niha je přehledovou publikací nové, bouřlivě se rozvíjející ob-
sti informatiky. Podává základní informace o procesu dobývání
alostí a o principech z oblastí databází, statistiky a strojového
ení, o rozhodovacích stromech, asociačních a rozhodovacích
avidlech, neuronových sítích, genetických algoritmech, baye-
ovských metodách klasifikace, případovém usuzování a induk-
vním logickém programování i o dobývání znalostí z textů
z webu. Je určena vysokoškolským studentům, zájemcům o in-
rmatiku i odborníkům z praxe zabývajícím se analýzou dat.

ISBN 80-200-1062-9
9 788020 010629
www.academia.cz

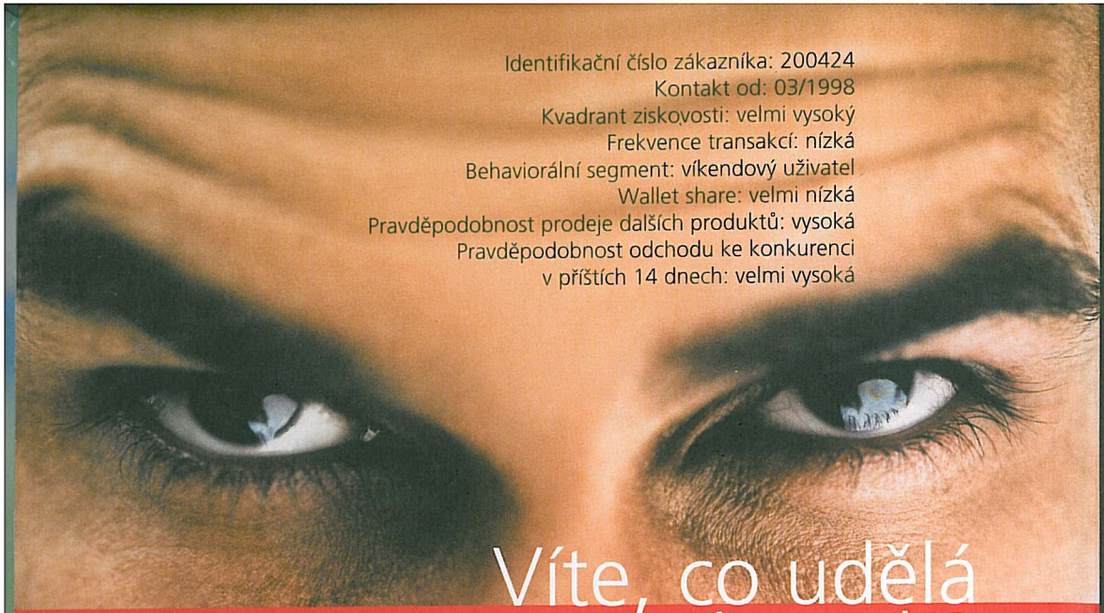
Petr Berka

Dobývání znalostí z databází

Petr Berka

Dobývání znalostí z databází

ACADEMIA



Víte, co udělá váš zákazník?

Co uděláte vy?

S řešeními Customer Intelligence od firmy Adastra se zorientujete v záplavě dat o vašich zákaznících. Můžete segmentovat zákazníky, předvídat jejich chování a ovlivnit ho. Můžete se ještě chytřeji rozhodovat. Základem jsou čistá, kompletní a integrovaná data, která poskytnou kvalitní vstupy pro analytické aplikace s využitím Data Mining.

www.adastra.cz

ADASTRA
CUSTOMER INTELLIGENCE SOLUTIONS

Data Mining projekty společnosti Adastra důsledně využívají metodologie ověřené na zkušenostech z reálného firemního prostředí. Realizovali jsme desítky Data Mining projektů pro společnosti z oblasti bankovnictví, finančnictví, telekomunikací a obchodních řetězců v České republice a Severní Americe. V ČR mezi naše klienty v oblasti Data Mining patří Aliatel, Citibank, Contactel, Česká spořitelna, ČSOB, ČSOB Pojišťovna, Kooperativa a řada dalších. Spolupráce se společností Adastra vám otevře přístup k těmto znalostem a dlouholeté praxi.

Řešení firmy Adastra

- Data Mining – kompletní řešení
- Analytické CRM
- Customer Data Warehouse
- Customer Data Mart
- Data Mining modely
- Skórovací procedury
- Operational Data Stores
- Integrace a čištění dat
- ETL, datové transformace
- Poradenství při výběru technologií
- Business Intelligence, OLAP
- Systémová integrace
- Správa metadat
- Quality Assurance
- Hodnotová segmentace
- Sociálně demografická segmentace
- Behaviorální segmentace
- Analýza a predikce odchodu zákazníků ke konkurenci (Churn, Attrition)
- Analýza pravděpodobnosti nákupu
- Analýza pravděpodobnosti reakce na nabídkovou kampaň
- Analytická podpora Cross-selling a Up-selling
- Prevence a odhalování podvodů (Fraud) analýza rizik
- Analýza a predikce pojistných událostí
- Řízení promocí pro obchodní řetězce

Přejeme Vám příjemné a inspirující čtení!

