DATA MINING A JEHO VYUŽITÍ

Petr Klímek

Abstrakt:
Tento příspěvek se zabývá problematikou získávání znalostí z dat. Jsou v něm definovány základní pojmy a postupné kroky, ve kterých probíhá dolování dat. Dále jsou zde uvedeny hlavní dodavatelé softwarových produktů v České republice - jedná se o renomované softwarové firmy. Příspěvek rovněž rozebírá nutné základní znalosti uživatelů, kteří provádějí data mining v podniku. Poslední odstavec se věnuje výzkumu v oblasti použití metod a speciálních softwarových produktů určených pro data mining.

Klíčová slova: *získávání znalostí z dat, data mining, softwarové produkty, statistika, metody data miningu*

1. Proces získávání znalostí z dat

Proces získávání znalostí z dat (knowledge discovery in databases - dále KDD) je chápán jako proces netriviálního objevování implicitních, dopředu neznámých a potenciálně použitelných vzorů z dat. Zatímco **dolování dat** (data mining - dále DM) **je pouze krokem v procesu KDD** založeným na aplikaci výpočetních technik, které na základě daných omezení (výpočetní efektivnost) poskytují enumeraci vzorů či modelů nad danými daty.

Existují i další, alternativní názvy, např. dolování znalostí z databází, extrakce znalostí, archeologie dat, bagrování znalostí, analýza dat apod. Objevování znalostí vede k extrakci zajímavých zákonitostí či informací vyšší úrovně, které mohou být studovány z dalších úhlů pohledu. Tento úkol je ve své podstatě interaktivní a iterativní. [1], [2]

Proces KDD zahrnuje tyto fáze:

- **selekce** - data se vybírají nebo segmentují podle nějakého kritéria. Selekcí je omezení např. všech osob na ty, kteří vlastní automatickou pračku. Pro některé algoritmy DM stačí selekcí vybrat pouze vzorky dat, není nutné zapojit do zpracování celý datový sklad (data warehouse - DW);
- **předzpracování** - znamená čištění dat, kdy některá data jsou odstraňována, protože nejsou potřebná a bránila by efektivnímu vyhodnocení dotazu. Např. při objevování znalostí o porodnosti je možné uvažovat z registru pacientů pouze ženy a není tedy nutné přejímat atribut pohlaví. Součástí čištění je také úprava formátů dat, např. kód pohlaví se unifikuje na binární atribut s hodnotami 0 nebo 1;
- **transformace** - nejsou přenášena pouze vyčištěná data, ale jsou rozšířena o další atributy např. z externích zdrojů (demografické atributy), které obohatí použitelnost dat;
- **dolování dat** - jde o stádium, které se zabývá extrakcí vzorů dat. Zde vybíráme vhodnou techniku DM (klasifikace, regrese, shlukování, neuronové sítě apod.). Dále v této fázi vybereme konkrétní algoritmus pro řešení DM úlohy. Nakonec tento krok obsahuje vlastní vyhledání zajímavých znalostí, jejichž forma závisí na zvolené metodě DM a může mít podobu klasifikačních pravidel nebo stromů, funkčních závislostí, logických pravidel atd.;
- **interpretace a vyhodnocení** - vzory identifikované systémem jsou vyhodnoceny jako znalosti, které mohou být použity k podpoře učinění rozhodnutí managera. Rozhodování je vztaženo k úlohám týkajících se predikce, klasifikace apod. tak, že je sumarizován obsah databáze nebo jsou vysvětleny pozorované jevy.[4]

