Ústav informačních studií a knihovnictví
Filozofická fakulta Univerzity Karlovy

Dobývání znalostí z databází

Seminární práce na předmět
Matematické a informatické modely v ontologii

LS 2003/2004

Linda Skolková
2. ročník bak. studia
program *Informační studia a knihovnictví*
kontakt: skolkova@chello.cz

DOSTUPNOST PRÁCE

Ø práce je v elektronické podobě dostupná na URL:
http://www.sweb.cz/lin.skl/kdd.pdf

OBSAH

1. Úvod

Rychle se rozvíjející informační a komunikační technologie (ICT) umožňují ukládat a komunikovat velké množství dat nejrůznějšího charakteru. Přesto často nejsou schopny splnit všechny požadavky, se kterými se na ně obracíme (nebo bychom se mohli obracet). Rozvoj oblastí *dobývání znalostí v databázích* nebo úžeji *dolování dat* svědčí o tom, že situace se začíná měnit. Tato práce si proto klade za cíl podat vysvětlení některých pojmů, s nimiž se v těchto a souvisejících oblastech setkáváme, a případně také při ústní prezentaci vyvolat diskusi o korektnosti či nekorektnosti používaných termínů z pohledu *informační vědy*.

2. Dobývání znalostí z databází

O problematice dobývání znalostí z databází (**Knowledge Discovery in Databases**, též zkráceně **KDD**) se začalo se začalo výrazněji diskutovat počátkem 90. let 20. století, přičemž první impulzy přišly z konferencí o umělé inteligenci, které se konaly v USA. Podle [4] byla fráze *knowledge discovery in databases* vytvořena na prvním workshopu KDD již v roce 1989.

Znalosti se však nezačaly dobývat na zelené louce, ale byly využity již existující metody, postupy či technologie, zejména pak:

- **metody strojového učení** (součást umělé inteligence)
- **databázové technologie** (prostředek uchování rozsáhlých dat a vyhledávání v nich)
- **statistické postupy** (prostředek modelování a analýzy závislostí v datech)

Tyto disciplíny se předtím vyvíjely nezávisle. S nárůstem objemu zpracovávaných dat a se vznikem potřeby shromážděná data využívat pro podporu strategického rozhodování ve firmách se vytvořilo příhodné prostředí pro jejich propojení.

KDD můžeme podle definice ve [4] z roku 1996 i podle následných prací definovat jako **netriviální získávání implicitních, dříve neznámých a potenciálně užitečných informací z dat**. KDD bývá chápáno také jako **interaktivní a iterativní proces tvořený kroky selekce, předzpracování, transformace, vlastního *dolování* (data mining) a interpretace.**

V [1] si sice můžeme v předmluvě přečíst následující motto:
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.[1] Nepochybně však platí, že čím lépe jsou data v počáteční fázi připravena, tím lepších konečných výsledků dosahujeme. Jinak samozřejmě dochází ke známému jevu *garbage in → garbage out.* Na čištění dat dnes existují nejrůznější profesionální nástroje. Samotný proces čištění dat probíhá podle [19] v těchto fázích:

- **analýza** (*investigation*) – zkoumání charakteru vstupních dat (formáty a typy záznamů, identifikace obecných údajů v záznamech, frekvenční analýza atd.)
- **standardizace** – jednotná reprezentace informací vhodná pro další zpracování (identifikace jednotlivých položek v záznamu, jejich transformace, konverze apod.)
- **obohacení** (*enrichment*) – kompletace dat (včetně kontroly správnosti) a jejich rozšíření z jiných interních/externích zdrojů (např. číselníků)

[1] Christieho-Daviesův teorém, Murphyho zákony.

- **hledání souvislostí** (*linking*) – identifikace vazeb mezi individuálními záznamy, seskupování (agregace) údajů, odstraňování duplicit či dokonce multiplicit
- **integrace** – umožnění implementace jednotného procesu kontroly a zušlechťování kvality dat

Impulzem pro zahájení procesu dobývání znalostí je nějaký **reálný problém**. **Cílem** procesu dobývání znalostí je **získat co nejvíce relevantních informací vhodných k řešení problému.**

Jako příklady problémů si můžeme uvést například nalezení skupin zákazníků obchodního domu nebo skupin klientů banky, jimž by bylo možné nabídnout speciální služby. Jedná se o tzv. *segmenty trhu*.

