**Česká zemědělská univerzita v Praze**

**Provozně ekonomická fakulta**

**Katedra informačních technologií**


# Bakalářská práce

# Data Mining

**Bohuslav Endt**

# ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

Katedra informačních technologií

Provozně ekonomická fakulta

# ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Bohuslav Endt

Informatika

Název práce

**Data Mining**

Název anglicky

**Data Mining**

---

**Cíle práce**

Bakalářská práce má za cíl zrekapitulovat informace z oblasti data miningu z databázových systémů a na základě získaných znalostí provést analýzu dat a sestavit predikční model použitím zvolené metody a softwaru pro data mining. Výsledky analýzy dat a predikce ověřit na testovacích datech, zhodnotit výsledky predikce a popsat
nejdůležitější části tohoto modelu.

**Metodika**

Na začátku práce bude provedeno studium literárních zdrojů, následně bude zpracována rešerše těchto zdrojů. Postupy užíváné pro DM objasní postup a princip algoritmů využívaných při dolování dat. Fáze úpravy dat se bude zabývat připravením dat pro zpracování zvoleným algoritmem, který bude, společně s vhodným softwarem pro zpracování, zvolen na základě nově získaných znalostí. Výsledky DM budou interpretovány běžným jazykem. Závěrem budou zesumarizovány nejdůležitější informace získané celým procesem DM.

**Doporučený rozsah práce**

30 – 40 stran

**Klíčová slova**

Datamining, Dolování dat, Získávání znalostí z databází, Informace, Databázové systémy

**Doporučené zdroje informací**

BERKA, Petr. Dobývání znalostí z databází. 1. vyd. Praha: Academia, 2003. 366s + Computer Press, 2003. 486str + 1CD ROM. ISBN 80-7226-969-0
LACKO, Luboslav. Databáze: datové sklady, OLAP a dolování dat. 1. vyd. Brno: 1CD ROM. ISBN 80-200-1062-9

**Předběžný termín obhajoby**

2015/06 (červen)

**Vedoucí práce**

Ing. Martin Havránek, Ph.D.

Elektronicky schváleno dne 31. 10. 2014

**Ing. Jiří Vaněk, Ph.D.**

Vedoucí katedry

Elektronicky schváleno dne 11. 11. 2014

**Ing. Martin Pelikán, Ph.D.**

Děkan

V Praze dne 15. 03. 2015

# Čestné prohlášení

Prohlašuji, že svou bakalářskou práci "Data Mining" jsem vypracoval samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou citovány v práci a uvedeny v seznamu literatury na konci práce. Jako autor uvedené bakalářské práce dále prohlašuji, že jsem v souvislosti s jejím vytvořením neporušil autorská práva třetích osob.

V Praze dne 16. 3. 2015 _____

## Poděkování

Rád bych touto cestou poděkoval Ing. Martinu Havránkovi, Ph.D. za trpělivost a rady při zpracování bakalářské práce.

# Data Mining

---

# Data Mining

**Souhrn**

Bakalářská práce se zabývá Data miningem. V teoretické části práce jsou na základě studia odborné literatury zrekapitulovány základní pojmy z oblasti data miningu a z databázových systémů. V práci jsou přiblíženy postupy užívané pro data mining a objasněny postupy a principy algoritmů využívaných při dolování dat.

Praktická část se zabývá přípravou dat pro dvě analytické metody data miningu a jejich následným zpracováním. V závěru jsou zesumarizovány výsledky celého procesu.

**Summary**

My Bachelor's thesis deals with data mining. In the theoretical part, based on research of expert literature, there are mentioned basic concepts from the area data mining and database systems. In the text there are described the procedures used for data mining and explained principles of algorithms used in data mining.

The practical part deals with data preparation for two analytical methods of data mining and its subsequent processing. In conclusion, the results of whole process are summarized.

**Klíčová slova:** Datamining, Dolování dat, Získávání znalostí z databází, Informace, Databázové systémy

**Keywords**: Datamining, Data mining, Knowledge discovery in databases, Information, Database systems

# Obsah:

# 1 Úvod

O Data Mining-u jako oboru se začalo ve vědeckých kruzích mluvit na počátku 90. let minulého století, na konferencích v Americe. Rozsah automaticky sbíraných dat začínal převyšovat kapacitu uživatelů na jejich zpracování. Definic Data Miningu (DM) je mnoho, mezi nejrozšířenější však patří tato, jež pronesl U. M. Fayyad v roce 1996: *„Data Mining je netriviální dobývání skrytých, předem neznámých a potencionálně užitečných informací z dat.“*

Mezi základní disciplíny oboru DM patří, kromě mnoha dalších, především databázové technologie, statistika, neuronové sítě, rozhodovací stromy a klasifikace. Většina těchto technik užívaných při DM byla popsána před samotným vznikem DM, avšak nebyly užívány společně. V současnosti je DM velmi rozšířený a bývá implementován ve všech komerčních databázových serverech a mnoho firem má vlastní řešení zabývající se průzkumem trhu nebo podporou rozhodování.
V dnešní době je DM velmi rozšířené, užívá se nejen ve vědeckých kruzích, bankovnictví či energetice, jako tomu bylo zpočátku 90. let, ale i k analýze trhů a predikci jejich chování.

# 2 Cíl práce a metodika

Bakalářská práce má za cíl zrekapitulovat informace z oblasti data miningu z databázových systémů a na základě získaných znalostí provést analýzu dat a sestavit predikční model použitím zvolené metody a softwaru pro data mining. Výsledky analýzy dat a predikce ověřit na testovacích datech, zhodnotit výsledky predikce a popsat nejdůležitější části tohoto modelu.

Na začátku práce bude provedeno studium literárních zdrojů, následně bude zpracována rešerše těchto zdrojů. Postupy užívané pro DM objasní postup a princip algoritmů využívaných při dolování dat. Fáze úpravy dat se bude zabývat připravením dat pro zpracování zvoleným algoritmem, který bude, společně s vhodným softwarem pro zpracování, zvolen na základě nově získaných znalostí. Výsledky DM budou interpretovány běžným jazykem. Závěrem budou zesumarizovány nejdůležitější informace získané celým procesem DM.