Je třeba zdůraznit, že proces je řízený uživatelem a využívá jeho schopnosti a znalosti. Apriorní, předem známá fakta hrají klíčovou roli především v přípravě dat. Každá databáze je připravována s určitým cílem. Uživatel má tedy alespoň přibližnou představu o tom, jaká data obsahuje a jaký typ znalostí by pro něj mohl být užitečný. To ovšem neznamená, že by první fáze byla nepodstatná

nebo jednoduchá. Na vhodné volbě cílů a přípravě dat často závisí úspěch celého KDD procesu. Proto je tento proces často iterativně opakován. Již získané znalosti pomáhají lépe specifikovat cíle a metody při opakovaném hledání. Situaci navíc často ztěžuje heterogenní prostředí. Různé druhy dat jsou uchovávány v různých typech databází - relačních, objektových, deduktivních, aktivních, hypertextových a multimediálních, časových, prostorových a jiných. Zajímavou výzvou je také hledání znalostí v distribuovaných databázích, například v prostředí Internetu (tzv. webmining).

Dodržení systematického přístupu k úspěšnému modelování při získávání znalostí z dat si uvědomily rovněž velké společnosti jako SAS, který využívá procesu SEMMA (**S**ample, **E**xplore, **M**odify, **M**odel, **A**ssess) a SPSS, který zavedl systém **5A´s** (**A**ssess, **A**ccess, **A**nalyze, **A**ct, **A**utomate). [7]

V literatuře [1] nejčastěji uváděná metodologie CRISP-DM vznikla na základě potřeb řešitelů data miningových úloh. Jejím úkolem je poskytnout strukturovaný nezávislý popis jednotlivých kroků data miningového projektu. V roce 1996 bylo založeno konsorcium společností, které ve své praxi využívali data mining, za účelem vypracování úlohově a softwarově nezávislého popisu kroků při řešení takových úloh.

Výsledkem projektu byl v roce 2000 velmi úspěšný procesní model CRISP-DM 1.0. Na jeho výstavě se podílely čtyři významné společnosti: NCR Systems Engineeering Copenhagen (USA, Dánsko), DaimlerChrysler AG (Německo), SPSS, Inc. (USA) a OHRA Verzekerigen en Bank Groep B.V. (Nizozemí).

Metodologie CRISP-DM je hierarchicky členěna do čtyř úrovní podle abstrakce činností prováděných na jednotlivých úrovních: fáze projektu, obecné úlohy, specializované úlohy a nastavení procesů. Na nejvyšší úrovni je proces rozdělen do šesti vzájemně propojených fází.

Řešení data miningových úloh se skládá z mnoha dílčích a vzájemně provázaných kroků. Mnohé projekty naráží na technické problémy jaké jsou například transformace dat, propojení různých softwarových nástrojů či potřeba vkládat programový kód. Takové problémy mohou nepřiměřeně zatěžovat a uživatel se nemůže plně věnovat hledání optimálního řešení.

Data miningový software Clementine (bude analyzován dále) nabízí integraci všech klíčových nástrojů a pokrývá tak celý proces řešení data miningových úloh. Uživatel metodou vizuálního programování, kdy pomocí myši klade na pracovní plochu jednotlivé nástroje a naznačuje jejich propojení, vytváří procesní diagram řešení dané úlohy. Navíc je veden metodologií CRISP-DM, která zajišťuje dosažení optimálního řešení v co možná nejkratším čase. Clementine je naprogramována jako formátově nezávislý software, tzn. není nadstavbou žádné databáze. Clementine není uzavřený systém, modely a procesy vybudované v Clementine lze exportovat a poskytovat dalším aplikacím či provozovat na počítačích, kde není Clementine instalovaná. Uživatel může do tohoto systému integrovat vlastní nástroje jako například speciální modelovací algoritmy.

Spojení metodologie CRISP-DM a data miningového softwaru Clementine, který tuto metodologii plně podporuje, zajišťuje uživatelům maximální efektivitu při zpracování data miningových projektů. Při řešení může uživatel optimalizovat čas potřebný pro nalezení řešení a zároveň získat velmi kvalitní data miningový model.

Jako poslední zde zmiňované lze použít metodologie společnosti Two Crows Corporation v [1], která využívá výše popsané metodologie CRISP-DM. Základní kroky v procesu data miningu jsou podle ní následující:

1. Definuj obchodní problém.
2. Vytvoř data miningovou databázi.
3. Získej data.
4. Připrav data pro model.
5. Sestav model.
6. Interpretuj model.
7. Využij model a výsledky.