Řešení problému probíhá postupně v několika etapách, jimiž jsou:

1) **vytvoření řešitelského týmu**, mezi jehož členy by se měly objevit tyto osoby.
   - expert na řešenou problematiku
   - expert na data
   - expert na metody KDD
2) **specifikace problému**
3) **získání** veškerých **dostupných dat** (i externích), která mohou být použita pro řešení problému
4) **výběr metod analýzy dat**,[2] například:
   - klasifikační metody
   - různé klasické metody explorační analýzy dat
   - metody pro získávání asociačních pravidel[3]
   - rozhodovací stromy
   - genetické algoritmy
   - bayesovské sítě
   - neuronové sítě
   - hrubé množiny (*rough sets*)
   - metody vizualizace
5) **předzpracování dat** (příprava dat do formy vyžadované pro aplikaci vybraných metod; odstranění odlehlých a doplnění chybějících hodnot)
6) **dolování dat** (*data mining, modeling, analysis*; aplikace vybraných analytických metod pro vyhledávání zajímavých vztahů v datech, přičemž metody data miningu jsou obvykle aplikovány vícekrát)
7) **interpretace** (zpracování množství výsledků jednotlivých metod – některé z výsledků jsou z hlediska uživatele nezajímavé nebo samozřejmé)

Mezi úlohy KDD patří:

- **klasifikace nebo predikce** – cílem je nalézt znalosti použitelné pro klasifikaci nových případů, přednost má pokrytí na úkor jednoduchosti; rozdíl mezi klasifikací a predikcí je v tom, že u predikce hraje důležitou roli čas
- **deskripce** – cílem je nalézt dominantní strukturu nebo vazby, které jsou skryté v daných datech

[2] Obecně rozlišujeme analýzu dat na *konfirmační*, kdy se pokoušíme vyvrátit/potvrdit dříve formulovanou hypotézu, a *explorační*, kdy hledáme zajímavé souvislosti mezi daty.
[3] Pozornost si zasluhuje metoda explorační analýzy dat, která je známa pod zkratkou GUHA (General Unary Hypotheses Automation), kterou cca před třiceti lety vytvořila skupina českých vědců kolem P. Hájka.

- **hledání *nuggetů*** – požadujeme nové, překvapivé znalosti, které nemusí plně pokrývat daný koncept

Úlohy dobývání znalostí lze nalézt v celé **řadě aplikačních oblastí**, mezi něž například patří:

- segmentace a klasifikace klientů banky či pojišťovny
- analýza nákupního košíku (*market basket analysis*)
- analýza důvodu změny poskytovatele služeb (poskytovatel připojení k internetu, mobilní operátor apod.)
- analýza příčin poruch v telekomunikačních sítích
- predikce vývoje kurzů akcií na burze
- rozbor databáze pacientů v nemocnici

K nejpoužívanějším metodikám patří metodika 5A, SEMMA a CRISP-DM. Metodika ***5A*** zahrnuje těchto pět kroků:

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

Metodika ***SEMMA*** (zde odkazuji na *Enterprise Miner*, softwarový produkt firmy SAS) se rovněž skládá z pěti kroků:

- **Sample** - vybírání vhodných objektů
- **Explore** - vizuální explorace a redukce dat
- **Modify** - seskupování objektů a hodnot atributů, datové transformace
- **Model** - analýza dat
- **Assess** - porovnání modelů a interpretace

Metodika ***CRISP-DM (CRoss-Industry Standard Process for Data Mining)*** se snaží nalézt univerzálně použitelný postup pro dolování dat. Z použité literatury lze také usuzovat, že terminologie uváděná v rámci této metodiky tvoří v oboru již standard, alespoň ve své anglické jazykové mutaci. Metodika CRISP-DM uvádí následující etapy:

- **porozumění problematice** (*Business Understanding*)
- **porozumění datům** (*Data Understanding*)
- **příprava dat** (*Data Preparation*)
- **modelování** (*Modeling*)
- **vyhodnocení výsledků** (*Evaluation*)
- **využití výsledků** (*Deployment*)