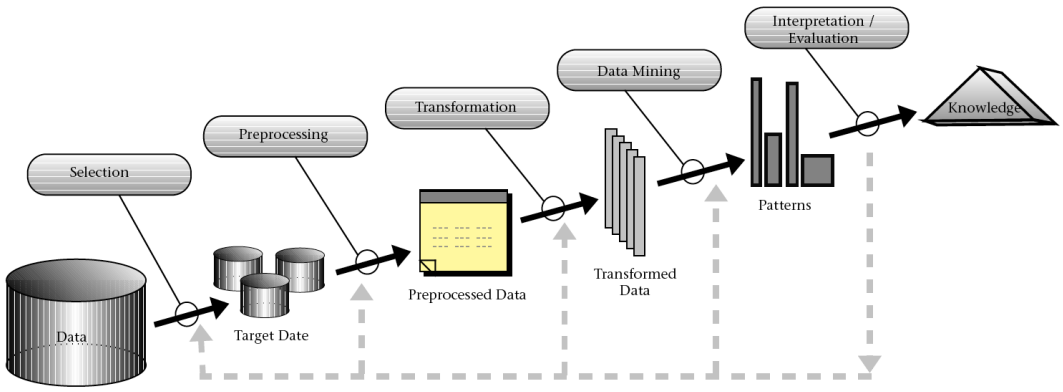
# 3 Přehled problematiky

S rostoucím počtem uživatelů moderních technologií a chápáním důležitosti DM jako konkurenční výhodou vzniká každý den více a více dat. Tato data, uložené v databázových systémech, jsou samy o sobě takřka bezcenné. Důležité informace však lze „vytěžit“. DM vyhledává vzory v datech nasbíraných a uložených v databázových systémech.

*„Data mining je proces analýzy dat z různých perspektiv a jejich přeměna na užitečné informace. Z matematického a statistického hlediska jde o hledání korelací, tedy vzájemných vztahů nebo vzorů v datech. Data mining je proces, jehož cílem je těžba informací v databázích. Využívá statistické metody a další metody hraničící s oblastí umělé inteligence.“(*Lacko, 2009)

## 3.1 Metodologie

V uplynulých letech vzniklo několik metodik, které mají za cíl poskytnout uživatelům jednotný proces pro řešení různých úloh z oblasti DM. Tyto metodiky jsou dále předávat zkušenosti z projektů. Některé byly produkovány producenty programových systémů (metodika 5A firmy SPSS nebo metodika SEMMA formy SAS), jiné vznikly jako „softwarově nezávislé“ spoluprácí komerčních a výzkumných institucí (CRISP-DM) zpracováno podle (Berka, 2001).




**Obrázek 1: Proces DM dle Fayyad 1996**

**zdroj http://www.infovis-wiki.net/images/4/4d/Fayyad96kdd-process.png**

### 3.1.1 Metodika 5A

Metodika užitá v produktu Clementine je pohledem firmy SPSS na celý proces DM. Název metodiky 5A vznikl jako akronym pro názvy jednotlivých kroků procesu:

- Assess – posouzení potřeb projektu,
- Access – shromáždění potřebných dat,
- Analyze – provedení analýz,
- Act – přeměna znalostí na akční znalosti,
- Automate – převedení výsledků analýzy do praxe.

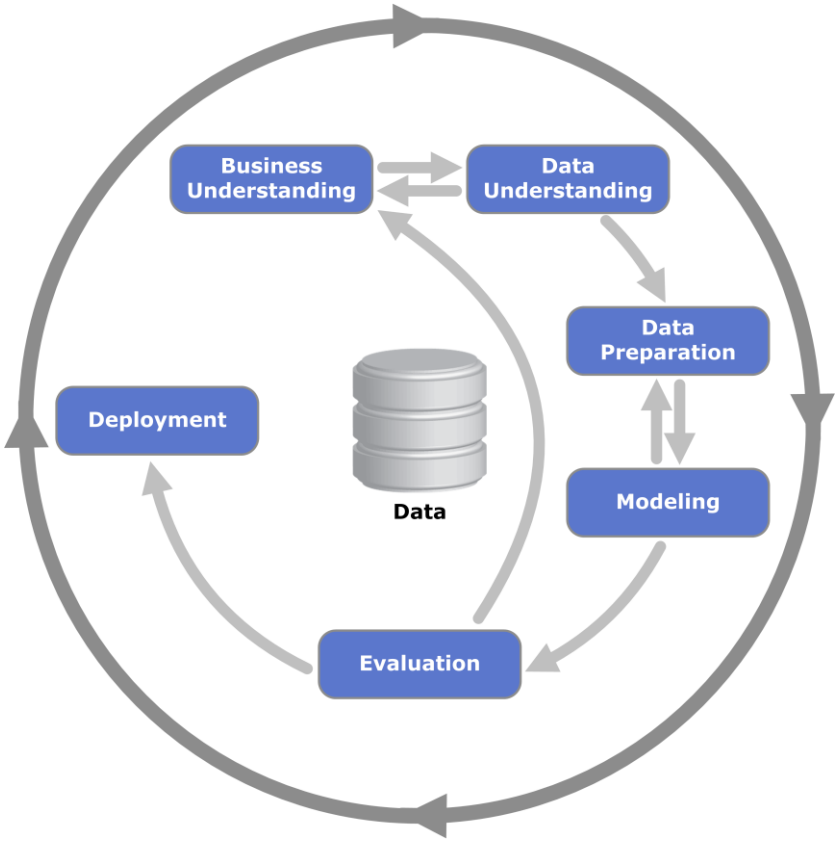
### 3.1.2 Metodika SEMMA

Metodiku SEMMA využívá produkt Enterprise Miner firmy SAS. Je zde kladen důraz na snadnou interpretaci výstupů ve formě přístupné a srozumitelné obchodnímu uživateli. Její název je opět akronymem jednotlivých kroků:

- Sample – vybírání vhodného objektu,
- Explore – vizuální explorace a redukce dat,
- Modify – seskupování objektů a hodnot atributů, datové transformace,
- Model – analýza dat, užití jednotlivých metod strojového učení
- Assess – porovnání modelů a interpretace

### 3.1.3 Metodika CRISP-DM

CRISP-DM (CRoss-Industry Standard Process for Data Mining) vznikla v rámci Evropského výzkumného projektu, jehož cílem bylo navrhnout univerzální postup , který bude možno použít ve vícero komerčních aplikacích. Nabízí soubor návodů, úkolů a cílů pro každou část procesu. Další verze, CRISP-DM 2.0, byla ohlášena roku 2006, ale velice pravděpodobně byl její vývoj ukončen.
CRISP-DM rozděluje životní cyklus projektu DM do šesti etap.




**Obrázek 2: CRISP-DM Porozumění problematice – Business Understanding**

**zdroj http://upload.wikimedia.org/wikipedia/commons/b/b9/CRISP-DM_Process_Diagram.png**

Úvodní fáze se zabývá pochopením cíle projektu a manažery předpokládaných vlastností řešení. Tyto poznatky se následně převádějí do zadání úlohy pro DM. Jsou brány v potaz možná rizika, přínos metody DM a celkové náklady. Stanovuje se předběžný plán prací a je prováděna inventura zdrojů.