Adastra – www.adastra.cz – info@adastra.cz

AKADEMIE VĚD ČESKÉ REPUBLIKY

Tato publikace vyšla s podporou
Akademie věd České republiky
a dále těchto firem:
Adastra, s. r. o., Benešovská 10, 101 00 Praha 10
SAS Institute, s. r. o., Na Pankráci 17–19, 140 21 Praha 4

Petr Berka

Dobývání znalostí z databází

ACADEMIA

ISBN 80-200-1062-9

Laskavému čtenáři

Obsah

Předmluva ... 11

I. Dobývání znalostí z databází

1 Dobývání znalostí z databází ... 15
1.1 Úlohy ... 18
1.2 Metodiky ... 22
1.2.1 Metodika 5A ... 22
1.2.2 Metodika SEMMA ... 23
1.2.3 Metodika CRISP-DM ... 24
Literatura ... 28

II. Tři zdroje

2 Databáze ... 33
2.1 Relační databáze ... 33
2.2 EIS ... 35
2.3 OLAP ... 35
2.4 Datové sklady a datová tržiště ... 41
2.5 Dotazovací jazyky pro dobývání znalostí z databází ... 43
Literatura ... 45

3 Statistika ... 46
3.1 Kontingenční tabulky ... 46
3.2 Regresní analýza ... 49
3.3 Diskriminační analýza ... 53
3.4 Shluková analýza ... 55
Literatura ... 59

4 Strojové učení ... 60
4.1 Základní pojmy ... 60
4.2 Učení jako prohledávání ... 69
4.3 Učení jako aproximace funkcí ... 78
Literatura ... 81

III. Proces dobývání znalostí

5 Modelování ... 85
5.1 Rozhodovací stromy ... 86
5.1.1 Základní algoritmus ... 86
5.1.2 Převod stromu na pravidla ... 93
5.1.3 Prořezávání stromů ... 94
5.1.4 Numerické atributy ... 95
5.1.5 Chybějící hodnoty ... 98
5.1.6 Ceny atributů ... 98
5.1.7 Regresní stromy ... 99
5.1.8 Systémy ... 100
5.1.9 Použití rozhodovacích stromů ... 101
5.2 Asociační pravidla ... 102

5.2.1 Základní charakteristiky pravidel ... 103
5.2.2 Generování kombinací ... 106
5.2.3 Počet kombinací ... 107
5.2.4 Algoritmus apriori ... 109
5.2.5 Zobecněná asociační pravidla ... 111
5.2.6 Pravidla s výjimkami ... 113
5.2.7 Časové sekvence ... 114
5.2.8 Více tabulek ... 116
5.2.9 Implikace, dvojité implikace a ekvivalence ... 118
5.2.10 Metoda GUHA ... 121
5.2.11 Kombinační analýza dat ... 125
5.2.12 Chybějící hodnoty ... 128
5.3 Rozhodovací pravidla ... 130
5.3.1 Pokrývání množin ... 130
5.3.2 Rozhodovací seznam ... 134
5.3.3 Pravděpodobnostní pravidla ... 138
5.3.4 Algoritmus ESOD ... 139
5.3.5 Chybějící hodnoty ... 147
5.3.6 Numerické atributy ... 147
5.3.7 Numerické třídy ... 149
5.3.8 Koncepty proměnlivé v čase ... 150
5.3.9 Integrace znalostí ... 153
5.3.10 Hierarchie hodnot atributů ... 155
5.4 Neuronové sítě ... 157
5.4.1 Model jednoho neuronu ... 157
5.4.2 Perceptron ... 163
5.4.3 Topologie soudobých sítí ... 166
5.4.4 Metoda SVM ... 171
5.4.5 Neuronové sítě a dobývání znalostí z databází ... 173
5.5 Evoluční algoritmy ... 176
5.5.1 Základní podoba genetických algoritmů ... 177
5.5.2 Použití genetických algoritmů ... 180
5.5.3 Genetické programování ... 181
5.6 Bayesovská klasifikace ... 182
5.6.1 Základní pojmy ... 182
5.6.2 Naivní bayesovský klasifikátor ... 185
5.6.3 Bayesovské sítě ... 187
5.6.4 Systémy a aplikace ... 196
5.7 Metody založené na analogii ... 197
5.7.1 Podobnost mezi příklady ... 198
5.7.2 Podobnost mezi časovými řadami a sekvencemi ... 201
5.7.3 Učení založené na instancích ... 203
5.7.4 Nejbližší soused ... 205
5.7.5 Případové usuzování ... 209
5.7.6 Systémy IBL ... 210
5.8 Induktivní logické programování ... 211
5.8.1 Základní pojmy ... 212
5.8.2 Systémy ILP ... 214
Literatura ... 217

6 Vyhodnocení výsledků ... 223
6.1 Testování modelů ... 224
6.1.1 Celková správnost ... 227
6.1.2 Správnost pro jednotlivé třídy ... 227

