

Knihovnické systémy a standards (VIKBA10)

7. Bibliomining

Martin Krčál

Data mining

Data mining

- = dolování/dobývání/vytěžování dat
- různé výklady a definice
- definice
 - ❖ získávání dat přímo z databáze nebo datového skladu
 - ❖ aplikace statistických a logických metod na data
 - ❖ uplatnění v managementu při
 - rozhodovacích procesech
 - strategickém řízení

Data mining

- jde o samostatnou metodu nebo je součástí knowledge discovery in databases (KDD)???
- v odborné literatuře často oboje
- původně z ekonomie
- dnes různé obory
 - ❖ IT a technologie, obchod a bankovníctví, medicína, vzdělávání, informační věda a knihovnictví

Souvislost s ISK

- datová analytika
- informační chování a potřeby
- organizace znalostí
- expertní systémy
- přesahy do IT, statistiky a ekonomie

Historie data miningu

- počátky v 60. letech s rozvojem IT
 - ❖ první pokusy o analýzu uložených dat
- 80. léta – počátky expertních databázových systémů
 - ❖ vyšší výkon PC
 - ❖ systematictější data mining
 - ❖ spíše vyhledávání vzájemných vztahů ve velkých datasetech
- 1989 – použit pojem knowledge discovery in databases

Historie data miningu

- 90. léta – samostatný obor aplikované vědy
 - ❖ nové statistické metody
 - ❖ přesun z akademické sféry do komerčního prostředí
 - ❖ využití ve velkých firmách provozující velké systémy s množstvím dat
- dnes big data
 - ❖ obrovské datasety a jejich analýza, výzkumy chování uživatelů, rozhodovací proces, umělá inteligence a strojové učení