#### 3.1.3.1 Porozumění datům – Data Understanding

V této fázi dochází ke sběru dat, k hodnocení kvality dat a k získání jejich základní charakteristiky.

#### 3.1.3.2 Příprava dat – Data Preparation

Jedná se zpravidla o nejnáročnější fázi projektu. Obsahuje všechny činnosti, vedoucí k vytvoření datového souboru, který bude nadále zpracován analytickými metodami. Tento soubor musí obsahovat údaje význačné pro danou úlohu a musí mít podobu, vyžadovanou danými analytickými algoritmy.

Příprava dat tedy zahrnuje selekci dat, čištění dat, transformaci dat, vytváření dat, integrování dat a formátování dat. Jednotlivé tyto úlohy jsou prováděny opakovaně a v různém pořadí.

#### 3.1.3.3 Modelování – Modeling

V této fázi se užívá vhodný algoritmus či vícero algoritmů, který je k řešení dané problematiky nejvhodnější. V případě využití více algoritmů je třeba dbát na to, aby byla data řádně připravena, či je dodatečně připravit. Výsledkem je jeden či více modelů.

#### 3.1.3.4 Zhodnocení výsledků – Evaluation

Hotový model či modely, analyticky hodnotný, je třeba vyhodnotit z manažerského hlediska, zda a jakým způsobem byly cíle zadané na počátku splněny. Na konci této fáze je rozhodnuto o způsobu využití dosažených výsledků.

#### 3.1.3.5 Využití výsledků – Deployment

Vytvoření modelu není vždy konečnou fází projektu. Ve většině případů je třeba upravit získané znalosti do formy využitelné zákazníkem (zadavatelem úlohy). Může se jednat o zpracování závěrečné zprávy nebo o zavedení systému pro automatickou klasifikaci nových případů. V mnoha případech bude zákazník sám využívat model v praxi, a tak je pro něj důležité pochopit, jaké kroky je třeba učinit, aby mohl výsledky efektivně využít. Zpracováno dle (Chapman a kol., 2000).

## 3.2 Zdroje dat

Data musí být uložené v předem definované struktuře, aby bylo zajištěno jejich zpětné získání. Na začátku databázových technologií byly užívány tzv. flat soubory. Tyto soubory byly indexovány na základě předpokládaných způsobů dotazování. Jejich nevýhodou byla velká redundance dat a jejich nízká variabilita.

Mezi nejpoužívanější zdroje dat v dnešní době patří relační databáze, datové sklady a OLAP krychle. Tyto zdroje jsou přítomny ve většině komerčních nástrojů.

### 3.2.1 Relační databáze

Oproti flat souborům jsou v relačních databázích data ukládána do tabulek propojených za pomoci identifikátorů. Operace s daty se provádí jazykem SQL, což bylo pro analytiky, kteří neuměli tento jazyk značným problémem a museli využívat programátory, kteří jejich dotazy přepisovali do SQL.

### 3.2.2 EIS – Executed Information System

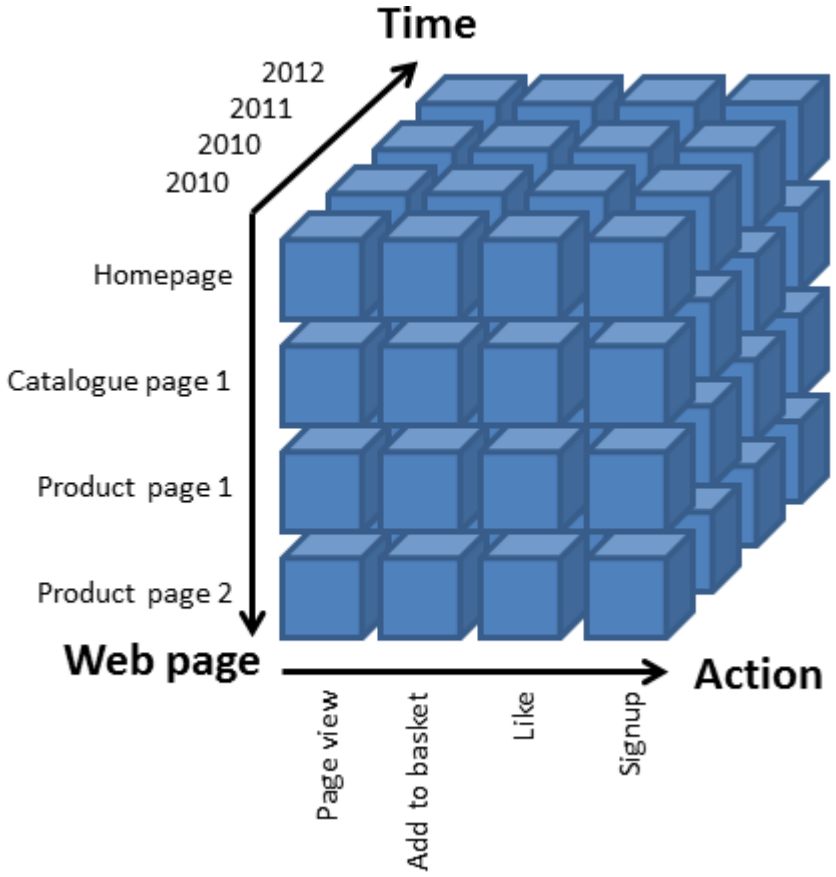
Systém vznikl na žádost uživatelů a obsahuje přívětivé interface. Užíval předdefinovanou množinu příkazů a zjednodušil tím uživatelům bez znalosti jazyka SQL přístup k datům. Programátora bylo potřeba jen v případě nutnosti rozšíření předdefinované množiny dotazů.

### 3.2.3 Krychle OLAP

OLAP (On-Line Analytical Processing) nabídl uživatelům výhody jak z relačních databází, tak z EIS; velkou flexibilitu a snadné, intuitivní ovládání. OLAP je vícedimenzionální uložení dat do datových krychlí. Jednotlivé atributy představují jednotlivé dimenze krychle a jednotlivé položky jsou buňkami krychle. Toto uložení dat umožňe rychlé provádění analytických operací jako jsou různá natáčení, provádění řezů, výběry částí a zobrazení agregovaných hodnot.

Mezi nevýhody OLAP krychlí patří zvýšená výpočetní náročnost, jelikož krychle obsahují nadbytečná data i prázdné buňky. Proto se užívají dva druhy OLAP – multidimenzionální OLAP (MOLAP) a relační OLAP (ROLAP). Případně lze využít i hybridní OLAP (HOLAP), což je kombinace výše zmíněných druhů. ROLAP převádí dotazy na SQL dotazy, jelikož jsou data skladována v relační databázi. MOLAP již pracuje s daty uloženými přímo v podobě krychlí.

