Uživatelská podpora v prostředí WWW

Jiří Jelínek

Katedra managementu informací

Fakulta managementu Jindřichův Hradec

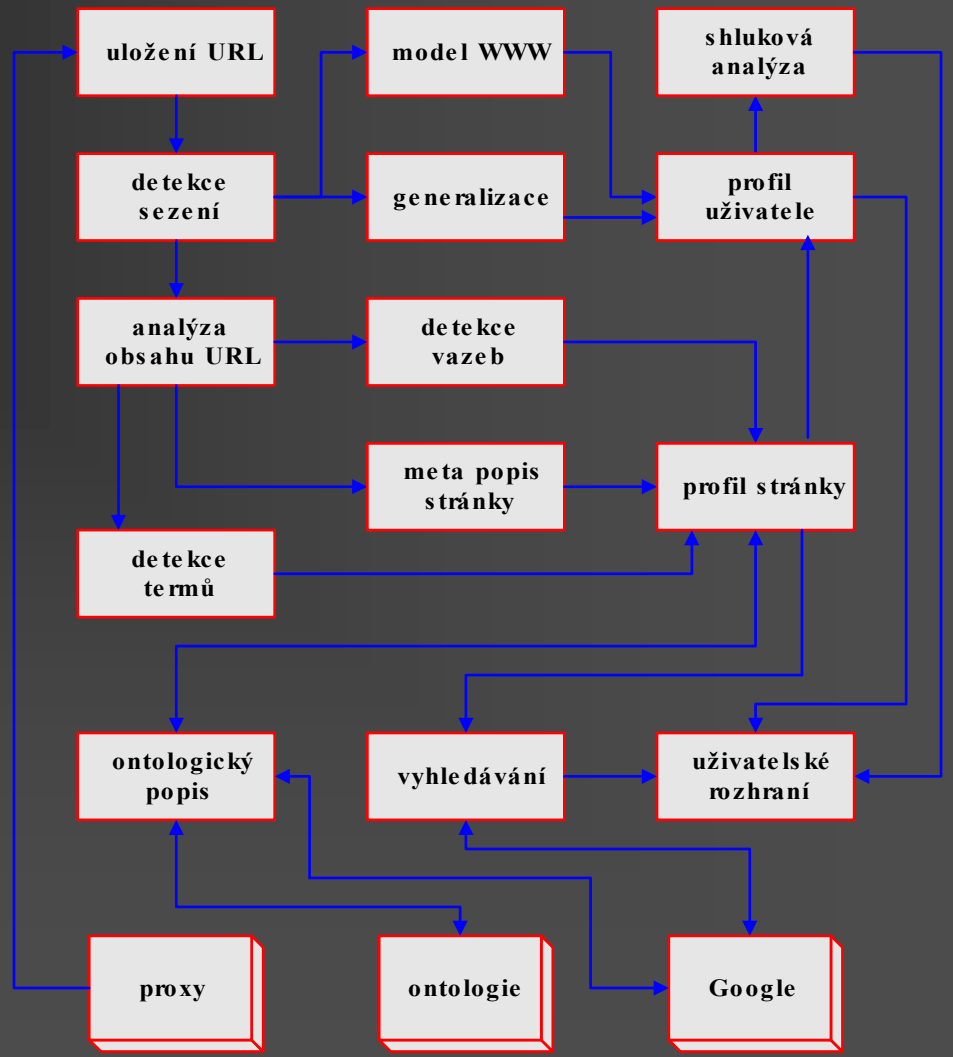
Vysoká škola ekonomická Praha

Úvod

- WWW
  - obsáhlost
  - obsahová i formátová pestrost
    - dokumenty, data, obrázky, video, atd.
  - dynamika
  - malá strukturovanost
  - absence sémantického popisu
  - dostupnost
- hlavní cíl
  - pomoci uživatelům WWW najít užitečné, zajímavé, relevantní informace
  - základem profil uživatele a jeho trvalá aktualizace

Systém podpory uživatele

- cíl
  - nabídka WWW stránek k tématu
- požadavky
  - jednoduché WWW rozhraní
  - rychlá odezva
  - využití co nejširšího spektra metod a jejich kombinací
  - svobodné rozhodnutí uživatele o užití systému



Web Mining

- data mining
  - netriviální extrakce implicitních, předtím neznámých a potenciálně užitečných znalostí z dat ve velkých datových skladech
- shromažďování a zpracování dat dostupných na WWW nebo dat generovaných v průběhu užívání webu
- užití
  - predikce chování a zájmů uživatele založená na předběžně naučených pravidlech a uživatelských profilech

Zdroje dat pro WM

- data o chování uživatele
  - automaticky ukládaná data v logovacích souborech:
    - server – více uživatelů, jedna prezentace
    - klient – obvykle nutná úprava prohlížeče
    - proxy – více uživatelů, více prezentací
  - aplikační záznamy
    - generované záměrně na vyšší programové úrovni
- obsahy WWW stránek
  - zobrazovaný obsah
  - meta popis stránky
  - WWW odkazy
  - URL a jeho struktura

Metody Web Miningu

- *„The Web is an Experimental Laboratory“*

  Ron Kohavi, Mining E-commerce Data, KDD 01

- Web Content Mining
  - zpracování obsahu WWW stránek
- Web Structure Mining
  - získávání informací ze struktury WWW prostoru
- Web Usage Mining
  - analýza chování uživatele (clickstream analýza)