6.1.3 Přesnost a úplnost ... 228
6.1.4 Senzitivita a specificita ... 228
6.1.5 Spolehlivost klasifikace ... 229
6.1.6 Křivka učení ... 230
6.1.7 Křivka navýšení ... 232
6.1.8 Křivka ROC ... 233
6.1.9 Analýza DEA ... 235
6.1.10 Numerické predikce ... 235
6.2 Vizualizace ... 236
6.2.1 Vizualizace modelů ... 236
6.2.2 Vizualizace klasifikací ... 238
6.3 Porovnávání modelů ... 239
6.3.1 t-test ... 239
6.3.2 Použití křivek ROC ... 240
6.3.3 Occamova břitva ... 241
6.4 Volba nejvhodnějšího algoritmu ... 241
6.4.1 STATLOG ... 242
6.4.2 METAL ... 243
6.5 Kombinování modelů ... 243
Literatura ... 245

7 Příprava dat ... 247
7.1 Strukturovaná data ... 247
7.2 Více vzájemně propojených tabulek ... 250
7.3 Odvozené atributy ... 251
7.4 Data s příliš mnoha objekty ... 252
7.5 Data s příliš mnoha atributy ... 253
7.6 Numerické atributy ... 257
7.7 Kategoriální atributy ... 266
7.8 Chybějící hodnoty ... 267
7.9 Závěr ... 267
Literatura ... 267

IV. Systémy a úlohy

8 Systémy pro dobývání znalostí z databází ... 271
8.1 Clementine ... 272
8.2 Enterprise Miner ... 275
8.3 Intelligent Miner ... 277
8.4 Systém Kepler ... 278
8.5 KnowledgeSTUDIO ... 280
8.6 LISp-Miner ... 281
8.7 MineSet ... 284
8.8 Statistica Data Miner ... 286
8.9 Weka ... 287
8.10 Který systém zvolit? ... 289
Literatura ... 290

9 Dobývání znalostí v praxi ... 291
9.1 Příklad úlohy ... 291
9.1.1 Porozumění problematice ... 291
9.1.2 Porozumění datům ... 291
9.1.3 Příprava dat ... 295
9.1.4 Modelování ... 296
9.1.5 Vyhodnocení výsledků ... 300

9.1.6 Využití výsledků ... 301
9.2 Obecné zkušenosti ... 302
Literatura ... 303

10 Nové směry ... 304
10.1 Dobývání znalostí z textů ... 304
10.1.1 Reprezentace dokumentu ... 304
10.1.2 Podobnost dokumentů ... 306
10.1.3 Typy úloh ... 307
10.1.4 Systémy ... 312
10.2 Dobývání znalostí z webu ... 312
10.2.1 Obsah webu ... 313
10.2.2 Struktura webu ... 319
10.2.3 Používání webu ... 320
10.3 Co bude dál ? ... 322
Literatura ... 322

Příloha
A Stručný popis PMML ... 327
A.1 Pravidla pro zápis syntaxe ... 327
A.2 Struktura dokumentu PMML ... 329
A.2.1 Element Header ... 329
A.2.2 Element DataDictionary ... 329
A.2.3 Element TransformationDictionary ... 330
A.2.4 Element pro popis modelu ... 330
A.3 Příklady dokumentů PMML ... 331
A.3.1 Rozhodovací strom ... 331
A.3.2 Asociační pravidla ... 332
A.3.3 Neuronové sítě ... 334
A.3.4 Naivní bayesovský klasifikátor ... 337
A.3.5 Model k-NN ... 340
A.4 Úplný popis DTD ... 341

B Obsah CD ... 352
B.1 Systémy dobývání znalostí ... 353
B.1.1 Systémy na CD ... 353
B.1.2 Informace o dalších nekomerčních systémech ... 354
B.1.3 Informace o komerčních systémech ... 355
B.2 Data a úlohy ... 356
B.2.1 PKDD Discovery Challenge – finanční data ... 357
B.2.2 Soutěžní úlohy dobývání znalostí ... 357
B.2.3 Referenční data ... 358
B.3 Výzkumné projekty EU ... 358
B.3.1 Networks of Excellence ... 358
B.3.2 Konkretní výzkumné projekty ... 359
B.4 Zdroje z Internetu ... 359
B.4.1 Dokumenty na CD ... 359
B.4.2 Informační portály ... 360
B.4.3 Katalogy ve vyhledávačích ... 360
B.4.4 Časopisy ... 361
B.4.5 Různé ... 361

Rejstřík ... 363

Předmluva

Christieho–Daviesův teorém:

Máte-li špatné údaje, ale dokonalou logiku, pak jsou vaše závěry zcela jistě mylné. Dopřejete-li si tudíž sem tam nějakou trhlinu v logickém uvažování, můžete díky náhodě dospět ke správnému závěru.

