

## DATA MINING A JEHO UPLATNĚNÍ PŘI POZNÁVÁNÍ KNIHOVNICKÝCH JEVŮ A ZÁKONITOSTÍ

PhDr. Beáta Sedláčková \*

V současné společnosti je tvorba informací a znalostí základním krokem úspěšného managementu, výraznou podporou osobního růstu a profesního postupu. Z těchto důvodů je data mining mnoha odborníky považován za jednu z velmi perspektivních a atraktivních činností budoucnosti ve všech oblastech a různých úrovních lidské aktivity.

Je to proces transformace dat přes informace ke znalosti, přesněji řečeno k akční znalosti. Příspěvek zmiňuje data mining jako nový nástroj zpracování a využívání informací a sleduje možnosti aplikace v knihovnicko-informační oblasti.

Pojem **data mining** definují různí autoři různě. Jednou z nejjednodušších a nejkratších může být definice: Data-mining je hledání hodnotných informací ve velkých objemech dat. O něco složitěji zní definice: Data mining je netriviální proces zjišťování platných, neznámých, potenciálně užitečných a snadno pochopitelných závislostí v datech (Vítek, 2002). Tato definice vystihuje nosnou myšlenku pojmu data mining, jelikož podstatnou částí procesu data miningu je datové modelování. Jde více o to „dolovat z dat“ než o „dolování dat“.

Z praktického hlediska lze říct, že data mining je obecná metodologie, která se používá k řešení různých problémů, sledováním proudu dat, monitorováním procesu s predikcí vývoje či selhání, odhadem budoucího chování individuálních případů, odhalením neurčitých cílů apod. Jedná se o systematickou činnost, o metodu průzkumu v datech, která umožňuje získat vzhled do věcného problému. Tento vzhled je užitečný už samotný, ale také přináší další výhody jako například schopnost předvídaní, protože data miningové příslušenství vytváří „sadu nástrojů“, které se používají různými a často překvapivými způsoby pro vyřešení problému.

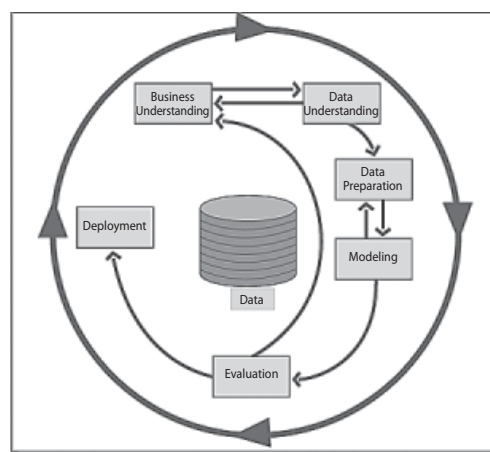
Zmíním jednu nejrozšířenější představu o data miningu, a to, že ho může dělat pouze IT expert. Ve skutečnosti v data miningu je rozhodující znalost problematiky, protože jinak produkuje nesmyslné nebo nepoužitelné výsledky. Rozsáhlá znalost problematiky se ovšem velmi zřídka kombinuje se znalostmi informačních technologií. Proto je nanejvýš vhodné, aby data miningové nástroje byly přijatelné a srozumitelné více pro odborníky než experty na technologie. Je na místě objasnit potenciálním uživatelům, že data mining poskytuje spíše vzhled a prospěšná řešení než matematickou jistotu.

### METODOLOGIE A TECHNIKY DATA MININGU

Data miningový proces vyžaduje určité množství zdrojů, od lidských přes datové až po softwarové a to vyžaduje značné peněžní prostředky. Jedním ze způsobů, jak šetřit finanční prostředky, je provádět tuto činnost standardi-

zovaným způsobem. Takové řešení nabízí souhrnná data miningová metodologie CRISP-DM (*Cross Industry Standard Proces for Data Mining*), která obsahuje návody krok po kroku, úkoly a cíle pro každou část celého procesu a tím umožňuje provádět rozsáhlé data miningové projekty rychleji, efektivněji s menšími náklady prostřednictvím osvědčených postupů, co umožňuje vyhnout se běžným chybám.