Web Content Mining

- analýza textové složky (obsah a meta popis) stránek
  - detekce sémanticky významných termů a jejich další užití
    - v dané množině stránek se často objevují termy “e-commerce” a “web mining”
  - často založeno na vektorovém modelu dokumentu
  - orientace na klíčová slova, ne sémantický obsah stránek

- práce s vnitřní strukturou stránek
- ontologický přístup
  - užití sémantických sítí a ontologií
  - problémem dostupnost ontologií a nedostatečný sémantický popis stránek
- aplikace: rozšíření vyhledávacích dotazů

vivace (202, 11, 18) - allegro (1202, 25, 48) - (201, 1152, 10)
vivace (202, 11, 18) - midi (2552, 137, 18) - (201, 2214, 10)
vivace (202, 11, 18) - molto (213, 12, 17) - (201, 207, 10)
vivace (202, 11, 18) - adagio (297, 9, 33) - (192, 293, 8)
vivace (202, 11, 18) - allegretto (158, 7, 22) - (182, 158, 7)
vivace (202, 11, 18) - menuetto (136, 5, 27) - (155, 136, 5)

Web Structure Mining

- analýza vzájemného propojení WWW stránek
  - transformace WWW prostoru do orientovaného grafu
  - techniky pro práci s grafy (např. nejkratší cesta)
  - co-citation analysis
    - sémantický obsah stránky je v úzké vazbě k obsahu hypertextově svázaných stránek
  - aplikace: detekce významných stránek (pagerank)

- analýza struktury WWW prostoru pomocí dekompozice URL
  - generalizované hierarchické stromy
    - užití při porovnávání uživatelů a jejich shlukování

```
celek ( 11 - 100.00 - 100.00)
        .com ( 2 - 18.18 - 18.18)
                .mysql ( 1 - 9.09 - 50.00)
                        .www ( 1 - 9.09 - 100.00)
                                / ( 1 - 9.09 - 100.00)
                .xnef ( 1 - 9.09 - 50.00)
                        /xnef ( 1 - 9.09 - 100.00)
                                /report ( 1 - 9.09 - 100.00)
        .cz ( 8 - 72.73 - 72.73)
                .idnes ( 4 - 36.36 - 50.00)
                        .zpravy ( 4 - 36.36 - 100.00)
                                / ( 1 - 9.09 - 25.00)
                                /foto.asp ( 3 - 27.27 - 75.00)
                .msmt ( 1 - 9.09 - 12.50)
                        .www ( 1 - 9.09 - 100.00)
                                / ( 1 - 9.09 - 100.00)
                .seznam ( 1 - 9.09 - 12.50)
                        .www ( 1 - 9.09 - 100.00)
                                / ( 1 - 9.09 - 100.00)
                .vse ( 2 - 18.18 - 25.00)
                        .fm ( 1 - 9.09 - 50.00)
                                .www ( 1 - 9.09 - 100.00)
                        .www ( 1 - 9.09 - 50.00)
                                / ( 1 - 9.09 - 100.00)
        .org ( 1 - 9.09 - 9.09)
                .w3c ( 1 - 9.09 - 100.00)
                        .www ( 1 - 9.09 - 100.00)
                                / ( 1 - 9.09 - 100.00)
```

Web Usage Mining

- detekce vzorů v datech generovaných v průběhu spojení mezi klientem a serverem WWW
- asociační pravidla a statistické metody
  - odkrývání závislostí mezi užitím WWW stránek
    - 60% uživatelů, kteří navštívili URL-A, také navštívilo URL-B
    - 75% uživatelů stahujících soubory z URL-A tak činilo mezi 19:00 a 23:00 během víkendů
- techniky shlukování
  - seskupování uživatelů s podobnými vzory chování
  - seskupování stránek navštěvovaných stejnou skupinou

- detekce sekvenčních vzorů (FGS)
  - objevování častých sekvencí URL charakteristických pro uživatele či sezení
  - možnost predikce chování uživatele

```
0 - http://www.php.net/
1 - http://www.php.net/search.php
-----------------------------------------------
0 - http://zpravy.idnes.cz/
1 - http://www.php.net/
-----------------------------------------------
0 - http://www.google.com/advanced_search?hl=cs
1 - *
2 - http://146.102.250.208/homes
-----------------------------------------------
0 - http://cz.php.net/search.php
1 - *
2 - http://cz.php.net/search.php
-----------------------------------------------
0 - http://www.sweb.cz/root/index
1 - http://www.google.com/advanced_search?hl=cs
2 - *
3 - http://www.php.net/
4 - *
5 - http://opravy..idnes.cz/
```

Problémy užití WM

- vstupní data
- logovací soubory
  - původně vytvářeny pro účely ladění
  - obsahují velké množství neužitečných informací
  - absence některých dat (použití cache)
  - další zdroje, pokud jsou k dispozici
  - detekce sezení, uživatele nebo významné události
  - uložení URL stránek, nikoliv sémantický popis
  - problémy s dynamickými stránkami
  - absence dat z webových formulářů
- ochrana soukromí uživatelů
- náročnost výpočtů

Diskuse