Bibliomining

Bibliomining



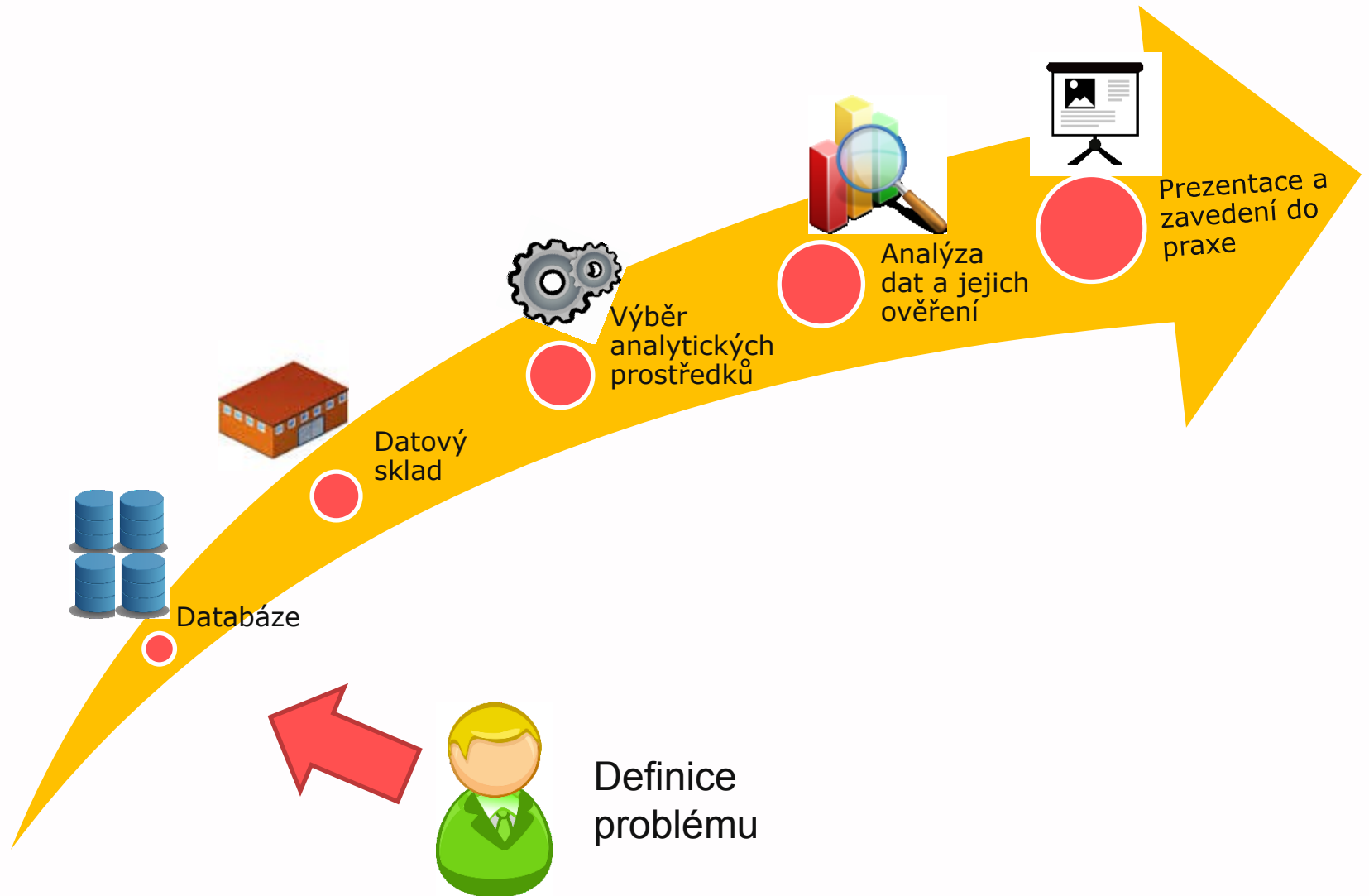
■ Scott Nicolson a Jeffrey Stanton

- ❖ Syracuse University
- ❖ weby
 - <http://www.bibliomining.com>
 - <http://www.scottnicholson.com>
- ❖ diskuze o dataminingu 2003 → analýza dat v knihovních systémech
- ❖ library datamining → obvykle vyhledává programovací knihovny pro datamining, snaha odlišit tuto oblast → pojem bibliomining

Bibliomining

- analýza dat produkovaných knihovními službami a systémy
- hledání zákonitostí a aplikace statistických metod na data v AKS
- uplatnění v managementu knihoven, v rozhodovacích procesech

Proces bibliominingu



1. Definice problému

- specifický požadavek knihovny
- požadavek managementu
- co chceme zjistit?
 - ❖ predikce trendů (jak to bude v budoucnu)
 - ❖ současný stav (přesné info)

**Co můžeme
zkoumat?**

Co můžeme zkoumat?

- trendy v oblasti výpůjček
 - využívanost fondu
 - chování uživatelů
 - využívanost služeb
 - využívanost EIZ
-
- sledují se hlavně vztahy a vazby

Konkrétní otázky

- Kolik prváků přišlo do knihovny a půjčilo si knihu k některému z povinných předmětů?
- Vede pravidelná docházka do knihovna k lepším studijním výsledkům?

2. Databáze

- identifikace datových zdrojů
- nalezení vhodných dat pro proces bibliominingu
- problém citlivých údajů
 - ❖ data o čtenářích a jejich chování
 - lze tím lépe cílit své služby na uživatele
 - lze lépe odhalit problémy v knihovních službách a procesech

Dělení zdrojů

■ interní

- ❖ vlastní data v systémech knihovny
- ❖ data o uživateli, výpůjčky, přístupy do katalogu, využívání služeb,...

■ externí

- ❖ mimo vlastní systémy
- ❖ může se k nim dostat nebo je odvodit
- ❖ odkud se uživatelé přihlašují (dle IP adres), data z Google Analytics,...