Murphyho zákony

Mým cílem bylo napsat přehledovou publikaci, která by postihla hlavní rysy dobývání znalostí z databází a strojového učení. Snažím se tak částečně vyplnit mezeru v nabídce českých publikací zaměřených na tuto tak bouřlivě se rozvíjející oblast informatiky. Zatímco anglicky psaných knih věnovaných tomuto tématu existuje veliké množství, nepočítám-li dílčí kapitoly v knihách Mařík, Štěpánková, Lažanský a kol.: *Umělá inteligence* a Sklenák a kol.: *Data, informace, znalosti a Internet*, popř. publikace zaměřené například na rozpoznávání obrazů nebo neuronové sítě, ucelenější česká práce zatím chybí.

Kniha je členěna do čtyř částí. Úvodní část „Dobývání znalostí z databází" podává základní informace, část druhá, nazvaná „Tři zdroje", ukazuje ty principy z oblastí databází, statistiky a strojového učení, které nejvíce ovlivnily dobývání znalostí. Třetí, nejobsáhlejší část, se věnuje těm krokům procesu dobývání znalostí, které lze nejlépe popsat v obecné rovině, tedy nezávisle na konkrétní úloze a aplikaci. Moji snahou bylo co možná nejlépe pokrýt zejména používané analytické metody. Vycházel jsem přitom především z vlastních, více než desetiletých zkušeností s prací v oblasti symbolických metod strojového učení. Tomu odpovídá i důraz na tuto oblast. Při zpracování témat vzdálenějších jsem čerpal z reprezentativní, ale spíše přehledové literatury i z množství odborných článků. Za všechny zdroje bych chtěl na tomto místě uvést knihy Machine Learning od Toma Mitchella a Data Mining od Iana Wittena a Eibeho Franka. Jsem si vědom toho, že řadě témat, o kterých jsem se v této knize zmínil třeba jen letmo, jsou věnovány samostatné mnohasetstránkové publikace. Zvídavý čtenář však na závěr každé kapitoly nalezne odkazy na další literaturu, kde může hledat odpovědi na své otázky. Závěrečná část představuje stručný přehled některých známých systémů pro dobývání znalostí a ukazuje i podrobnější příklad aplikace.

Jeden z problémů, na které jsem při psaní narazil, byla otázka české terminologie. Ne ke všem anglickým termínům existují české ekvivalenty a pokud existují, nebývají jednoznačné. Vždyť i samotné základní pojmy knowledge discovery in databases nebo data mining se překládají různě. Proto jsem se přidržel vlastního rozumu: důsledně používám termín dobývání znalostí z databází (pro KDD), naopak data mining se v textu objevuje (ve shodě s metodikou CRISP-DM) jako modelování, analytické metody nebo analytické procedury a vyhýbal jsem se pojmu dolování dat, který není podle mého názoru příliš přesný.

Chtěl bych poděkovat všem kolegům, se kterými jsem měl možnost spolupracovat a kteří ovlivnili můj pohled na oblast dobývání znalostí a strojové učení. Jsou to zejména Ivan Brůha, Petr Hájek, Tomáš Havránek, Jiří Iványk, Radim Jiroušek, Vladimír Mařík, Katharina Morik, Emil Pelikán, Jan Rauch, Břetislav Stejskal, Vojtěch Svátek, Olga Štěpánková, Shusaku Tsumoto, Gerhard Widmer a Jan Zytkow.

Součástí knihy je i CD obsahující řadu dalších informací volně přístupných na Internetu. Mimo jiné se jedná o plné texty tří knih a o tři plně funkční systémy. V této souvislosti patří můj obzvláštní dík autorům těchto knih – Petru Hájkovi, Davidu Spielgelhalterovi, Jiřímu Šímovi a Romanu Nerudovi – a systémů – Petru Hájkovi, Janu Paraličovi, Janu Rauchovi a Milanu Šimůnkovi – za jejich laskavý souhlas s uveřejněním.

Děkuji rovněž pracovníkům a spolupracovníkům nakladatelství Academia, Praha, zejména A. Baďurovi a V. Havlíčkovi za trpělivost při převodu mého rukopisu do výsledné knižní podoby.

Závěrečné poděkování patří laskavému čtenáři, který snad zamhouří oči nad možnými chybami v této knize.