### 3.2.4




**Obrázek 3: OLAP krychle**

**zdroj http://snowplowanalytics.com/assets/img/olap/example-cube.png**

**Datové sklady**

„*Datový sklad je v podstatě databáze uspořádaná podle trochu jiných pravidel. Tabulky nemusí být například normalizované. Datový sklad je soubor technologií pro efektivní skladování údajů tak, aby tyto údaje po jejich přeměně na informace sloužily podpoře rozhodování*.“ (Lacko, 2003)

Datové sklady zakládáme za účelem uchování nasbíraných dat, data ve skladu se neupravují, ani neodstraňují. V pravidelných časových intervalech se k nim však přidávají data nová, takzvané série snímků.

### 3.2.5 Datové trhy

Datové trhy jsou menší, na sobě nezávislé části datových skladů, většinou obsahují data použitelná konkrétním oddělením, nebo například jednou pobočkou dané firmy. Datové trhy mohou vznikat samostatně a postupně utvořit datový sklad, nebo naopak, mohou být tvořeny uměle z již existujícího datového skladu.

## 3.3 Statistika

„*Matematická statistika je věda, která buduje metody pro analýzu dat a využívá při tom princip statistické indukce. Její součástí je teorie odhadu, testování statistických hypotéz a statistická predikce.*“ (Budíková a kol., 2010)

### 3.3.1 Kontingenční tabulky

Kontingenční tabulky slouží k přehlednému shrnutí vztahů mezi vícero kvalitativními proměnnými.

**Tabulka 1: Ukázková kontingenční tabulka**

| | praváci | leváci | ∑ |
|---|---|---|---|
| muži | 43 | 9 | 52 |
| ženy | 44 | 4 | 48 |
| ∑ | 87 | 13 | 100 |

### 3.3.2 Regresivní analýza

Pomocí regresivní analýzy lze zjistit, zda existuje závislost mezi proměnou závislou a jednou a více proměnnými nezávislými, případně k zjištění síly dané závislosti a určení regresní funkce. Pomocí regresní funkce lze predikovat hodnoty závisle proměnných v závislosti na hodnotách jejich nezávislých proměnných, vypočítat koeficienty nezávisle proděných určit hodnotu náhodné veličiny. Regresní diagnostika určuje míru kvality dat, podle jejích výsledků lze upravit soubor a tím zvýšit míru závislosti.

### 3.3.3 Diskriminační analýza

Pomocí diskriminační analýzy lze rozdělit objekty do předem zadaných tříd. Tedy je to úloha hledání závislosti nominální veličiny na dalších numerických veličinách. Po určení závislosti mezi nominálními a dalšími veličinami slouží k rozdělení dalších případů do jednotlivých tříd.

### 3.3.4 Shluková analýza

Shluková analýza rozděluje data do shluků na základě jejich podobností. Data ve shluku by si měla být podobnější než data mimo něj. Podobnost určujeme na základě vzdálenosti dvou prvků. Vzdálenost lze vyjadřovat různými mírami, například Hammingovou vzdáleností, eukleidovskou vzdáleností a Čebyševovou. Metod pro shlukovou analýzu je vícero, například „metoda nejbližšího souseda“, kdy je vzdálenost určována vzdáleností dvou nejbližších objektů z různých shluků.

## 3.4 Strojové učení

Strojové učení je vědecký obor, který se zabývá algoritmy schopnými se učit z dat. Takové algoritmy utvářejí model na základě ukázkových vstupů a následně ho používají k rozhodování, či predikci. Strojové učení je úzce spjato se statistikou a matematickou optimalizací.

Hlavní užití strojového učení nacházíme tam, kde návrh a programování explicitních, na pravidlech založených algoritmů bývá nesmírně náročný. Mezi tyto příklady patří například filtrování spamu v elektronické poště, rozpoznávání znaků (OCR), některé online překladače, automatická detekce virů a vyhledávače. (Bishop, 2003)

Metody strojového učení lze dělit dle vynaloženého úsilí potřebného k získání nových znalostí:

1. Učení zapamatováním (rote learning) – systém zaznamenává data nebo znalosti
2. Učení se z instrukcí (learning from instruction) – systém provádí integraci se znalostmi již získanými
3. Učení se z analogie (learning by analogy) – systém si pamatuje již vyřešené příklady
4. Učení se na základě vysvětlení (explanation based learning) – systém užívá několik příkladů a rozsáhlé znalosti z dané oblasti
5. Učení se z příkladů (learning from examples) – systém využívá velké množství příkladů a indukci
6. Učení se pozorováním a objevováním (learning from observations and discovery) – pracuje s velkým množstvím dat, které jsou získány pozorováním či objevováním

Verifikace učení je velmi důležitým a obtížným prvkem celého procesu, existuje několik typů učení:

1. Učení s učitelem (supervised learning) – příklady jsou zařazené do tříd, systém se je postupně učí
2. Posilované učení (reinforcement learning) – odměny za správné chování a postihy za nesprávné, též známo jako zpětnovazební učení
3. Učení se napodobováním (apprenticeship learning) – učitel poskytuje nepřímé náznaky, systém z nich vychází
4. Učení bez učitele (unsupervised learning) – nejsou poskytovány doplňkové informace, systém hledá informace za pomoci metod DM

## 3.5 Data mining software

DM software můžeme dělit dle dvou kriterií, buď dle jejich typu, nebo dle licence pod kterou jsou jednotlivé programy distribuovány.

Dělení dle typu považuji za relevantnější. Statistický software, software pro podporu rozhodování s prvky DM, databázové s implementací business inteligence (BI) a aplikace pro DM.

Statistický software
- SPSS Modeler firmy IBM
- Statistica Data Miner firmy Statsoft
- SAS Enterprise miner pocházející od firmy SAS

Software pro podporu rozhodování
- SAP R/3 vyvíjený firmou SAP od roku 1992

Databázové servery s prvky BI
- IBM DB2
- Oracle Database 11g
- Microsoft SQL server

Aplikace pro DM
- Orange, Slovinský univerzitní projekt
- WEKA, Novozélandský univerzitní projekt
- RapidMiner, dříve YALE, jeden z nejrozšířenějších nástrojů
- LISp-Miner, DM program VŠE v Praze vyvíjený od roku 1996

# 4 Analytické data miningové metody

## 4.1 Rozhodovací stromy

Rozhodovací stromy jsou nejrozšířenějším typem vyobrazení znalostí. Jedná se zároveň o jeden z nejčastěji používaných algoritmů pro Data mining. Jsou srozumitelné a přehledné, díky tomu uživatel může jednoduše a velmi rychle vyhodnocovat výsledky a blíže zkoumat jednotlivé části případů. Algoritmus má za cíl identifikovat entity a rozdělit je do tříd. (Lacko, 2009)

„*Při tvorbě rozhodovacího stromu se postupuje metodou „rozděl a panuj“ (divide and conquer). Použití rozhodovacích stromů pro klasifikaci odpovídá analogii s klíči k určování rostlin nebo živočichů. Od kořene stromu se na základě odpovědí na otázky (umístěné v nelistových uzlech) postupuje příslušnou větví stále hlouběji, až do listového uzlu, který odpovídá zařazení příkladu do třídy*.“ (Berka, 2003)

Rozhodovací stromy jsou vhodné zejména pro úlohy klasifikace a predikce, ve kterých převažují kvalitativní data. Numerická data je třeba agregovat do konečného počtu skupin, aby obsahovaly v ideálním případě jen jednu hodnotu predikované veličiny. Tento proces provádí algoritmus automaticky. Díky hierarchické struktuře jsou algoritmy velmi rychlé a jednoduché. Algoritmus začíná u kořene stromu, a dále postupuje větvemi stromu až do koncového uzlu neboli listu stromu.