Vývoj metodologie byl zahájen jako projekt Evropské komise definující model standardního postupu při vytváření data miningových projektů. Metodologie CRISP-DM rozděluje celý proces data miningového projektu do šesti základních etap, v rámci nichž rozlišuje další kroky. Těmito etapami jsou: definování cílů; porozumění datům; příprava dat; modelování; hodnocení výsledků; implementace vytvořeného modelu (viz obrázek). Není záměrem příspěvku zabývat se jednotlivými etapami, tyto jsou podrobně popsány v příslušné odborné literatuře (Vítek, 2002).



Metodologie CRISP-DM  
(CRISP-DM: Process Model. [online]. London : SPSS,2007)

Zvláštní zmínku si zaslouží techniky využívající vizualizaci pro data mining, které dovolují provádět pozorování bez předešlé představy. To znamená, že nemusíme vědět, co hledáme, místo toho necháváme data, aby nám ukázala, co je v nich důležitého a zajímavého. Během vizualizační techniky můžeme rychle vidět zajímavou strukturu a různými

\* Slezská univerzita Opava, Česká republika  
e-mail: beata.sedlackova@fpf.slu.cz

né datové závislosti.. Právě použití vizualizační techniky umožňuje objevení zajímavého vzoru nebo trendu, který by jinak zůstal nepovšimnut. Výhody vizualizace tkví v kognitivní schopnosti mozku. Zatímco při tradiční analýze jsou vyšší nároky kladeny na pozornost a paměť, při vizualizaci se využívají poznávací schopnosti. Mozek je schopný během několika milisekund rozeznat v obrázku či grafu důležité znaky, zatímco ve sloupci čísel by mu to trvalo mnohem déle, protože by každou jednotlivou položku musel zpracovávat samostatně.

## DATOVÝ SKLAD A BIBLIOMINING

Zatímco u nás se v současné době uplatňuje data mining především v takových oblastech, jako jsou například bankovníctví pro detekci a prevenci pojišťovacích podvodů, v telekomunikacích pro analýzu zákaznických skupin se sklony k migraci mezi poskytovateli služeb, v medicíně pro diagnostiku a terapii na základě známých symptomů nebo v e-businessu pro získávání dat o prohlížení webových stránek uživatelů a k analýze jejich chování (web mining), v zahraničních odborných časopisech se objevují první zmínky a případové studie o aplikaci data miningu v našem oboru a v této souvislosti se objevuje i pojem **bibliomining**, kterého definice zní „application of data mining for libraries“ (Nicholson, 2006). Autor zvažuje smysl a účel datového skladu v knihovnách, vymezuje pojem bibliomining, který chápe jako kombinaci bibliometrie a data miningových technik pro výzkum a porozumění optimálního fungování knihoven a jejich služeb, a nastiňuje koncepci bibliominingu.

V souvislosti s data miningem nelze nezmínit datový sklad, který sice není nezbytnou podmínkou data miningu, jak je obecně mylně rozšířeno (Khabaza, 2002), nicméně s data miningem úzce souvisí. Datovou základnou jakékoliv organizace je datový sklad (*Data Warehouse*). Zvětšující se množství zpracovávaných a ukládaných dat vedlo ke snahám o sjednocení metodiky přístupu k mnohdy nehomogenním datům často rozptýleným v různých aplikacích. A navíc jsou zde požadavky na uchování dat, které vytváří něco takového, jako je *institucionální paměť*.

Tak jako jiné instituce a organizace i knihovna si budeje svoji vlastní provozní databázi, která podchycuje veškeré knihovní aktivity. A stejně tak jako ostatní instituce i knihovny musí hledat rovnováhu mezi ochranou soukromí svých čtenářů na jedné straně a zachováním historie knihovních transakcí, které jsou důležité pro zmapování a vyhodnocení činnosti knihovny a zdůvodnění jednotlivých informačních služeb. Datový sklad je databáze oddělená od provozního systému, která obsahuje vyčištěnou verzi provozních údajů, upravených pro analýzu. Může být využitý pro bibliomining, kterým lze vyprofilovat informační potřeby a informační chování uživatelů a odhalit vzory informačního chování pro vyhodnocení a opodstatnění jednotlivých typů služeb knihoven. A navíc je to možnost uchovávat informace důležité pro tvorbu rozhodnutí. Prozatím mnoho knihovníků neuvazuje o tom, že takové data mohou být velmi užitečná pro potřebu rozhodování. Tradiční vyhodnocování celkové činnosti a služeb knihovny je zaměřeno na obecné shrnutí a průměry.