2. Nabídka na českém trhu

Jak již bylo řečeno, DM je prudce se rozvíjející oblastí, do které investuje v současné době mnoho softwarových společností. Aktuální stav na trhu v oblasti DM je tedy poměrně dynamický; v tabulce 1 jsou uvedeni největší dodavatelé DM softwarových nástrojů v ČR.

Na trhu je v současnosti kromě výše uvedených mnoho dalších produktů, které implementují nejrůznější algoritmy. Jejich aktuální přehled nalezneme například na [6]. Například soutěže programů pro získávání informací KDD-CUP98, konané v rámci konference KDD-98, se zúčastnilo 57 produktů. Vítězem se stala firma Urban Science Application s programem pro prediktivní

Tab. 1: Přehled největších světových dodavatelů technologií pro DM

Dodavatel	Hlavní produkt	Kontakt pro ČR
SAS Institute, Inc.	Enterprise Miner	SAS Institute ČR, s.r.o; www.sas.com
IBM Corporation	DB2Intelligent Miner for Data	IBM ČR, spol.s r.o.; www.ibm.cz
SPSS, Inc.	Clementine	SPSS ČR, spol. s r.o; www.spss.cz
Silicon Graphics, Inc.	Mine Set	Silicon Graphics, s.r.o.; www.sgi.cz
Angoss Software Corp.	Knowledge Studio	Speedware, s.r.o.; www.speedware.cz
StatSoft, Inc.	STATISTICA Data Miner	StatSoft CR, s.r.o; www.statsoft.cz

Zdroj: vlastní zpracování podle [6]

modelování GainSmart. Její program, který je specializovanou nadstavbou nad statistickými nástroji firmy SAS Institute, těsně zvítězil nad obecným řešením pro dolování dat Enterprise Miner právě od SAS Institute. Nástroje pro získávání znalostí a jejich využití jsou ve stejné situaci, jako byla relační databázová technologie ke konci 80. let. Komerční uživatelé z nejprogresivnějších společností již léta používají vlastní speciálně vyvinuté programy, např. pro modelování marketingových kampaní či analýzu úvěrového rizika. Další společnosti přivádí k zavádění technik dolování dat zostřená konkurence na trhu, zvyšující se počty cílových zákazníků. Obě tyto skupiny dnes hledají standardizované řešení, která pokrývají nejrůznější typy úloh a poskytují výstupy snadno srozumitelné managementu. Zároveň tato řešení musí být schopná nasazení v prostředí IT současných komerčních společností.

Zájem takových výrobců, jako SAS Institute s produktem Enterprise Miner, IBM nabízející produkt Intelligent Miner, či Mineset firmy Silicon Graphics ukazuje, že přední softwarové firmy si uvědomují potenciál získávání informací z dat na trhu informačních technologií.

3. Nutné znalosti uživatelů při obsluze softwarových nástrojů

Data mining představuje skupinu vzájemně dosti odlišných metod a technik, jejichž dobrou kombinací i vhodnou parametrizací lze dosáhnout optimálních výsledků. Bylo již zmíněno, že vytváření určitého modelu je iterativní záležitostí, změna jednoho parametru či způsobu výběru a normalizace vstupních dat může dramaticky změnit vlastnosti (kvalitu) výsledného modelu.

Pro praktickou práci se softwarovými nástroji nejsou potřebné zcela detailní teoretické znalosti algoritmů, které využívají jednotlivé nástroje. Nezbytné minimum ovšem podle [5] je:

1) alespoň základní průprava v matematické statistice,
2) zručnost při připojování a operacích s různými datovými zdroji, ovládání některého z nástrojů pro transformaci a integraci dat,
3) porozumění obchodnímu cíli a dobrá komunikativnost při spolupráci s ne-technicky a s neanalyticky zaměřenými kolegy v průběhu vzniku a ověřování modelu a provozního používání zjištěných znalostí,
4) analytické myšlení a praktické zkušenosti umožňující, leckdy i s přispěním intuice, zvolit správnou cestu pro nalezení vhodného modelu pro řešení daného obchodního zadání.

