Web mining

PV056 Strojové učení a dobývání znalostí

Fakulta Informatiky, Masarykova Univerzita
Juraj Jurčo, 173001@mail.muni.cz

Jar 2012

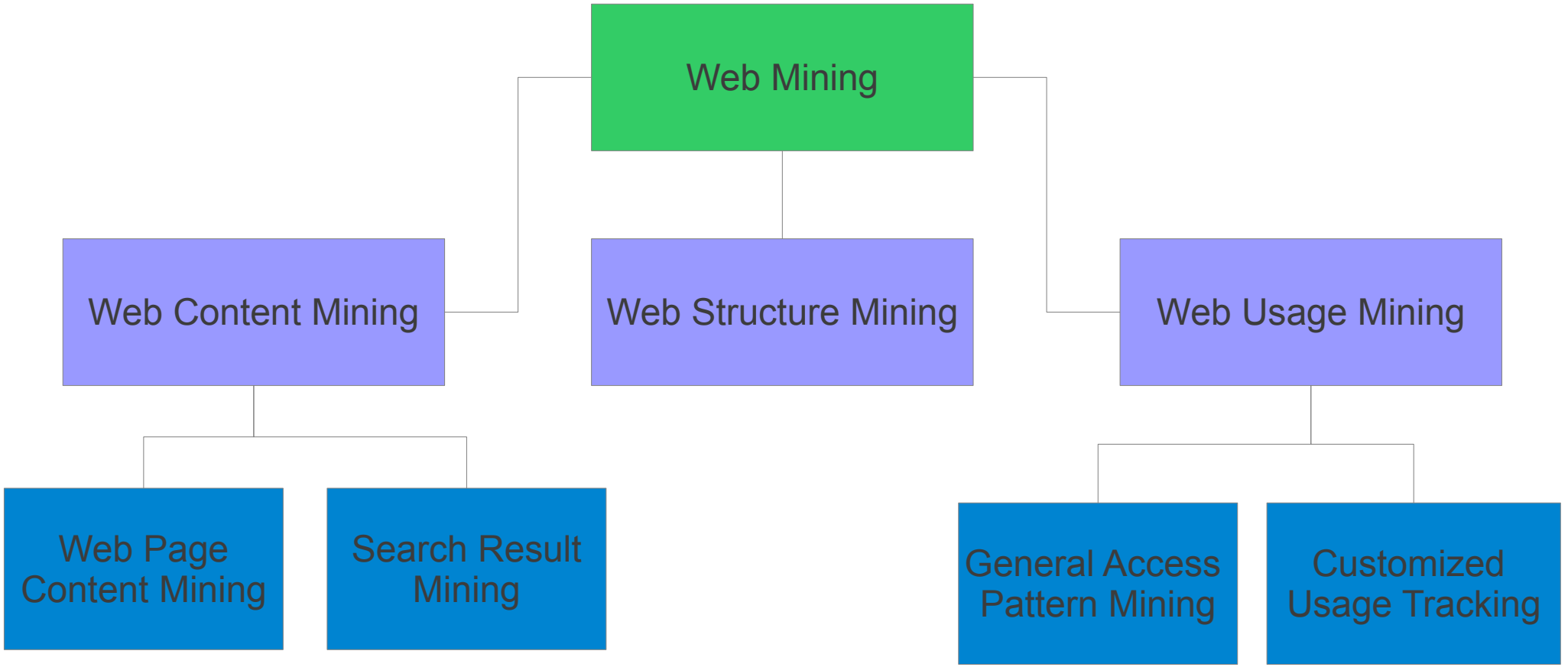
Motivácia

- Internet je bohatý zdroj informácií
- V roku 2005 Eric Schmidt (Google, SEO) predpokladal, že veľkosť internetu je približne 5EB [1]
  - Veľké množstvo dát je skrytých (DBMS)
- Na internet je dnes pripojené takmer všetko (pc, mobily, tablety, herné konzoly, TV, čítačky kníh, roboti, chladničky, hodinky, kamery...)

Čo je web mining?

- Aplikácia techník dolovania z dát na obsah, štruktúru (odkazov) a získavanie užitočných informácií a vzorov z dokumentov, zdrojov a služieb na Internete.

Taxonómia Web miningu[2]



Interakcie na internete

- Spoločenský pohľad
  - Užívateľ – server
    - Prezeranie stránok a dokumentov, Vyhľadávanie, Digitálne knižnice...
  - Užívateľ – užívateľ
    - Sociálne siete, chaty, diskusné fóra
- Technický uhol pohľadu
  - Prístup k obsahu (čítanie)
  - Vytváranie obsahu (zápis)
  - Navigácia užívateľa na webe

Zdroje dát

- Web stránky a online dokumenty
- Štruktúra stránok a dokumentov
- Štatistiky využívania zdrojov
- Ďalšie dáta
  - Užívateľské profily
  - Registračné údaje
  - Cookies

Problémy spojené s Web Miningom

- Veľké množstvo dát na internete
- Veľa dát je irelevantných
  - Môže byť problém nájsť relevantné dáta
- Personalizácia stránok
  - Podľa zariadenia – rôzne dáta
  - Jazyková lokalizácia – viac jazykov v dokumente
  - Irelevantný obsah