Hovoříme-li o dobývání znalostí v databázích a o dolování dat, je vhodné připomenout si ještě existenci následujících dvou oblastí:

- **knowledge discovery in texts** (resp. **text mining**), což není nic jiného než speciální typ úlohy dobývání znalostí z databází (hlavním problémem je vhodná reprezentace původního nestrukturovaného textového dokumentu)
- **web mining**, který bývá dále členěn na *web content mining*, *web structure mining* a *web usage mining*

3. Datový sklad

Bez. ***datového skladu (data warehouse)*** se při dolování dat a návazně dobývaní znalostí v databázích neobejdeme, proto si jeho charakteristické rysy nyní detailněji představíme.

Datový sklad představuje ucelené řešení, které poskytuje jednak prostředky pro **ukládání** dat, jednak sadu nástrojů pro jejich **analýzu**.[4] Klíčovou roli v datovém skladu hrají relační databáze. Pro datový sklad je dále typické, že je neustále rozšiřován bez redukce obsahu (toto tvrzení však neplatí ve všech případech, viz [10]). V datových skladech se setkáváme také s tzv. **datovými pumpami** neboli ETL (extract, transform & load) nástroji, které umožňují plnění databází daty (nejprve **získávají** data ze vzájemně nekompatibilních zdrojů, poté je **transformují** do nových struktur a konečně je **ukládají** do datového skladu).

Datový sklad je **orientován** na subjekty, kterými se firma zabývá (zákazník, dodavatel, produkt, aktivita) a uchovává data pro podporu rozhodování na manažerské úrovni. Lze jej také označit za **integrovaný** (jedná se o integraci a sjednocení dat) a **časově proměnný** (všechna data v datovém skladu představují *časový snímek* dat z produkčních databází sejmutý v určitém okamžiku[5]). Datový sklad je aktualizován offline v určitých časových intervalech (např. měsíčně, čtvrtletně, ročně) a analyzován odděleně od produkčních bází (ty uchovávají data potřebná pro operativní řízení), takže případný nešetrný zásah do datového skladu neovlivní operativní řízení firmy. Datový sklad je rovněž **stálý**, což znamená, že dotazy, které do datového skladu směřují uživatelé-analytici, nezpůsobují změnu uložených dat. Data uložená v datovém skladu představují **neutrální datový prostor**, který není vytvářen s myšlenkou konkrétních analýz. Z toho důvodu se doporučuje vytvářet v návaznosti na datový sklad řadu specializovanějších **datových tržišť (data marts)**, kam se z datového skladu přesunou data relevantní pro určitý typ analýz (resp. pro určité oddělení firmy).

V [12] nalezneme osm základních doporučení, kterých bychom se při tvorbě datového skladu měli držet:

1. **Začněte v malém** (např. data mart pro jedno oddělení, výhodou je mj. rychlost implementace)..
2. **...ale myslete ve velkém**
3. **Stanovte cíle a vyčíslete přínosy**
4. **Zapojte nejvyšší vedení**
5. **Přehnaný perfekcionismus není přínosem**
6. **Vyberte systém, který se přizpůsobí potřebám uživatelů**
7. **Ujistěte se, že komponenty datových skladů skutečně spolupracují** (pro datové sklady totiž zatím neexistují standardy v podobném rozsahu jako v ostatních oblastech IT: zejména TCP/IP pro sítě, SMTP pro elektronickou poštu a HTML a Java pro web)
8. **Váha mezilidských vztahů**

K poslednímu uvedenému bodu dodejme, že s datovými sklady je nerozlučně spjato sdílení informací, což pro mnoho uživatelů může znamenat ztrátu kontroly. S tímto

[4] Je užitečné uvědomit si také vzájemné vztahy mezi nástroji OLAP (*Online Analytical Processing*), datovými sklady a dolováním dat. Nástroje OLAP umožňují analýzu (a vizualizaci) dat o firmě, kdežto datový sklad je místem, kde jsou analyzovaná data uložena. Zatímco v případě používání nástrojů OLAP získáváme z dat sumární charakteristiky na zvolené podrobnosti pohledu, při dolování dat hledáme v datech zajímavé souvislosti.
[5] V tomto ohledu se podobá internetovému archivu dostupnému z www.archives.org.

fenoménem, který autor v [12] výstižně označuje jako *datový provincialismus*, se je potřeba vyrovnat hned zpočátku.