Petr Berka

I. DOBÝVÁNÍ ZNALOSTÍ Z DATABÁZÍ

1 Dobývání znalostí z databází

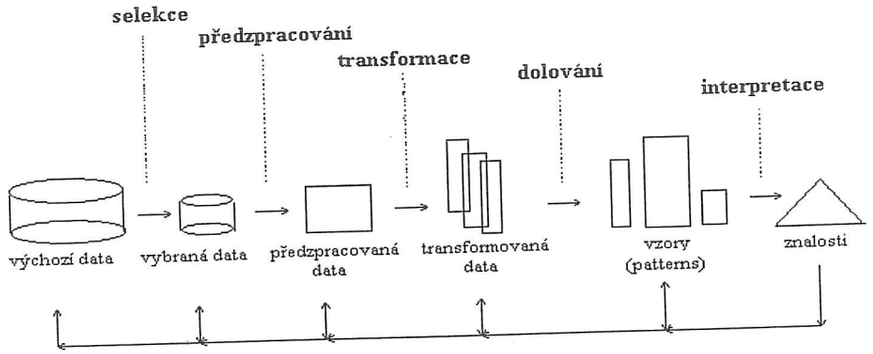
O dobývání znalostí z databází (Knowledge Discovery in Databases, KDD) se začalo ve vědeckých kruzích mluvit počátkem 90. let minulého století. První impuls přišel z Ameriky, kde se na konferencích věnovaných umělé inteligenci (mezinárodní konference o umělé inteligenci IJCAI'89 nebo konference americké asociace umělé inteligence AAAI'91 a AAA'93) pořádaly první workshopy věnované této problematice. Nebyla to ale jen umělá inteligence (přesněji řečeno metody strojového učení), které stály u zrodu dobývání znalostí z databází. Databázové technologie představují osvědčený prostředek jak uchovávat rozsáhlá data a vyhledávat v nich informace, statistika představuje osvědčený prostředek jak modelovat a analyzovat závislosti v datech. Po léta se tyto disciplíny vyvíjely nezávisle, až přišla ta chvíle, kdy rozsah automaticky sbíraných dat začínal uživatelům přerůstat přes hlavu. Současně s tím také vznikla potřeba tato data používat pro podporu (strategického) rozhodování ve firmách. Zájem finančně silných uživatelů o aplikace pak stimuloval ono propojení a dal vzniknout (a hlavně popularitu) dobývání znalostí z databází. Neustálý nárůst zájmu odborné komunity dokládá množství konferencí (americké konference KDD, asijské konference PAKDD, evropské konference PKDD), vznik odborných skupin (např. special interest group for KDD – SIGKDD – při americké asociaci ACM) i vznik samostatných odborných časopisů (časopis Data Mining and Knowledge Discovery vydávaný nakladetelstvím Kluwer). Tematika dobývání znalostí si postupně našla cestu i do šířeji zaměřených počítačových časopisů. Dnes již není nic neobvyklého, že na pojmy knowledge discovery, data mining, nebo business intelligence[1] narazíme i v reklamách počítačových firem.

Dobývání znalostí z databází (KDD) lze definovat jako *netriviální získávání implicitních, dříve neznámých a potenciálně užitečných informací z dat* (Fayyad a kol, 1996). Zpočátku se pro tuto oblast razily nejrůznější názvy: information harvesting, data archeology, data destilery. Nakonec ale převládla hornická metafora; dobývání znalostí a dolování z dat (data mining). Po jistém období tápání se ustálilo i chápání KDD jako interaktivního a iterativního procesu tvořeného kroky selekce, předzpracování, transformace, vlastního „dolování" (data mining) a interpretace (obr. 1).

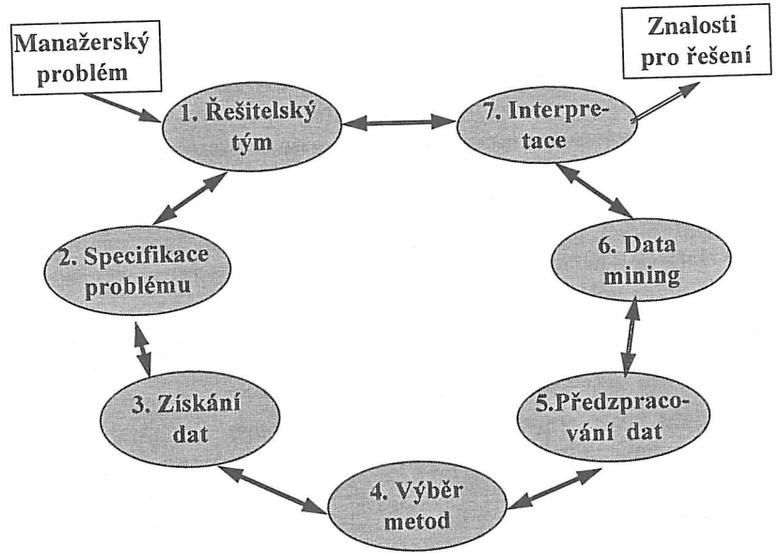
Na rozdíl od „prostého" použití statistických metod a metod strojového učení se v procesu dobývání znalostí již klade důraz i na přípravu dat pro analýzu a na

[1] Význam pojmu business intelligence je možné (s trochou nadsázky) interpretovat touto rovnicí: business intelligence = artificial intelligence + business.

interpretaci výsledných znalostí. Při přípravě dat se obvykle z dat uložených ve složité struktuře, např. datového skladu, vytváří jedna tabulka obsahující relevantní údaje (hodnoty atributů) o sledovaných objektech (např. klientech banky nebo zákaznících obchodního domu). Při interpretaci se nalezené znalosti[2] hodnotí z pohledu koncového uživatele.



Obr. 1 Proces dobývání znalostí z databází dle knihy (Fayyad a kol, 1996).



Obr. 2 Manažerský pohled na proces dobývání znalostí z databází.

[2] Fayyad rozlišuje mezi znalostmi získanými jako výstup z kroku dolování (nazývá je vzory – patterns) a mezi znalostmi interpretovanými uživatelem. My toto rozlišení nebudeme provádět.