3. Datový sklad

- nutno zajistit ochranu osobních dat
- vyčištění dat a jejich příprava
 - ❖ info o uživatelích
 - ❖ hledání obecných zákonitostí pro určité skupiny uživatelů
 - ❖ potřebná data se uloží do datového skladu, zbytek se vymaže
- data o uživatelích nelze zpětně z datového skladu obnovit
- nejsložitější proces na bibliominingu

Datový sklad

■ cyklický proces

- ❖ zohledňují se výsledky analýzy
- ❖ přiřazování nových parametrů
- ❖ ladění → učení se z chyb
- ❖ nejprve úzce specifikované otázky
 - odhalí špatný výběr dat do DS
 - chyby v algoritmech

■ důležitá přesnost dat → drobná nepřesnost může ovlivnit výsledky

Otázky

- Lze na vybraných datech provést analýzu?
- Jsou data přesná?
- Jak a kdy se data sbírala?
- ...

4. Výběr analytických prostředků

- vhodná statistická metoda
- software
 - ❖ statistické programy (SPSS, SAS nebo opensource Weka)
 - ❖ algoritmus – vytvoření na míru našim potřebám

5. Analýza dat a jejich ověření

- provedení analýzy

- náročné na čas

- ❖ může se několikrát opakovat, ladit data na vstupu, algoritmy apod.

- ověření modelů a výsledků

- ❖ asistence odborníků

- knihovníci, informační specialisté
- znají procesy, dokáží ověřit
- vypadají výsledky reálně???
- překvapivé výsledky – nedošlo k chybě???, prověření analýzy, dat, algoritmů,...
- snaha vysvětlit anomálie, vysvětlení vybraných trendů
- „typický představitel“ - validace

Nebezpečí analýzy

- snaha o úpravu dat
- manipulace při analýze
- nevhodný výběr dat pro analýzu
- nespolupráce knihovníků a informačních specialistů

6. Prezentace výsledků

- vypracování zprávy
- popsán podrobný postup analýzy
- prezentace veřejnosti
 - ❖ předtím konzultovat s knihovníky
 - ❖ ne na úrovni uživatelů (filtr v DS)
- vhodná vizualizace dat
 - ❖ grafy, tabulky, animace, schémata,...
- nutno vnímat pozitivně
 - ❖ zkvalitnění služeb a procesů

Opakování analýzy

- opakovat v pravidelných intervalech
 - ❖ možnost srovnání
 - ❖ ekonomický rozměr – vyvinutí algoritmů a ladění vstupů náročné na čas a peníze

Výhody a nevýhody

Výhody bibliominingu

- sledování trendů
- analýza současného stavu
- možnost porovnání s jinými institucemi
- analýza interních procesů a služeb
- pomoc při rozhodování
 - ❖ podpora strategického managementu

Nevýhody bibliominingu

- zkreslování dat
- problém při nevhodném výběru dat
- zneužití dat
- časová náročnost procesu
- nutno přesvědčit knihovníky o důležitosti bibliominingu

Bibliomining v knihovných systémech

Staré systémy

- obvykle nemají speciální nástroje
- nutná improvizace
- znalost programování, databází a statistiky
- vlastní algoritmy
 - ❖ kic.ics.muni.cz
 - ❖ Z. Kadlec, M. Sošková
- ARC (Aleph Reporting Center)
 - ❖ omezená funkcionálnita, nejsou trendy