Z provedených studií však zároveň vyplývá, že jestliže uživatelé získají přístup k informacím z nového skladu, zanedlouho zjistí, že jich potřebují mnohem více a na mnohem větší úrovni podrobnosti.

Z [20] je patrné, že nejde především o samotnou implementaci datového skladu jako takového, ale zejména o to, že by jím organizace měla řešit skutečné potřeby, a to nejenom své, ale především svých zákazníků. Proto je nezbytné, aby datové sklady poskytovaly využitelné informace, tedy informace, které přinesou v krátké době měřitelné finanční výsledky (zvýšení obratu a zisku, snížení nákladů, udržení zákazníků apod.).

4. Závěrečné zamyšlení

Množství dat, které nás obklopuje, se čím dál více zvětšuje. S tím se také začíná stále více zviditelňovat ten fakt, že nejde primárně o shromažďování terabytů dat, ale jedná se hlavně o jejich interpretaci a praktické využití.

Uveďme si příklad z nedávného článku z Newsweeku[6] - hovoří se zde o tom, že máme stále lepší technické prostředky pro zpracování velkého množství dat, avšak často si od těchto prostředků slibujeme nereálné výsledky. Obyvatel Londýna je sice každodenně přibližně třistakrát vyfotografován, ovšem záznamy o teroristech, kvůli kterým se toto vše děje, se stejně v příslušných databázích nevyskytují. Podobně jsou na tom také různé systémy *face recognition*.[7]

Jinými slovy můžeme říci, že situaci vystihuje citát od Daniele Rizzi z Evropské komise (objevil se v rámci článku *EU aims to improve net searching*[8]): **Today's software makes us information rich but insight poor.** Může nám posloužit jako doklad dvou problémů – jednak konstatování omezenosti současných softwarových produktů, jednak rozkolísané terminologie (je možno *insight* vnímat jako téměř synonymní výraz k *knowledge* nebo je významově někde mezi *knowledge* a *wisdom*?).[9]

[6] Taking a closer look. *Newsweek.* March 8, 2004, s. 44-48.
[7] GILES, Jim. Smiles reveal secrets to security cameras. [online]. [cit. 2004-03-31]. Dostupné z World Wide Web: <http://www.nature.com/nsu/040322/040322-13.html>.
[8] EU aims to improve net searching. *BBC News* [online]. Wednesday, 17 March, 2004, 09:21 GMT. [cit. 2004-04-02]. Dostupné z World Wide Web: <http://news.bbc.co.uk/1/hi/technology/3516088.stm>.
[9] Nesmíme zároveň zapomenout, že i vzhledem k dnešním možnostem zůstávají nejrůznější úlohy výpočetně náročnými. Řešením může být tzv. *grid computing* (patrně nejznámějším příkladem využití této metody je projekt Seti@Home, mezi nejužitečnější by naopak patřilo hledání léku na rakovinu) nebo tzv. *utility computing* (využívá výpočetního výkonu serverů, které by jinak byly v danou dobu v nečinnosti). Blíže viz HLAVENKA, Jiří. Vyšší matematika. CONNECT! 2004. Roč. 9, č. 3, s. 35-36.

5. Použité zdroje

[1] BERKA, Petr. *Dobývání znalostí z databází.* Vyd. 1. Praha : Academia, 2003. 366 s. ISBN 80-200-1062-9.

[2] BIANCHI-BERTHOUZE, Nadia; HAYASHI, Tomofumi. Subjective interpretation of complex data : requirement for supporting Kansei mining process. *Mining multimedia and complex data.* Ed. O.R. Zaïane et al. Berlin ; Heidelberg : Springer, 2003, s. 1-17. LNAI 2797.

[3] BRADLEY, P.S. Data mining as an automated service. *PAKDD 2003.* Ed. Whang, K.-Y. et al. Berlin ; Heidelberg, 2003, s. 1-13. LNAI 2637.