Zatímco schéma na obr. 1 popisuje „technologický" pohled na dobývání znalostí, Anand (Anand a kol., 1996) nabízí pohled manažerský (obr. 2). Impulsem pro zahájení procesu dobývání znalostí je nějaký r e á l n ý p r o b l é m . Cílem procesu dobývání znalostí je získat co nejvíce relevantních informací vhodných k řešení daného problému. Příkladem reálného problému je otázka nalezení skupin zákazníků obchodního domu nebo skupin klientů banky, kterým by bylo možné nabídnout speciální služby. U zákazníků obchodního domu se může jednat o zjištění, že zákazník kupuje potravinářské zboží odpovídající jisté dietě, v případě klientů banky může jít o potenciální zájemce o hypoteční úvěr. Nalezené skupiny jsou interpretovány jako takzvané segmenty trhu v dané oblasti.

Prvním krokem při řešení problému je v y t v o ř i t ř e š i t e l s k ý t ý m . Jeho členy musí být *expert na řešenou problematiku*, *expert na data* – jak v organizaci, tak popřípadě i na externí data – a *expert na metody KDD*. V případě rozsáhlejších problémů je obvyklé, že jednotliví experti mají k dispozici vlastní tým, nebo alespoň využívají konzultací s dalšími experty.

Prvním úkolem sestaveného týmu je s p e c i f i k a c e p r o b l é m u , který je třeba řešit v souvislostech dobývání znalostí. U zákazníků obchodního domu nakupujících potravinářské zboží odpovídající jisté dietě je mimo jiné třeba specifikovat položky zboží odpovídající různým dietám. U skupin zákazníků nakupujících položku A a nenakupujících položku B je krom jiného třeba vytipovat vhodné skupiny položek atd.

Po specifikaci problému je třeba z í s k a t v š e c h n a d o s t u p n á d a t a , která mohou být použita pro řešení problému. Znamená to posoudit všechna dostupná data a zvážit, zda odpovídají danému problému. Tento proces může vyvolat menší či větší přeformulování problému. V některých případech je třeba pracovat i s daty, která jsou archivována po delší dobu ve formě datových souborů a ne v databázi, data jsou někdy dokonce uložena v několika různých systémech. Náročnost získání dat je nepřímo úměrná úrovni datové základny, která je k dispozici.

V mnohých případech je vhodné uvažovat i *externí data* popisující prostředí, ve kterém se analyzované děje odehrávají. V případě klientů banky i zákazníků obchodního domu je důležitou informací kalendářní období (např. vánoce, velikonoce, období letních a zimních dovolených, den kdy zákazníci dostávají výplatu, pondělí, úterý, ...). Na zákazníky bude mít jistě vliv i počasí, reklama probíhající ve sdělovacích prostředcích, v některých případech i politické události.

Cílem *výběru metody* je zvolit vhodné metody analýzy dat. V rámci dobývání znalostí z databází je používána řada typů metod analýzy dat, ve většině případů je k řešení konkrétní úlohy zapotřebí kombinovat více různých metod. Mezi používané typy metod patří například klasifikační metody, různé klasické

metody explorační analýzy dat, metody pro získávání asociačních pravidel, rozhodovací stromy, genetické algoritmy, bayesovské sítě, neuronové sítě, hrubé množiny (rough sets), velmi používané jsou i metody vizualizace. Dá se také předpokládat vývoj dalších metod.

V rámci *předzpracování dat* se data získaná k řešení specifikovaného problému připravují do formy vyžadované pro aplikaci vybraných metod. V řadě případů se může jednat o značně náročné výpočetní operace. Do této fáze patří i odstranění odlehlých hodnot, popř. doplnění chybějících hodnot.

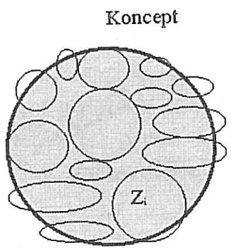
Krok *data mining* zahrnuje aplikaci vybraných analytických metod pro vyhledávání zajímavých vztahů v datech. Obvykle jsou jednotlivé metody aplikovány vícekrát, hodnoty vstupních parametrů jednotlivých běhů závisí na výsledcích předchozích běhů. Zpravidla se nejedná o aplikace metod jenom jednoho typu, jednotlivé typy se kombinují na základě dílčích výsledků.

Cílem *interpretace* je nezbytné zpracování obvykle značného množství výsledků jednotlivých metod. Některé z těchto výsledků vyjadřují skutečnosti, které jsou z hlediska uživatele nezajímavé nebo samozřejmé. Některé výsledky je možné použít přímo, jiné je nutné vyjádřit způsobem srozumitelným pro uživatele. Jednotlivé výsledky je často vhodné uspořádat do analytické zprávy. Analytická zpráva však není jediným možným výstupem procesu dobývání znalostí. Výstupem může být i provedení vhodné akce jako například zapnutí monitorovacího programu.

1.1 Úlohy

V případě dobývání znalostí z databází můžeme mluvit o různých typech úloh. Jsou to především (Klosgen a Zytkow, 1997)[3]:

- *klasifikace* nebo *predikce*,
- *deskripce*,
- *hledání „nuggetů"*.