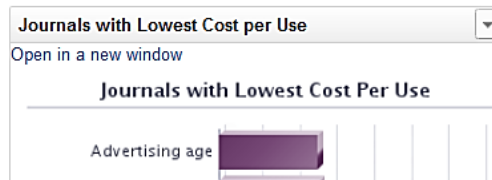
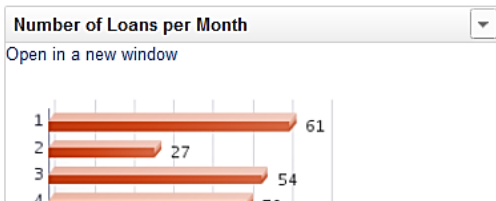
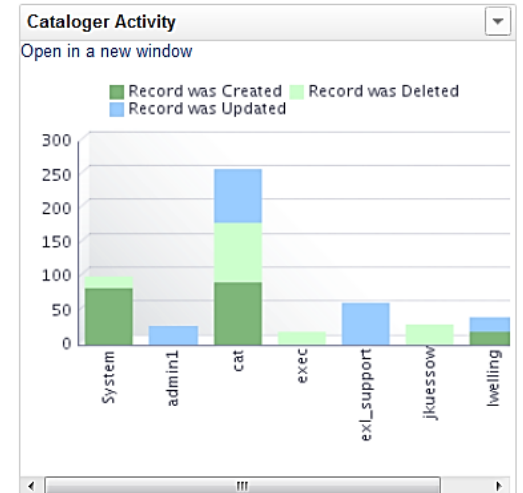
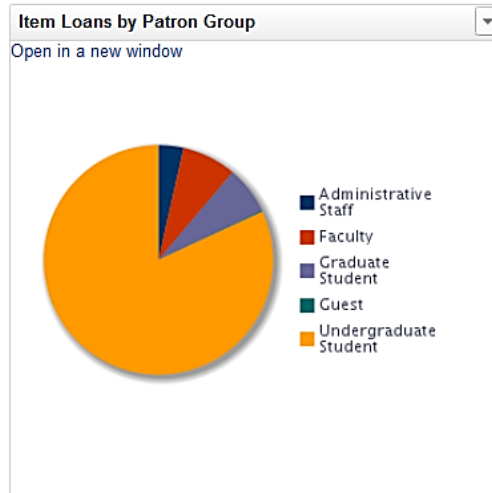
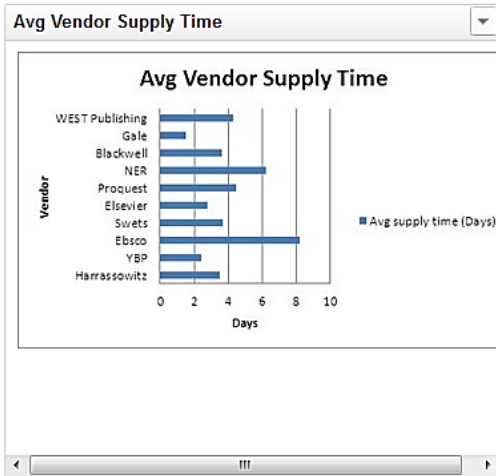
Systemy nové generace

- samostatné moduly
- predikce trendů
- např. ALMA
 - ❖ cloud
 - ❖ modul statistik
 - ❖ šablony
 - ❖ vizualizace dat
 - ❖ jednorázové i opakované reporty
 - ❖ informace pro management a rozhodovací proces

Alma - statistiky

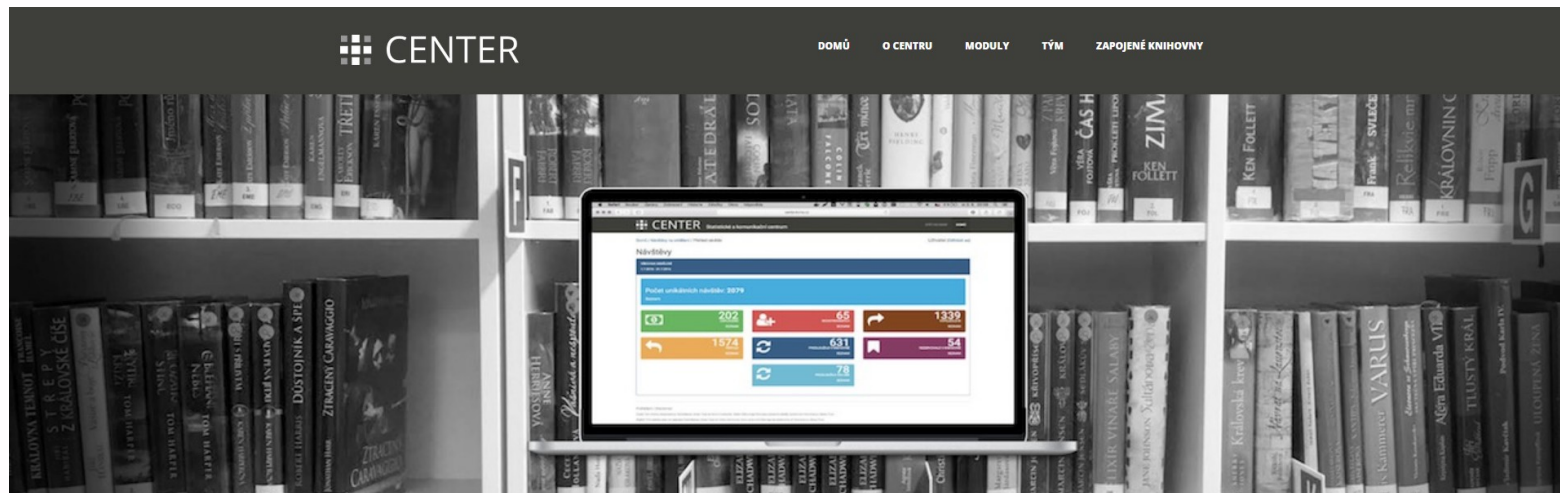
ExLibris **Alma** Tasks Analytics Currently at: Main Library - Main Circula... Help Jim Benson