## 4.2 Rozhodovací pravidla

Rozhodovací pravidla jsou známy všem programátorům z jejich oblíbených programovacích jazyků, tak i ostatním lidem, jelikož jsou nedílnou součástí našeho běžného života, například: má-li pršet, vezmi si deštník. Obecná forma těchto pravidel zní: **IF (předpoklad)THEN (třída).**

Tato pravidla patří k velmi rozšířeným způsobům vyobrazení znalostí, podobně jako rozhodovací stromy. Pravidla lze tvořit několika algoritmy, například odvozením z rozhodovacích stromů. Oproti rozhodovacím stromům nabízejí rozhodovací pravidla lehčí interpretaci výsledků a při jejich vytváření lze postupovat oběma směry.

## 4.3 Asociační pravidla

Asociace je funkce Data miningu, která zjišťuje pravděpodobnost společného výskytu položek v datech. V asociačních pravidlech není žádný atribut vyčleněn jako cíl klasifikace, nýbrž se hledají všechny asociace mezi hodnotami různých atributů. (Berka, 2003)

Asociační pravidla se často užívají k analýze nákupního košíku k nalezení souvislostí mezi dvěma kupovanými produkty, a následném účelnějším rozmístění zboží v obchodě, či relevantnějším akčním nabídkám a letákům zasílaným zákazníkům. Většina informací zjištěná pomocí asociačních pravidel nepřináší žádný užitek, protože jsou to asociace známé.

## 4.4 Shlukování

Při analýze údajů na základě shluků se seskupují údaje podle podobných charakteristik. Shlukování se využívá pro identifikaci zákaznických segmentů, které jsou založené na společných charakteristikách například demografických, sociálních, profesních a podobně. Tento algoritmus se používá pro odhalování shluků dat ve vícedimenzionálních prostorech. Pomocí něj lze rozdělit množinu případů na co nejohraničenější skupiny takzvané „ostrovy podobnosti“.

Shluky jsou odhalovány na základě aplikování analýzy křivek pravděpodobnosti, přičemž se zkoumá, zda jednotlivé údaje nebo skupiny údajů nesplňují podmínku statistického rozložení například Gaussova normálního rozložení a podobně.
(Lacko, 2003)

## 4.5 Neuronové sítě

Neuronové sítě se hojně využívají v oboru umělé inteligence a jsou inspirovány lidským mozkem. Neuronové sítě jsou tvořeny navzájem propojenými elementy, neurony. Neuron se skládá ze vstupů, těla a výstupu. Jednotlivé vstupy jsou ohodnoceny vahou a každý neuron dále obsahuje prahovou hodnotu, které je potřeba dosáhnout aby se neuron aktivoval, a převodní funkci, která převádí vstupní hodnotu na výstupní hodnotu neuronu.

Neuronové sítě se skládají z vrstev uspořádaných neuronů. Vrstvy jsou děleny na vstupní, konečné a skryté. Data je pro učení nutno rozdělit na data učící a testovací. Při učení je síti postoupen vzor, pokud se výsledek liší od výsledku vzorového výsledku, je spočítána korekce a následně jsou upraveny hodnoty vah či prahů. Proces je opakován znovu s upraveným nastavením sítě, dokud není nedosaženo požadované maximální chyby.

Neuronové sítě jsou běžně užívány k predikci časových řad či klasifikaci. Dokáží predikovat číselné hodnoty či klasifikovat objekty na základě popisu. Nevýhodou pro uživatele může být, že oproti rozhodovacích stromů či pravidel, jsou zde znalosti rozprostřeny v podobě vah a jednotlivých vazeb mezi neurony, a tudíž není okamžitě zřejmé, jaké procesy uvnitř sítě probíhají.

## 4.6 Genetické algoritmy

Jsou po neuronových sítích druhým zástupcem algoritmů založených na bilogických principech. Vychází se z přirozené selekce přírody *„Genetické algoritmy nalezly uplatnění v řadě oblastí: numerická optimalizace a rozvrhování, strojové učení, tvorba modelů (ekonomických, populačních, sociálních), apod. Z hlediska dolování dat z databází je zajímavé využití genetických algoritmů přímo pro učení se konceptům, nebo použití genetických algoritmů pro optimalizaci neuronových sítí.“* (Berka, 2003) Jedinci v populaci odpovídají počtem a parametry neuronů, a kriteriální fit funkcí, která slouží k zjištění kvality jedinců.

Na počátku algoritmu je vygenerována náhodná populace jedinců a postupně se za použití mutace, reprodukce a selekce zlepšují další generace až k požadované hranici. Mutace jsou tvořeny náhodnou změnou, reprodukce představuje proces kombinace dvou jedinců, nebo klonování jednoho jedince. Kombinací těchto jedinců vzniká nový jedinec, nové generace. Selekce se užívá k výběru jedinců vhodných pro křížení.

## 4.7 Časové řady

Data určená k analýze jsou hodnoty určité veličiny v čase, například cena akcií, nebo nezaměstnanost. Dle analýzy minulých období a současnosti, lze sestavit pravidla, dle kterých lze predikovat vývoj veličiny. Model časových řad z pohledu statistického obsahuje 4 složky: trend, sezónní, cyklickou a nahodilou složku. Algoritmus analýzy časových řad vychází z tohoto dělení.

# 5 Příprava dat

Jedná se zpravidla o nejnáročnější fázi projektu. Obsahuje všechny činnosti, vedoucí k vytvoření datového souboru, který bude nadále zpracován analytickými metodami. Tento soubor musí obsahovat údaje význačné pro danou úlohu a musí mít podobu, vyžadovanou danými analytickými algoritmy.

## 5.1 Redukce počtu dimenzí

Redukcí počtu atributů dochází ke zvýšení rychlosti a efektivity, ale také ke snížení nákladů na tvorbu modelu.

