Jak využít metody data miningu v knihovně:

příklad Univerzitní knihovny Slezské univerzity

Mgr. Anna Janíková, anna.janikova@fpf.slu.cz

Affiliation: Slezská univerzita v Opavě, Filozoficko-přírodovědecká fakulta, Ústav informatiky; Karlova univerzita v Praze, Filozofická fakulta, Ústav informačních studií a knihovnictví

Abstract

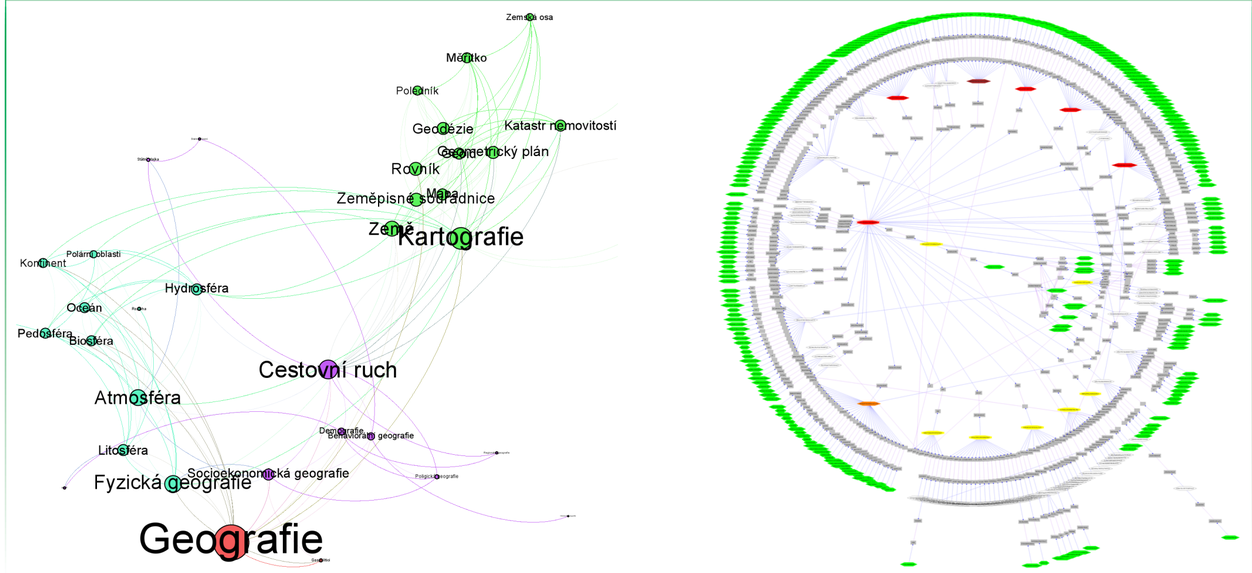
Knihovny jsou ve své podstatě úložišti dat našeho kulturního dědictví. A obsah je uložen fyzicky na regálech s knihami. Ovšem aby knihovna efektivně fungovala, musí vznikat také elektronický odraz jejích činností v automatizovaném systému knihovny a nejen tam. Tyto informace jsou uloženy a dále se vrší bez možnosti s nimi pracovat jiným způsobem, než přidávat nové a nové záznamy.
Tento příspěvek se zabývá možnostmi využití sofistikovaných analýz, které se provádí nad daty, v tomto případě nad daty knihovny, a které by nám mohly přinést potenciálně zajímavé informace. Možnosti využití jsou jak při vytváření modelu uživatele, který je v naší knihovně nejčastější, tak při zvyšování efektivnosti knihovních procesů, evaluaci atd.
Jednou z metod data miningu, je vizualizace dat. Na příkladu Univerzitní knihovny Slezské univerzity si ukážeme, jak by vypadala využitelnost knihovních dat metodami vizualizace.

Keywords: Data mining, Vizualizace dat, Univerzitní knihovna

Úvod do situace

Univerzitní knihovna se jako mnoho jiných knihoven potýká s mnoha problémy. Otázkou zůstává, zda jí v této nelehké situaci mohou pomoci data, která sama vlastní. Mluvíme o datech, která jsou uložena v rámci jejího Automatizovaného knihovnického systému, jsou to data o knihách, výpůjčkách, data o přístupech do katalogu atd.
Problémy, které chceme v rámci data miningu řešit jsou problémy komplexního charakteru, se kterými nemohou pomoci klasické statistiky. A to proto, že statistiky jsou vázány vždy jen na určitý den a fixně na jednu činnost. Kontext v těchto číslech zaznamenán není.
V rámci tohoto posteru bude představena metoda vizualizace dat, která do metod data miningu spadá, a která nám může pomoci už jen ze samotné její podstaty. Vizualizace dat se provádí se záměrem odhalit zajímavé události v datech. Její vlastností je to, že na počátku procesu mnohdy nemusíme vědět, co přesně hledáme. Provedení procesu nám může odhalit něco, na co bychom z pouhých čísel v tabulce mohli těžko usuzovat. Zároveň není na škodu, když tato metoda potvrdí naše domněnky či zkušenosti, které v reálném životě máme, a které pak můžeme právě díky vizualizaci lehce prezentovat.
Například víme-li, že část fondu naší knihovny se zabývá geologií, nepřekvapí nás že část knih bude mít v klíčových slovech popis geologie. Na druhou stranu si v datech můžeme všimnout vazeb, které bychom si těžko sami představovali.
Také pokud chceme zjistit, do jaké části fondu se za poslední dva roky nejvíce investovalo a v jaké části knihovny tyto fondy skončily, můžeme využít vizualizace. Jinak by nás stálo mnoho času a úsilí procházet statistiky či fakturační údaje a data extrahovat.