## TOTAL QUALITY MANAGEMENT A BIBLIOMINING

Trendem řízení v současné době je koncepce *Total Quality Management* (TQM). Jedná se o otevřený manažerský systém zahrnující vše pozitivní, co může být použito pro rozvoj organizace. Základní principy TQM jsou orientace na zákazníka, vedení lidí a týmová práce, partnerství se spolupracujícími organizacemi, rozvoj a angažovanost lidí, orientace na procesy, zlepšování a inovace, měřitelnost výsledků a odpovědnost vůči okolí. Kromě přístupů zabezpečování jakosti vycházejících z požadavků norem ISO řady 9000 existuje řada názorových proudů TQM, společné rysy ovšem lze odvodit již z názvu:

- *total* – jde o úplné zapojení všech pracovníků organizace, ve smyslu zahrnutí všech činností,
- *quality* – jde o pojetí jakosti zahrnující nejen výrobek či službu, ale i proces, činnost,
- *management* – řízení je zahrnuto jak z pohledu strategického, taktického a operativního řízení, tak i z pohledu manažerských aktivit jako jsou plánování, motivace, vedení, kontroly atd.

V současném tržně orientovaném prostředí knihovnic na řídicí a rozhodující úrovni musí rozumět více tomu, jak jsou jejich služby a zdroje využívány, aby mohli úspěšně konkurovat dalším poskytovatelům informačních služeb. Datový sklad a bibliomining umožní odpovědným pracovníkům knihoven mít k dispozici rozsáhlé údaje pro zdůvodnění a získání finančních prostředků a může se stát také mocným zdrojem manažerského informačního systému v knihovně. Pro řídicí pracovníky poslouží zprávy vyhotoveny na základě těchto dat jako klíč k variabilitě, kterou je možné identifikovat a monitorovat *chod a puls knihovny* a dělat rozhodnutí založené na doložených modelech a evidenci z minulých let.

### LITERATURA

1. BIBLIOMINING: *Data Mining for Libraries*. [online]. Ed. Scott Nicholson. Syracuse: School for Information Studies, 2005. [cit. 10-11-2007]. Dostupné z <http://www.bibliomining.com/>
2. CHUI-Cheng Chen. *Using Data Mining Techniques to Discover Personalized Book Recommendation for Library*. Journal of Educational Media and Library Sciences, 2005, Vol. 43, Issue (No) 1, pp. 87 – 107.
3. KHABAZA, Tom. *Hard Hats for Data Miners: Myth and pitfalls of data mining*. London : SPSS Advanced Data Mining Group, 2002.
4. NENADÁL, Jaroslav. *Aplikace norem ISO řady 9000 ve službách*. [online]. Ostrava : VŠB – technická univerzita, 2007. [cit. 10-11-2007]. Dostupné z <http://spbi.hgf.vsb.cz/html/clan26.htm>
5. NICHOLSON, Scott. *The basis for bibliomining: Frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services*. Information Processing and Management, May 2006, Vol. 42, Issue 3, pp. 785 – 804.
6. RUD, Olivia Parr. *Data Mining: praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků*. Praha : Computer Press, 2001.
7. SYSTÉM managementu jakosti. [online]. Praha: Nexus Group, 2004. [cit. 10-11-2007]. Dostupné z <http://businessinfo.cz/cz/clanek/kvalita-jakost/system-managementu-jakosti/1000513/16924/#b02>
8. VÍTEK, Martin. *Metodologie CRISP-DM*. [online]. Praha : Fakulta informačních technologií VUT, 2002. [cit. 10-11-2007]. Dostupné z <http://datamining.xf.cz/view.php?cisloclanku=2002102807>