Redukci počtu dimenzí lze provádět takzvanou selekcí, kdy některý atribut je možno úplně odstranit, nebo konverzí, kdy se užívá určitý algoritmus ke sloučení vícero atributů do jednoho, a původní atributy jsou odstraněny. Užíváním těchto operací jsou informace nenávratně ztráceny, je tudíž nutno jich užívat s rozmyslem.

### 5.1.1 Selekce

Nejprve je potřeba provést analýzu vstupních dat a selektovat duplicitní, či nepotřebné atributy. Nepotřebné atributy buď neobsahují relevantní informace, nebo obsahují žádné významné informace. K určení těchto atributů se užívá filtračních metod vycházejících z analýzy vlastností atributů.

V případě, že je nutno odstranit atribut, který obsahuje informaci důležitou k tvorbě modelu, je nutno stanovit akceptovatelnou chybu modelu a náklady na model.

### 5.1.2 Konverze

Opět je nutno nejprve provést analýzu vstupních dat, nezkoumají se však duplicity ale vztahy mezi jednotlivými atributy. Pokud je zjištěna korelace, atributy jsou sloučeny do nového atributu a původní jsou vymazány.

## 5.2 Chybná data

Pokud je hodnota některého z atributů chybná, může dojít k nesprávnému zařazení objektu do prostoru. Prostorem je míněn n-rozměrný prostor, ve kterém data miningové algoritmy zkoumají polohu objektu. Počet rozměrů prostoru je definován počtem atributů.

V případě, že některé hodnoty atributů chybí, jsou nahrazovány průměrem, mediánem nebo střední hodnotou ostatních hodnot. Případně mohou být nekompletní atributy smazány kompletně. (Berka, 2003)

# 6 Vlastní práce

Vlastní práce se zabývá analýzou dat pomocí algoritmů data miningu. Cílem je analýza dat generovaných návštěvníky webových stránek nejmenované cestovní kanceláře.

Pro tuto úlohu využívám software RapidMiner verze 5.3.015, který využívá bloků, které zastupují určitou fázi procesu. Tyto bloky, nazývané operátory, se skládají do blokového schématu. Je snadné s tímto programem pracovat a díky zdařilému grafickému prostředí je i intuitivní.

## 6.1 Analýza a úprava dat

Zdrojová data byla dodána ve třech souborech, přičemž největší tabulka obsahovala přes 38 000 řádků pro atributy: identifikátor události, identifikátor zobrazené stránky, identifikátor session, relativní uri navštívené stránky, jméno kategorie stránky (z navigace), identifikátor kategorie (z navigace), jméno kategorie stránky (z obsahu), identifikátor kategorie (z obsahu), název tématu, identifikátor tématu, čas strávený na stránce v sekundách, váha stránky odvozená dle času stráveném na stránce a pořadí stránky v clickstreamu. Rozhodl jsem se na datech provést shlukovou analýzu a vyhledat asociační pravidla.

Úprava dat probíhala převážně v Microsoft Excelu, za pomoci kontingenčních tabulek byly data pro clustering sloučeny následovně:

- názvy sloupců převzaty z atributu název tématu
- každý řádek odpovídá jednomu identifikátoru session
- v potaz byli bráni pouze uživatelé, kteří navštívili nejméně 3 stránky
- v práných buňkách byla hodnota null nahrazena hodnotou 0

Data pro asociační pravidla byly upraveny následovně:

- všechny číselné hodnoty byly nahrazeny slovní hodnotou (low, medium, high, very high)
- časový údaj nahrazen slovní hodnotou (morning, day, night)
- referer identifikátor nahrazen jeho názvem z jiné tabulky

## 6.2 Data mining

### 6.2.1 Asociační pravidla

Pro tvorbu asociačních pravidel jsem používal zejména bloky “FP-Growth”, “Create Association Rules” a “Normal to Binominal”.

Zde jsem pracoval s následujícími sloupci:
- Referer
- DobaNaStrance
- SumSkore
- PocetStranek
- Hodina

**Výsledek: Referer = fulltext**

| No. | Premises | Conclusion | Support | Confiden... | LaPlace | Gain | p-s | Lift | Convi... |
|---|---|---|---|---|---|---|---|---|---|
| 120 | Hodina = day, SumSkore = medium | DobaNaStrance = medium, Referer = Fulltext | 0.077 | 0.298 | 0.856 | -0.181 | 0.036 | 1.897 | 1.200 |
| 123 | SumSkore = medium | DobaNaStrance = medium, Referer = Fulltext | 0.109 | 0.299 | 0.813 | -0.255 | 0.052 | 1.906 | 1.203 |
| 127 | Hodina = day | Referer = Fulltext | 0.214 | 0.300 | 0.709 | -0.498 | -0.004 | 0.983 | 0.993 |
| 128 | Hodina = day, Pocet stranek = medium | Referer = Fulltext | 0.088 | 0.301 | 0.842 | -0.204 | -0.001 | 0.986 | 0.994 |
| 132 | Pocet stranek = medium | Referer = Fulltext | 0.125 | 0.305 | 0.798 | -0.285 | -0.000 | 0.999 | 1.000 |
| 145 | Hodina = day, SumSkore = medium | Referer = Fulltext | 0.084 | 0.323 | 0.861 | -0.175 | 0.005 | 1.059 | 1.026 |
| 147 | Hodina = day, DobaNaStrance = medium, SumSkore = medium | Referer = Fulltext | 0.077 | 0.326 | 0.871 | -0.159 | 0.005 | 1.068 | 1.031 |
| 148 | Hodina = day, Pocet stranek = low | Referer = Fulltext | 0.084 | 0.326 | 0.862 | -0.173 | 0.005 | 1.068 | 1.031 |
| 149 | SumSkore = medium | Referer = Fulltext | 0.119 | 0.327 | 0.821 | -0.245 | 0.008 | 1.071 | 1.032 |
| 152 | DobaNaStrance = medium, SumSkore = medium | Referer = Fulltext | 0.109 | 0.328 | 0.832 | -0.223 | 0.007 | 1.072 | 1.033 |
| 155 | Pocet stranek = low | Referer = Fulltext | 0.119 | 0.334 | 0.825 | -0.237 | 0.010 | 1.093 | 1.042 |
| 156 | DobaNaStrance = medium | Referer = Fulltext | 0.157 | 0.336 | 0.789 | -0.310 | 0.014 | 1.100 | 1.046 |
| 162 | Hodina = day, DobaNaStrance = medium | Referer = Fulltext | 0.113 | 0.339 | 0.835 | -0.220 | 0.011 | 1.110 | 1.051 |
| 167 | DobaNaStrance = medium, Pocet stranek = low | Referer = Fulltext | 0.087 | 0.346 | 0.868 | -0.165 | 0.010 | 1.132 | 1.062 |

**Obrázek 4: Výsledná asociační pravidla**

Zde jsem se zaměřil na pravidla, jejichž závěr je, že uživatel přišel na stránky pomocí vyhledávače. Jak je vidět z tabulky, pravidla mají obecně střední Confidence, lift parametr je okolo 1, což také není relativně špatné. Z těchto pravidel lze obecně vyčíst, že pokud uživatelé přijdou ve dne a jejich návštěva je z pohledu skóre stránky, doby na stránkách nebo průměrného počtu navštívených stránek „střední“, pak pravděpodobně přišli z fulltextu.