[4] FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. American Association ofr Artificial Intelligence, 1996

[5] FORMAN, George. Choose your words carefully : an empirical study of feature selection metrics for text classification. *PKDD.* Ed. T. Elomaa et al. Berlin ; Heidelberg : Springer, 2002, s. 150-162. LNAI 2431.

[6] HABÁŇ, Jaromír. Co se skrývá pod povrchem : technologie datových skladů. *EBiz : byznys, management a technologie.* Prosinec 2003, s. 44-46. ISSN 1213-063X.

[7] HABÁŇ, Jaromír. Poklady skryté v datech. *EBiz : byznys, management a technologie.* Prosinec 2003, s. 47—48. ISSN 1213-063X.

[8] HABÁŇ, Jaromír. Trhy a trendy : aktuální přehled řešení datových skladů na českém trhu. *EBiz : byznys, management a technologie.* Prosinec 2003, s. 52—58. ISSN 1213-063X.

[9] HIPP, Jochen; GÜNTZER, Ulrich; NAKHAEIZADEH, Gholamreza. Data mining of association rules and the process of knowledge discovery in databases. In *Advances in data mining 2002.* Ed. P. Perner. Berlin ; Heidelberg : Springer, 2002, s. 15-36. LNAI 2394.

[10] KUČERA, Milan. Monitorování data warehouse. *Data security management : časopis o bezpečnosti, správě a řízení rizik informačních systémů.* 2001. Roč. 5, č. 4, s 44-45.

[11] KUČERA, Milan. Správné řešení není vždy snadné I-III. *Data security management : časopis o bezpečnosti, správě a řízení rizik informačních systémů.* 2000. Roč. 4, č. 4, s. 36-38, č. 5, s. 36-38, č. 6, s. 40-43.

[12] KYJONKA, Vladimír. Jak začít s prvním datovým skladem. *IT system.* 2004, č. 3, s. 36-37.

[13] MARTÍNEK, Tomáš; RADEMACHER, Malte. Information lifecycle management. *IT system.* 2004, č. 3, s. 40-41.

[14] PANEC, Zdeněk. Čtvrtý článek řetězu. *Business World.* Roč. 4. Červenec 2003, č. 7. s. 18-19.

[15] PARK, Seong-Bae; ZHANG, Byoung-Tak. Large scale unstructured document classification using unlabeled data and syntactic information. *PAKDD 2003.* Ed. Whang, K.-Y. et al. Berlin ; Heidelberg, 2003, s. 88-99. LNAI 2637.

[16] PIRKL, David. Dolování dat. *Business World.* Roč. 4. Červenec 2003, č. 7. s. 14-17.

[17] PŮLPÁN, Jaroslav. Hledání skrytých souvislostí. *EBiz : byznys, management a technologie.* Prosinec 2003, s. 49—51. ISSN 1213-063X.

[18] SHAH, Harshit S. et al. Mining eBay : bidding strategies and shill detection. In *WEBKDD 2002.* Ed. O.R. Zaïane et al. Berlin ; Heidelberg : Springer, 2003, s. 17-34. LNAI 2703.

[19] SCHILLER, Martin. Kvalitní data znamenají kvalitní rozhodování. *IT system.* 2004, č. 3, s. 38-39.

[20] SKOPAL, Vít. Neimplementujte datový sklad. *IT system.* 2004, č. 3, s. 47.

[21] VICKERY, Brian. Knowledge discovery from databases : an introductory review. *Journal of Documentation.* March 1997, vol. 53, no. 2, s. 107-122.

[22] VRÁNA, Jan. Správa rozsáhlých datových systémů. *IT system.* 2004, č. 3, s. 44-46.

[23] Využití dolování dat při řízení vztahů se zákazníky : role data miningu v CRM strategii. *IT system.* 2004, č. 3, s. 42-43.

[24] YUN, Bo-Hyun; LIM, Myung-Eun; Park, Soo-Hyun. An integrated system of mining HTML texts and filtering structured documents. *PAKDD 2003.* Ed. Whang, K.-Y. et al. Berlin ; Heidelberg, 2003, s. 350-355. LNAI 2637.