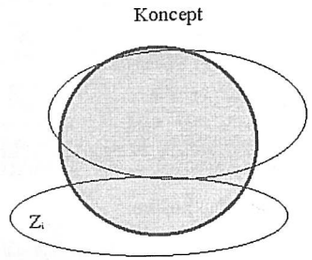


Obr. 3 Klasifikace nebo predikce.

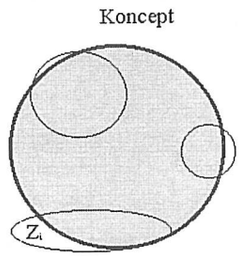
[3] Podrobnější členění lze nalézt v práci (Chapman a kol, 2000). Tvůrci metodiky CRISP-DM zde uvádějí úlohy deskripce dat a sumarizace, segmentace, deskripce konceptů, klasifikace, predikce a analýzy závislostí.

Při klasifikaci, popř. predikci, je cílem nalézt znalosti použitelné pro klasifikaci nových případů – zde požadujeme, aby získané znalosti co nejlépe odpovídaly danému konceptu; dáváme přednost přesnosti pokrytí na úkor jednoduchosti (připouštíme větší množství méně srozumitelných dílčích znalostí tak, jak je to naznačeno na obr. 3). Rozdíl mezi klasifikací a predikcí spočívá v tom, že u predikce hraje důležitou roli čas; ze starších hodnot nějaké veličiny se pokoušíme odhadnout její vývoj v budoucnosti (např. předpověď počasí nebo pohybu cen akcií).

Při deskripci (popisu) je cílem nalézt dominantní strukturu nebo vazby, které jsou skryté v daných datech. Požadujeme srozumitelné znalosti pokrývající daný koncept; dáváme tedy přednost menšímu množství méně přesných znalostí (viz obr. 4). Hledáme-li nuggety, požadujeme zajímavé (nové, překvapivé) znalosti, které nemusí plně pokrývat daný koncept (obr. 5).



Obr. 4 Popis (deskripce).



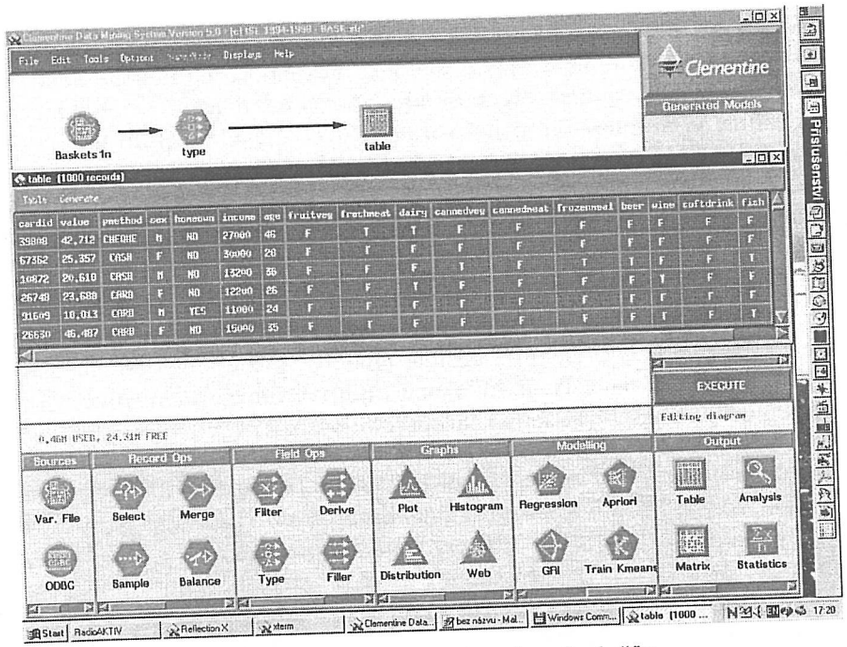
Obr. 5 Nuggety.

Úlohy dobývání znalostí lze nalézt v celé řadě aplikačních oblastí:

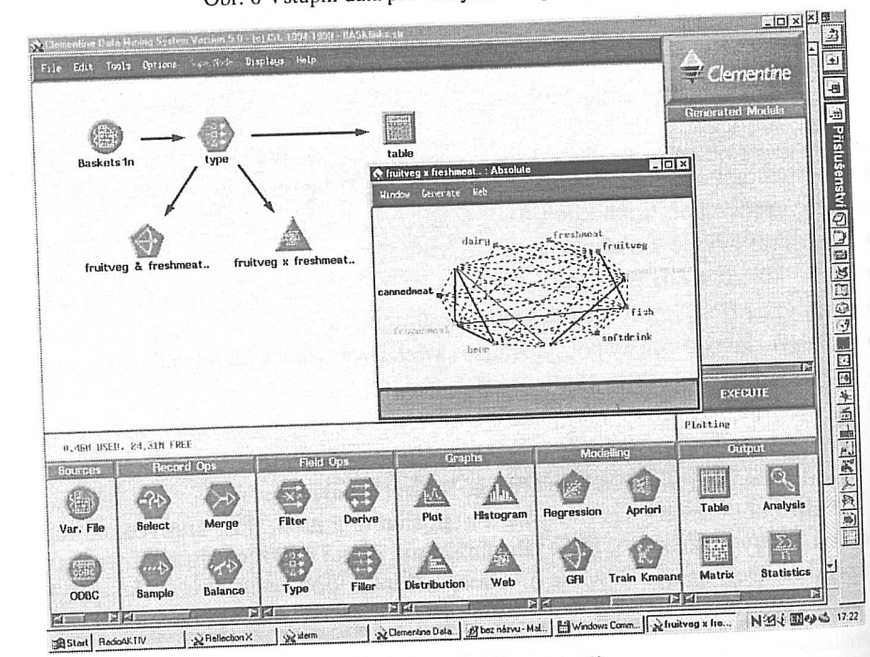
- segmentaci a klasifikaci klientů banky (např. rozpoznání problémových nebo naopak vysoce bonitních klientů),
- predikci vývoje kurzů akcií,
- predikci spotřeby elektrické energie,
- analýze příčin poruch v telekomunikačních sítích,
- analýze důvodů změny poskytovatele služeb (Internet, mobilní telefony),
- segmentaci a klasifikaci klientů pojišťovny,
- určení příčin poruch automobilů,
- rozboru databáze pacientů v nemocnici,
- analýze nákupního košíku (Market Basket Analysis).

Podrobněji se zde podívejme na poslední z nich. Při analýze nákupního košíku se vychází z dat, která shromažďují různé řetězce supermarketů (u nás například Delvita nebo Meinl). Data, alespoň podle následujícího příkladu[4],

[4] Příklad je převzat z ukázkového příkladu v systému *Clementine*. O tomto systému najde čtenář podrobnější výklad v kapitole 8.



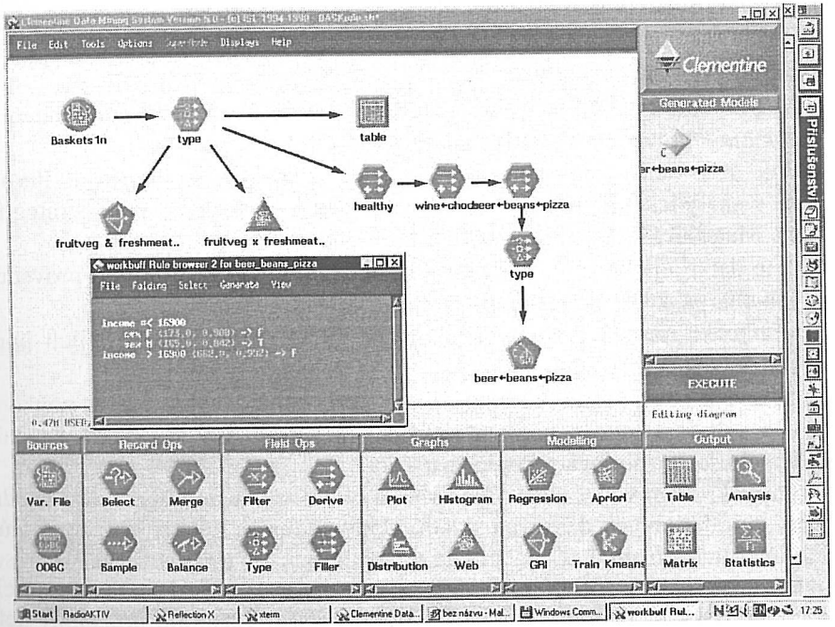
Obr. 6 Vstupní data pro analýzu nákupního košíku.



Obr. 7 Asociace mezi typy zboží.

tvoří jednak charakteristiky zákazníků (pohlaví, vlastnictví domu, příjem a věk), jednak údaje o jednotlivých nákupech (způsob placení, částka, zakoupený typ zboží). Data jsou již předzpracována do podoby relační tabulky. Co záznam, to jeden zákazník, typy zboží jsou pevně dány – uvádí se, zda konkrétní výrobek byl nebo nebyl zakoupen (obr. 6). V takovýchto datech můžeme například hledat souvislosti mezi jednotlivými typy zboží; bude nás zajímat, zda existují skupiny produktů, které si zákazníci kupují současně (např. pivo a párek). Obrázek 7 ukazuje, že velmi často se v nákupním košíku objevuje současně například pivo, zmrazené maso a konzervovaná zelenina nebo ryby, ovoce a zelenina[5].

Samozřejmě, že „zdravou" a „nezdravou" výživu nekupují titíž zákazníci. Může nás tedy zajímat, čím se tyto skupiny zákazníků vyznačují. Takové znalosti je možné získat například pomocí rozhodovacích stromů. Obrázek 8 ukazuje, že pizzu, pivo a fazole nakupují muži s nižším příjmem.



Obr. 8 Konzumenti nezdravé výživy.

[5] Systém Clementine zde použil grafický způsob prezentace těchto asociací; hrany v grafu odpovídají vazbám mezi produkty.