### 6.2.2 Shlukování

K shlukové analýze jsem použil bloku ”Clustering with K-means“, s nastavením:
- K=8
- max runs =100
- max optimalization steps = 1000

Ze shluků můžeme vyvodit, jaká témata uživatele spojují. Jako směrodatná data jsem vybíral TOP 5 navštěvovaných témat ve shluku a zároveň téma navštívilo více než 10% návštěvníků ze shluku.

- **Cluster 0**: Poznávací zájezdy s ubytováním, Turistika
- **Cluster 1**: Obecné, Poznávací zájezdy s ubytováním, Pobyty s výlety, Poznávací zájezdy s lechkou turistikou
- **Cluster 2**: Obecné, Turistika, Poznávací zájezdy s ubytováním, Lastminute
- **Cluster 3**: Obecné
- **Cluster 4**: Obecné, Horská turistika, Vysokohorská turistika, Last minute, Expedice náročné
- **Cluster 5**: Černá hora, Korsika, Obecné
- **Cluster 6**: Lastminute, Obecné, Lyže, Exotika
- **Cluster 7**: Obecné, Exotika, Expedice náročné

**Tabulka 2: Výsledky clusteringu**

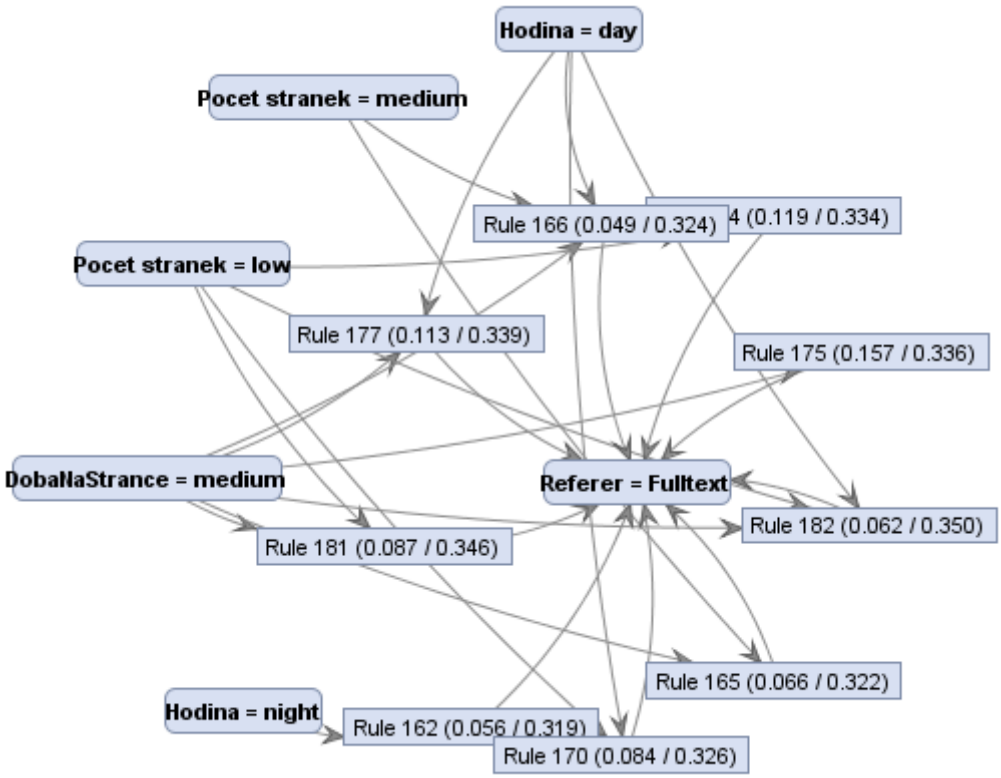
| | Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | Cluster 7 |
|---|---|---|---|---|---|---|---|---|
| **Obecne** | 0.000 | 0.837 | 1.000 | 1.000 | 0.864 | 0.354 | 0.657 | 0.379 |
| **Turistika** | 0.163 | 0.271 | 1.000 | 0.000 | 0.158 | 0.044 | 0.030 | 0.051 |
| **Horska turistika** | 0.076 | 0.199 | 0.118 | 0.000 | 0.998 | 0.048 | 0.014 | 0.014 |
| **VHT** | 0.038 | 0.041 | 0.032 | 0.014 | 0.355 | 0.003 | 0.030 | 0.042 |
| **Voda** | 0.029 | 0.034 | 0.033 | 0.010 | 0.061 | 0.007 | 0.022 | 0.013 |
| **Rafting** | 0.022 | 0.014 | 0.027 | 0.011 | 0.061 | 0.010 | 0.019 | 0.007 |
| **Exotika** | 0.000 | 0.110 | 0.128 | 0.000 | 0.118 | 0.010 | 0.107 | 1.000 |
| **Neni** | 0.060 | 0.947 | 0.057 | 0.010 | 0.067 | 0.010 | 0.063 | 0.038 |
| **Korsika** | 0.042 | 0.168 | 0.073 | 0.037 | 0.056 | 0.109 | 0.073 | 0.022 |
| **Pobyty s vylety** | 0.052 | 0.290 | 0.034 | 0.023 | 0.021 | 0.068 | 0.036 | 0.018 |
| **Hotelbusy** | 0.013 | 0.141 | 0.069 | 0.011 | 0.035 | 0.048 | 0.029 | 0.018 |
| **Lastminute** | 0.000 | 0.218 | 0.147 | 0.000 | 0.198 | 0.027 | 1.000 | 0.046 |
| **Poznavaci zajezdy U** | 0.211 | 0.475 | 0.178 | 0.057 | 0.041 | 0.044 | 0.089 | 0.057 |
| **Expedice Narocne** | 0.041 | 0.070 | 0.055 | 0.015 | 0.174 | 0.024 | 0.076 | 0.114 |
| **Lyze** | 0.067 | 0.019 | 0.009 | 0.016 | 0.098 | 0.020 | 0.136 | 0.011 |
| **Lipari** | 0.025 | 0.141 | 0.040 | 0.024 | 0.021 | 0.058 | 0.057 | 0.022 |
| **Cyklo** | 0.051 | 0.072 | 0.054 | 0.033 | 0.083 | 0.048 | 0.066 | 0.020 |
| **Bulharsko** | 0.040 | 0.077 | 0.018 | 0.033 | 0.032 | 0.088 | 0.055 | 0.013 |
| **Poznavaci zajezdy LT** | 0.048 | 0.700 | 0.043 | 0.007 | 0.015 | 0.112 | 0.014 | 0.008 |
| **Skolni zajezdy** | 0.018 | 0.014 | 0.006 | 0.005 | 0.012 | 0.010 | 0.009 | 0.002 |
| **Ubytovani a doprava** | 0.024 | 0.043 | 0.011 | 0.012 | 0.015 | 0.054 | 0.043 | 0.005 |
| **Pobytove** | 0.035 | 0.158 | 0.019 | 0.033 | 0.018 | 0.071 | 0.077 | 0.019 |
| **Horolezecka skola** | 0.008 | 0.012 | 0.005 | 0.007 | 0.135 | 0.003 | 0.016 | 0.006 |
| **Nezjisten** | 0.026 | 0.005 | 0.007 | 0.003 | 0.008 | 0.003 | 0.005 | 0.007 |
| **Cerna Hora** | 0.000 | 0.070 | 0.022 | 0.000 | 0.024 | 1.000 | 0.022 | 0.001 |
| **Alpy** | 0.028 | 0.177 | 0.012 | 0.007 | 0.070 | 0.031 | 0.025 | 0.007 |
| **Golf** | 0.011 | 0.007 | 0.005 | 0.005 | 0.003 | 0.007 | 0.016 | 0.002 |