Welcome Jim Benson



Zdroj: Christian Motovski. [Budoucnost je už dnes: odhalte svoje možnosti s Almou a Primo](#). BA 2014.

Koha CENTER



**STATISTICKÉ A KOMUNIKAČNÍ CENTRUM
PRO OTEVŘENÝ KNIHOVNÍ SYSTÉM KOHA**



Chci Center

Uplatnění v praxi

Jak lze bibliomining
uplatnit v praxi???

Jaké systémy lze
využít???

Systemy

- knihovní systém
- informační systém instituce
- digitální knihovny ([DigiQual](#))
- e-zdroje a databáze
 - ❖ přístupy a využívanost, standard Counter
 - ❖ statistiky AKVŠ
- citační manažery
- ostatní nástroje
 - ❖ Google Analytics na webu knihovny
 - ❖ ...

Uplatnění v praxi

- Data Driven Acquisition = řízená akvizice dle potřeb uživatelů
- podklady pro vedení
 - ❖ rozhodovací proces
 - ❖ provozní doba, posílení oddělení,...
- sledování výpůjček
 - ❖ i prezenční
 - ❖ akvizice, přesuny do/z skladu, vyřazování, posunování regálů (např. co se bude nakupovat???)
- návštěvnost webu a jeho úprava

Uplatnění v praxi

■ katalog

- ❖ návrh rozhraní vyhledávání
- ❖ komentáře a hodnocení, zakomponování do výsledků vyhledávání + relevance
- ❖ kvalita katalogizace – jsou záznamy vyhledatelné jazykem uživatelů
- ❖ funkce systému – co využívají, jak se systém využívá, co je zbytečné, co chybí,...

Zdroje

Zdroje

- <http://www.bibliomining.com>
- <http://www.scottnicolson.com>
- SOŠKOVÁ, Michaela. *Analýza a vyhodnocení činností uživatelů souborného online katalogu Masarykovy univerzity*. Brno: 2006. Dostupné také z Thesis.cz. Bakalářská práce, KISK FF MU, vedoucí práce Z. Kadlec.
- SHIEH, Jiann-Cherng. The integration system for librarians' bibliomining. *The Electronic Library* [online]. 2010, vol. 28, issue 5, s. 709-721 [cit.: 2014-11-13]. DOI: 10.1108/02640471011081988. Dostupné z:
<http://www.emeraldinsight.com/10.1108/02640471011081988>
- NICHOLSON, Scott. Approaching librarianship from the data: using bibliomining for evidence-based librarianship. *Library Hi Tech* [online]. 2006, vol. 24, issue 3, s. 369-375 [cit. 2014-11-13]. DOI: 10.1108/07378830610692136. Dostupné z:
<http://www.emeraldinsight.com/10.1108/07378830610692136>

Závěr



Děkuji Vám za pozornost

Martin Krčál
krcal@phil.muni.cz