4. Výzkum v oblasti softwarových produktů pro získávání znalostí z dat v ČR

Komerční data mining projekty jsou realitou a jejich počet roste. Z vlastního průzkumu mezi hlavními dodavateli softwarových produktů pro data mining [3] se zdá, že rok 2001-2002 byl v oblasti data miningu určitým průlomem - bylo realizováno několik standardních data miningových projektů včetně realizace skórovacích mechanismů a využití výsledků v obchodně-provozním chování organizací.

Komerční aktivity v oblasti softwarových produktů samozřejmě nejsou jediným okruhem: problematikou získávání znalostí z dat se zabývají rovněž i skupiny na akademické či výzkumné půdě (například v rámci VŠE Laboratoř inteligent-

ních systémů) a některé podniky řeší tuto problematiku interně, například s přispěním expertýzy mateřských zahraničních společností.

Pro informace ohledně využívání softwarových produktů se bohužel objevuje stejný trend, jako je tomu v zemích, kde jsou tyto produkty již běžnější: zákazníci význam těchto projektů kladou tak vysoko, že si vymiňují mlčenlivost o obsahu využití, ne-li přímo o faktu(!), že jej používají.

Zákazníci

Je třeba říci, že ještě v roce 2002 byl počet zákazníků poměrně nízký. Podle rozhovorů s dodavateli tito odhadují růst implementací softwarových produktů asi o 20-40 % v letech 2003-2005. Zákazníci v ČR byli ze stejných oblastí, jako byli první tzv. „early adopters" těchto softwarových produktů v zahraničí: především finanční instituce (banka, pojišťovna) a telekomunikační společnosti.

Zákaznické skupiny v rámci těchto organizací byly různé: hlavními účastníky projektu byli vedoucí oddělení IT na jedné straně, naopak u jiného zákazníka to byl například marketingový ředitel. Je zřejmé, že konečný úspěch může být dosažen pouze kooperací se specialisty jak z oblasti IT tak dotčené oblasti obchodní.

Dle zkušeností dodavatelů je pravidlem, že možnosti získávání znalostí z dat nejsou v organizacích známy a při přípravě projektů jsou i poměrně obtížně sdělitelné. Potenciál získávání znalostí (data miningu) se projevil teprve na základě konkrétních výsledků. Celkový přehled o možnostech získávání znalostí z dat (data miningu) pro obchodní či provozní potenciál podniků je dnes v ČR poměrně malý.

Řešené úlohy

Dá se říci, že typy řešených úloh se neliší od nejčastěji řešených úloh v zahraničí. Podle dodavatelů jsou nejčastějšími oblastmi segmentace zákazníků, detekce podvodů a web mining.

Pomocí softwarových produktů se řešily v ČR především následující úlohy:

a) segmentace zákazníků s následnými analýzami,
b) analýza a predikce odchodu zákazníků.

Přístup zákazníků k řešení

Zákazníci dle zkušeností dodavatelů ocenili, pokud:

a) Nebyli nuceni ihned k nákupu relativně nákladných softwarových produktů a dodavatel realizoval první projekt(y) jako službu. Existují nicméně i zákazníci, kteří od začátku počítají s tím, že budou své další aktivity v oblasti získávání znalostí z dat realizovat interně.
b) Součástí řešení bylo i „skórování", tj. realizace procedur, které periodicky na základě zjištěných data miningových modelů generují skóre typu marketingový segment zákazníka, pravděpodobnost odchodu, k ukončení smlouvy atd. Taková skóre jsou pak využívána koncovými uživateli ve formě například on-line reportů v Intranetu, jejichž formou jsou odstíněny složité formy jejich vzniku.

