každé zkušenosti
- Užití v marketingu – např. predikce reakce na nabídku
- Neuronové sítě mají schopnost vystihnout i nelineární vztahy

Genetické algoritmy

- Modely, které se v každém kroku upravují, až se nalezne nejlepší model pro danou úlohu
- Určení cíle modelu, výběr míry pro vyhodnocení, jak model odpovídá našim požadavkům

Klasifikační stromy

- Účelem je roztržít data do skupin
- Oproti regresi i nelineární závislosti
- Identifikace segmentů s požadovaným chováním

DM - postup

- Specifikace problému
- Získání dat
- Výběr metody
- Předzpracování dat
- Vlastní data mining
- Interpretace

Fáze při dolování dat

1. Data Understanding – porozumění úloze
2. Data Preparation - příprava
3. Modelling - dolování
4. Evaluation - vyhodnocení
5. Deployment - nasazení

Příklady DM

- Množina pravidel, které určují jak lze seskupit produkty do skupin
- Rozhodovací strom, který predikuje, zda si zákazník koupí produkt
- Matematický model předpovídající prodej

Microsoft Analysis Service - DM

- Classification – Decision Tree Algorithm
- Regression – Microsoft Time Series Alg.
- Segmentation – Clustering Algorithm
- Association – hledá korelace
- Sequence analysis algorithm

Klasifikační algoritmy

- Předvídají jednu nebo více diskretních veličin na základě vstupních algoritmů

Regresní algoritmy

- na základě atributů předvídají jednu nebo více spojitých veličin, jako je například zisk či ztráta.

Segmentační algoritmy

- Rozdělují data do skupin s podobnými charakteristikami
- „trs“ dat
- Např. Analýza nákupního koše

Sekvenční algoritmy

- Hledají sekvence v datech
- Např. analýza informací o uživateli webu – snaha vytvořit skupiny uživatelů s podobnými nakupovacími návyky

Metodiky

- 5 A (SPASS, 1996)
- SEMMA (Enterprise Miner, SAS)
- CRISP-DM (2000, Chapman)
 - DIME
 - RAMSYS (2001)
- Two Crows
- Annad & Buchner

Metodika 5A

- Assess – posouzení potřeb
- Access – shromáždění dat
- Analyze
- Akt – přeměna na znalosti
- Automate – převedení výsledků analýzy do praxe

SEMMA

- Sample
- Explore
- Modify – datová transformace, seskupení
- Model – analýza dat
- Assess – porovnání modelů, interpretace

CRISP-DM

- Standard podporovaný většinou systémů
- Metodologie pro plánování data miningových projektů

Predictive modeling Markup Language

- Na bázi XML
- Data Mining Group
- www.dgm.org
- Popis dat, datových transformací a vytvořených modelů

API pro Data Mining

- OLE DB for Data Mining – Microsoft
- Java Data Mining

Systemy používané pro Data Mining

SPM	Salford Systems	www.salford-systems.com
Clementine	SPSS	www-01.ibm.com/software/analytics/spss/products/modeler/
Enterprise Miner	SAS Institute	www.sas.com/technologies/analytics/datamining/miner/
GhostMiner	Fujitsu	www.fqs.pl/business_intelligence/products/ghostminer
Intelligent Miner	IBM	www-01.ibm.com/software/data/infosphere/warehouse/enterprise.html
Knowledge Studio	Angoss	www.angoss.com
Oracle Data Mining	Oracle	www.oracle.com/us/products/database/options/data-mining/index.html
PolyAnalyst	Megaputer	www.megaputer.com/
Statistica Data Miner	StatSoft	www.statsoft.com/products/data-mining-solutions/
LISp Miner	VŠE	lispminer.vse.cz
RapidMiner	Rapid-I	rapid-i.com/
Weka	University of Waikato	www.cs.waikato.ac.nz/ml/weka/index.html

Zdroje dat

- <http://www.registry.cz/index.php?pg=nabizen-e-sluzby--analyza-dat>