Obrázek č. 1 Klíčová slova, Obrázek č. 2 Akvizice



Metodika data miningu

Jako mnoho jiných, také data mining může mít mnoho přístupů, které ovlivňují metodiku práce s daty. Jedním z nich je metodika 5A nebo SEMMA. Pro účely našeho problému však použijeme metodiku CRISP-DM (CRoss-Industry Standard Process for Data Mining).[1]
Tato metodika vznikla v rámci evropského projektu, jehož cílem bylo navrhnout univerzální postup pro data mining. Výsledkem je šest fází, které se mohou libovolněkrát opakovat a jít jakkoli za sebou.[2] Jde o:

1. Porozumění problematice – obsahuje pochopení cílů a jejich přesnou formulaci, provádí se plánování, zvažuje se přínos a náklady.
2. Porozumění datům – prvotní sběr dat, získání přehledu o datech a jejich vizualizace.
3. Příprava dat – připravení souboru dat pro analýzu (úprava formátu dat, jejich čištění, selekce, integrace atd.)
4. Modelování – nasazení analytických metod na připravený soubor dat.
5. Evaluace – zhodnocení získaných výsledků z hlediska správnosti a vzhledem k cíli.
6. Využití výsledků – upravení výsledků vzhledem k plánovanému použití.[2]

Kroky vizualizace

V případě vizualizace dat můžeme aplikovat vybranou metodiku a stanovit si kroky jako:

1. Nejdříve musíme formulovat problém, který chceme vizualizovat. Některé problémy mají smysl vizualizovat pouze určitým způsobem. Vizualizace musí dávat jednoznačnou informaci.
2. K našemu cíli si vybereme data, která jsou relevantní.
3. Také pro vizualizaci musíme data připravit. Často pracujeme s externím softwarem, pro ten musíme data upravit. Často během přípravy dat zjistíme, že musíme změnit formu vizualizace. Samozřejmě je možné se k jakémukoli bodu vrátit.
4. Konstrukce vizualizace se provádí ve specializovaném softwaru. Kromě mnoha placených softwarů jsou k dispozici také programy pod svobodnou licencí.[3]
5. Už v průběhu provádění vizualizace je zřejmé, jak výsledek dopadne. Je potřeba jeho podobu konzultovat také s vedením knihovny, pro kterou je primárně zpracováván.
6. Poskytnutí vizualizací knihovně, která může vizualizace přímo využít, nebo upřesnit zadání a v tom případě se cyklus opakuje.

Modelové problémy k vizualizaci

Hotové vizualizace se předávájí knihovně. Mohou být využity k evaluaci nebo prezentaci knihovny. Je důležité mít na paměti, že ať se budeme snažit jakkoli, výsledek vizualizace bude vždy pouze výsek ze skutečnosti.
Na obrázku č. 1 Klíčová slova například můžeme rozdělit knihy, které mají stejná propojení s klíčovými slovy do stejné tématické skupiny. Tato informace může pomoci například pri klasifikaci dokumentů dle MDT, pokud si nejsme jistí, pod který klasifikační znak knihu zařadit. Nebo při reorganizaci fondu, kdy fondy, které jsou popsány stejnými klíčovými slovy, by měly být umístěny blízko u sebe. Tato vizualizace byla sestavena pomocí open software Gephi.[4]
Jako další modelový problém k vizualizaci můžeme chtít vizualizovat data z akvizice za jeden rok, na obrázku č. 2 Akvizice.[5] Zelené listy na okraji obrázku tvoří knihy. Na ně navázaný kruh zaplacené faktury, další kruh směrem ke středu jsou pak prodejci knih a vnitřní výplet grafu tvoří knihovna a její pobočky, sklady a součásti, kam byly knihy uloženy. Z tohoto grafu by bylo možné zjistit užitečné informace, které souvisí s funkcemi knihovny. Získali bychom představu o tom, co jsme v danou dobu koupili, komu všemu jsme za fond zaplatili a kde se momentálně naše nově koupené exempláře nachází. Vizualizace lze samozřejmě zpracovat také zpětně za jakýkoli předchozí rok, pro který se v systému nachází data.
Zde jsou uvedeny pouze dva příklady problémů, k jejichž řešení by mohla knihova využít vizualizaci. Dalších, podobných problémů, bychom však mohli nalézt nespočet. Stejně tak k vizualizaci byl použit jeden software, podobných softwarů s různými přístupy, které generují různorodé grafy a obrázky, je k dispozici nemalé množství, ať už placených, či svobodných.[3]

Reference

Reference

[1] CRISP Consortium. *CRISP-DM Home* [online]. [2000] [cit. 2008-03-02]. Dostupný z:http://www.crisp-dm.org/index.htm

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

[3] KDnuggets. Software: Visualization and Data Mining. In: *Data Mining Communitys Top Resource* [cit. 2012-05-03]. Dostupný z: http://www.kdnuggets.com/software/visualization.html

[4] GEPHI Consortium. *Gephi 0.8.1-beta* [software]. [cit. 2012-04-10]. Dostupný z:http://gephi.org

[5] Gallery Graphviz. In: *Graph Visualization Software* [online]. [cit. 2012-05-03]. Dostupný z:http://www.graphviz.org/Gallery.php.