Technologie a metody

Lze říci, že se ČR neliší od zahraničí: nejčastěji jsou používány softwarové produkty společností SPSS, SAS a StatSoft. Z pohledu metod jsme (v ČR i v celosvětovém měřítku) svědky prolínání statistických a speciálních - moderních metod a algoritmů. Ukazuje se, že přes přetrvávající místy nekompromisní argumenty statistiků a znalostních inženýrů vůči druhé straně je spojení obou oblastí nevyhnutelné a hlavně již probíhá, jak tomu naznačuje kromě vlastností dnešních softwarových produktů i přehled používaných metod v komerčních data miningových projektech v ČR:

- regresní a korelační analýza,
- analýza hlavních komponent (faktorová analýza),
- shlukování (tzv. clustering),
- rozhodovací stromy (algoritmy CHAID, CART, C 5.0),
- umělé neuronové sítě,
- logistická regrese (logit).

Vlastnímu modelování samozřejmě předchází transformace a normalizace dat (například logaritmická), imputace chybějících hodnot (missing value analysis) atd.

Využití softwarových prostředků zákazníky v ČR

V rámci vlastního průzkumu v oblasti odběratelů (zákazníků) ve [3] - uživatelů softwarových prostředků pro získávání znalostí z dat - bylo celkem osloveno 252 firem, z nichž na zaslaný dotazník odpovědělo 41. Návratnost dotazníků tedy činila 16,3 %, což je poměrně solidní výsledek vzhledem k citlivosti těchto údajů. Nejdříve však bylo osloveno 15 firem (odpověděly 4 firmy), na kterých byla odladěna dotazníková forma (formulace otázek,

Tab. 2: Rozdělení absolutních a relativních četností odpovědí na otázky dotazníku pro zákazníky

Otázka	Znění	Σ	%
1	**Jakým typem činnosti se Vaše firma zabývá?**	**41**	**100**
1a	výrobní	14	34,1
1b	obchodní	9	22,0
1c	služby	15	36,6
1d	finanční	3	7,3
1e	ostatní	0	0
2	**Jaká je velikost Vaší firmy?**	**41**	**100**
2a	do 50 zaměstnanců	16	39,0
2b	50 – 100 zaměstnanců	11	26,8
2c	100 – 500 zaměstnanců	10	24,4
2d	nad 500 zaměstnanců	4	9,8
3	**Zabýváte se problematikou DM?**	**41**	**100**
3a	ano	23	56,1
3b	ne	18	43,9
4	**Uvažujete do budoucnosti o zavedení produktu umožňující DM?**	**41**	**100**
4a	ano	18	43,9
4b	ne	23	56,1
5	**Pomocí kterých produktů provádíte v současnosti analýzu dat?**	**32**	**100**
5a	statistický software	12	37,5
5b	matematický software	1	3,1
5c	tabulkový procesor	18	56,3
5d	neprovádíme analýzy dat	1	3,1
6.	**Používáte některý z uvedených produktů umožňující DM?**	**41**	**100**
6a	Clementine	3	7,3
6b	Statistica Data Miner	2	4,9
6c	Intelligent Miner	2	4,9
6d	Enterprise Miner	1	2,4
6e	jiný	5	12,2
6f	žádný	28	67,7
7	**Jste s tímto produktem spokojeni, splnil Vaše očekávání?**	**13**	**100**
7a	ano, naprosto spokojeni	10	77,7
7b	ano, částečně	3	22,3

Pokračování tabulky 2

7c	ne	0	0
8	**Jste spokojeni se servisem dodavatele?**	**13**	**100**
8a	ano, naprosto spokojeni	8	61,5
8b	ano, částečně	5	28,5
8c	ne	0	0
9	**Přineslo zavedení produktu pro DM úspory v nákladech nebo zvýšení tržeb?**	**13**	**100**
9a	ano, finanční přínosy sledujeme	9	69,2
9b	nevíme, finanční přínosy nesledujeme	4	30,8
9c	ne	0	0

Zdroj: vlastní zpracování podle [3]

jejich pořadí apod.). Vyhodnocením zaslaných odpovědí pomocí vybraných statistických metod se zabývají následující odstavce.