Výsledky odpovídají tomu, co jsme očekávali, mezi zajímavé shluky patří například: Cluster 4, kde vidíme, že uživatele spojují témata Horská turistika, Vysokohorská turistika, Last minute, Expedice náročné. Další zajímavým shlukem je Cluster 1, kde uživatele spojují témata Poznávací zájezdy s ubytováním, Pobyty s výlety, Poznávací zájezdy s lehkou turistikou.

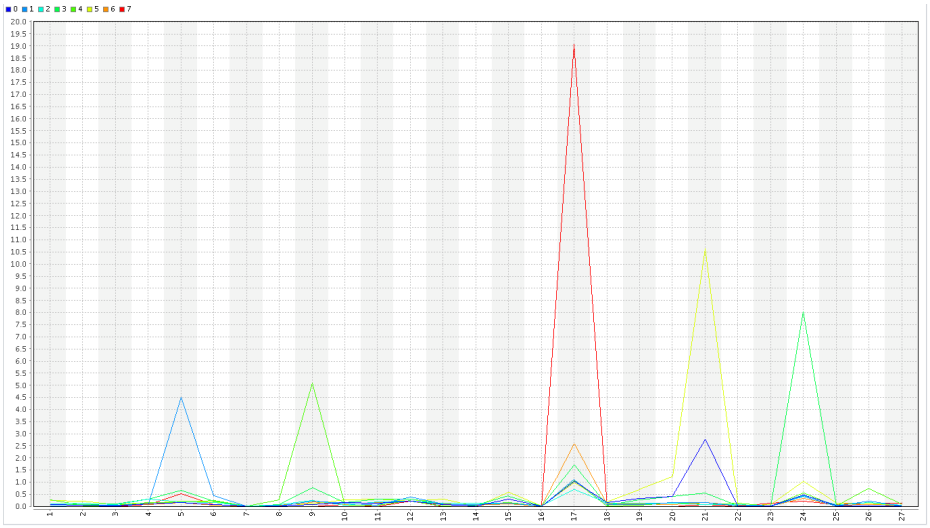
# 7 Zhodnocení výsledků

Závěry analýz u shlukování i asociačních pravidel jsem popsal přímo u výsledků. Celkově lze říci, že jsem poměrně úspěšně rozdělil návštěvníky dle jejich zájmů (tedy dle témat stránek, které navštívili) do shluků.

U asociačních pravidel jsem identifikoval pravidla, která naznačují situace, kdy návštěvníci splní definované cíle, ale také jak se chovají, pokud přijdou z vyhledávačů. Tato pravidla dle mého do jisté míry věrně kopírují reálnou situaci.




**Obrázek 5: Vizualizace Vztahů mezi asociačními pravidly**




**Obrázek 6: Vizualice průběhu clusteringu**

# 8 Závěr

Z dat jsem ověřil, že výsledky odpovídají očekávaným výstupům, byly zjištěny shluky témat, po kterých se jednotlivé skupiny uživatelů nejvíce pohybují a chování návštěvníků webových stránek, kteří přicházejí z vyhledávačů.

Pro přípravu dat bych do budoucna již nevolil MS Excel, jelikož úprava větších tabulek je v tomto nástroji poněkud těžkopádná. Některé operace s velkým množstvím dat trvají v desítkách vteřin, někdy bylo dokonce třeba data rozdělit a zpracovávat postupně. S pomocí RapidMineru se stejné úpravy prováděli mnohem lépe. Zpracování dat bylo nejnáročnější částí celého procesu. Časová náročnost by se dala zmenšit užitím datového skladu či trhu.

# 9 Sezam použitých zdrojů

## 9.1 Seznam užité literatury

BERKA, Petr. Dobývání znalostí z databází. 1. vyd. Praha: Academia, 2003. 366s + Computer Press, 2003. 486str + 1CD ROM. ISBN 80-7226-969-0

LACKO, Luboslav. Databáze: datové sklady, OLAP a dolování dat. 1. vyd. Brno: 1CD ROM. 2003. ISBN 80-200-1062-9

C. M. Bishop. Pattern Recognition and Machine Learning. Springer, 2006. ISBN 0-387-31073-8.

LACKO, Luboslav. Business Intelligence v SQL Serveru 2008. 1. Vyd. Brno: Computer Press, 2009. 456str. ISBN 978-80-251-2887-9 978-1-55860-901-3OO

BUDÍKOVÁ, Marie aj. Průvodce základními statistickými metodami. 1. vyd, Praha: Grada, 2010. ISBN 978-80-247-3243-5

COLIN, SHEARER aj,. CRISP-DM 1.0. *CRISP-DM 1.0* [online]. 2000, s. 78 [cit. 2015-02-16]. Dostupné z: http://www.crisp-dm.org/CRISPWP-0800.pdf

BERKA, Petr. Aplikace systémů dobývání znalostí pro analýzu medicínských dat. Aplikace systému KDD. [Online] 2001. [cit. 2015-02-16]. Dostupné z: http://euromise.vse.cz/kdd/index.php?page=proceskdd

## 9.2 Seznam obrázků



## 9.3 Seznam tabulek