V tabulce 2 jsou uvedeny celkové absolutní a relativní četnosti odpovědí na jednotlivé otázky dotazníku pro zákazníky.

K vyhodnocení dotazníku byly použity prostředky jak popisné tak i matematické statistiky (statistická indukce - testování statistických hypotéz). Použité metody jsou uvedeny v přílohách [3]. Pomocí popisné statistiky byly vypočteny absolutní a relativní četnosti odpovědí na otázky dotazníku.

Nyní následuje **přehled výsledků**. Produkty největších společností (dodavatelů) používá 8 z 41 firem (tj. 19,5 %), které odpověděly na zaslaný dotazník. Tento podíl bude zřejmě v základním souboru zřejmě nižší, poněvadž oslovené firmy byly vybrány na základě doporučení dodavatelů. Tento do jisté míry záměrný výběr byl proveden, jelikož pro zjištění odpovědí na otázky 7 až 9 dotazníku bylo potřeba oslovit firmy, které již implementovaly některý ze softwarových produktů. Tady je ze zjištěných údajů vidět, že trh s těmito produkty není ještě nasycen - mnoho firem používá k analýze dat pouze tabulkového procesoru. Také povědomí o data miningu není mezi firmami příliš silné (například v otázce 6e odpověděly 4 z 5 firem, že jako produkt umožňující data mining používají tabulkový procesor!). V následujících letech 2005-2006 se dá s vysokou pravděpodobností očekávat prudký nárůst implementací těchto produktů ve firmách.

Pokud již firmy používají daný produkt, jsou s ním spokojeny, splnil jejich očekávání. Rovněž firmy vyjádřily naprostou spokojenost se servisem dodavatelů. Také u 69 % firem zavedení tohoto produktu přineslo úspory v nákladech, popřípadě zvýšení tržeb. Zkušenosti firem s těmito produkty jsou vesměs pozitivní.

Dále byl dotazník podroben analýze závislostí dvou kvantitativních znaků pomocí (χ^2-testu v kombinační tabulce. Bylo použito statistické indukce, a to testování statistických hypotéz.

Statistická hypotéza je předpoklad o statistickém souboru, jehož platnost nemůžeme předem stanovit. Testování statistických hypotéz je součástí statistické indukce. Obecný postup při testování statistických hypotéz je uveden ve [3]. Těchto hypotéz existuje celá řada. Pro potřeby této práce byla zvolena hypotéza o nezávislosti dvou kvalitativních znaků.

Na základě dotazníku byly stanoveny klíčové otázky, u jejichž dvojic lze statisticky testovat existenci závislosti mezi jednotlivými odpověďmi. Toto jsou dvojice otázek, u nichž byla tato závislost testována pomocí (χ^2-testu v kombinační tabulce: otázky 1 a 3, 2 a 3, 1 a 4, 2 a 4, 3 a 4, 7 a 8 - bylo tedy provedeno celkem 6 testů. Při testování statistických hypotéz o nezávislosti se postupuje podle kroků, které jsou popsány v běžných učebnicích statistiky a také ve [4].

Na základě (χ^2-testů v kombinační tabulce byly testovány závislosti odpovědí na dvojice vybraných klíčových otázek. Pomocí matematické statistiky byly zjištěny následující závěry:

- Skutečnost, že se firma zabývá či nezabývá problematikou získávání znalostí z dat, **nezávisí** na typu činnosti, kterým se firma zabývá;

- Skutečnost, zda se firma hodlá či nehodlá zavést DM produkt, **nezávisí** na činnosti firmy;
- Skutečnost, že se firma zabývá či nezabývá problematikou získávání znalostí z dat, **nezávisí** na velikosti firmy;
- Skutečnost, že firma hodlá či nehodlá zavést DM produkt, **nezávisí** na velikosti firmy;
- Ochota firmy pořídit si některý DM produkt velmi **silně závisí** na tom, zda se firma touto problematikou vůbec zabývá;
- Celková spokojenost s produktem **závisí** na spokojenosti se servisem tohoto produktu.

5. Závěr

Úspěšnost prosazení výsledků získávání znalostí z dat často souvisí s úspěšností modelování. Dle zkušeností dodavatelů lze například dosáhnout velmi vysoké přesnosti predikce odchodu zákazníků ke konkurenci (např. 80 %) - taková situace důvěru v užitečnost softwarového produktu značně posiluje.

V ČR dodavatelé očekávají postupný nárůst komerčních projektů ze stávajících jednotek na desítky v průběhu nejbližších tří let. Největší rizika a faktory úspěchu softwarových produktů jsou podle dodavatelů tyto:

- dobrá analýza a kvalita vstupních dat a
- zajištění silného obchodního sponzora projektu.

Literatura:

[1] EDELSTEIN, H. A. *Introduction to Data Mining and Knowledge Discovery.* 1st ed. Potomac, Maryland: Two Crows Corporation, 1999. ISBN 1892095025

[2] FAYYAD, U. M., PIATETSKI-SHAPIRO, G. *Advances in Knowledge Discovery and Data Mining.* 1st ed. Cambridge: MIT Press, 1996. ISBN 0262560976

[3] KLÍMEK P. Získávání znalostí z dat (dat mining). *Disertační práce.* Zlín: UTB, 2003.

[4] PARR RUD, O. *Data mining (Praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM).* 1. vyd. Praha: Computerpress, 2002. ISBN 80-7226-577-6

[5] WADDELL, D., SOHAL, A.S. Forecasting: The Key to Managerial Decision Making. *Management Decision*, 1994, r. 32, č. 1, s. 41-49. ISSN 0025-1747

[6] [online]. KDNuggets Inc.[cit. 4. 5. 2003]. Dostupné na WWW: <http://www.kdnuggets.com/>

[7] [online]. SAS Institute Inc. [cit. 9. 6. 2003] Dostupné na WWW: <http://www.sas.com>

Ing. Petr Klímek, Ph.D.
Univerzita Tomáše Bati ve Zlíně
Fakulta managementu a ekonomiky
Ústav informatiky a statistiky
klimek@fame.utb.cz

Doručeno redakci: 17. 4. 2005
Recenzováno: 16. 6. 2005
Schváleno k publikování: 7. 7. 2005

SUMMARY

DATA MINING AND ITS USE

Petr Klímek

Databases today can range in size into the terabytes. Within these masses of data lies hidden information of strategic importance. But when there are so many trees, how do you draw meaningful conclusions about the forest? The newest answer is data mining, which is being used both to increase revenues and to reduce costs.

There are two keys to success in data mining. First is coming up with a precise formulation of the problem you are trying to solve. A focused statement usually results in the best payoff. The second key is using the right data. After choosing from the data available to you, or perhaps buying external data, you may need to transform and combine it in significant ways.

Data mining offers great promise in helping organizations uncover patterns hidden in their data that can be used to predict the behavior of customers, products and processes. However, data mining software tools need to be guided by users who understand the business, the data, and the general nature of the analytical methods involved. Realistic expectations can yield rewarding results across a wide range of applications, from improving revenues to reducing costs.

Building models is only one step in knowledge discovery. It's vital to properly collect and prepare the data, and to check the models against the real world. The „best" model is often found after building models of several different types, or by trying different technologies or algorithms.

Choosing the right data mining products means finding a tool with good basic capabilities, an interface that matches the skill level of the people who'll be using it, and features relevant to your specific business problems.

Key words: *knowledge discovery in databases, data mining, software products, statistics, methods for data mining*