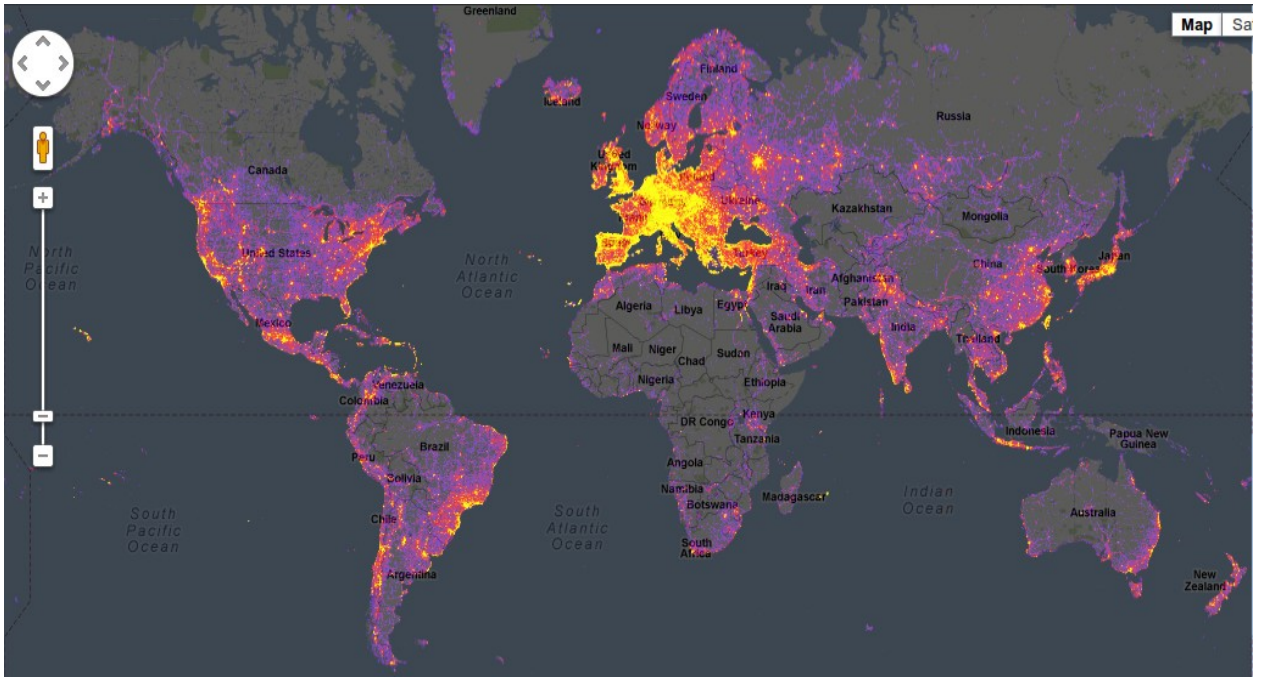
Web usage mining

- Taktiež nazývaný “log mining”
- Dáta sú zbierané na základe používania internetových zdrojov
  - IP adresy, dátum a čas prístupu
  - Užívateľské profily, session dáta, cookies
  - Dotazy užívateľov, záložky
  - Interakcia myši (kliknutia, skrolovania, poloha)
  - ...

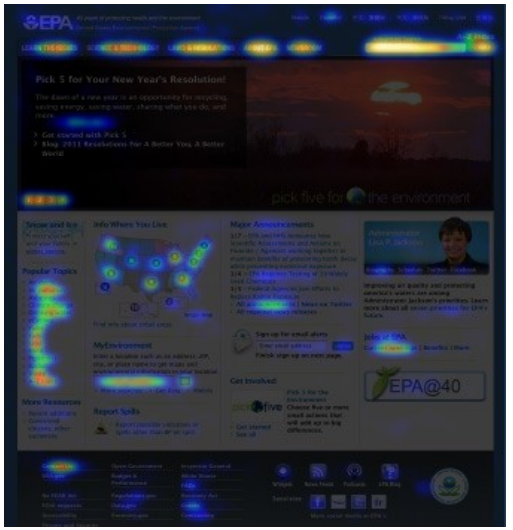
Využitie Log Miningu

- Analýza prístupu a vzorov užívateľského správania
  - Prihlásenie sa do systému
  - Analýza SSH útokov
- Analýza odkazujúcich stránok
- Personalizácia stránok, vyhľadávania
  - Nákup podobných produktov, zjemňovanie dotazu
- Analýza dôležitých častí stránky
  - Rozloženie stránky
- Využívania zdrojov
  - Rozloženie dát medzi servermi; neočakávané prístupy

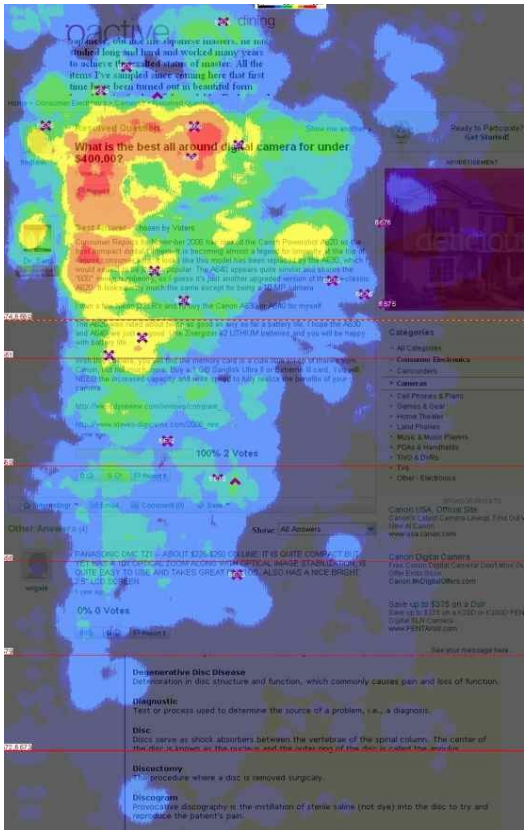
Log mining – vizuálna analýza



Najčastejšie fotené miesta sveta (http://www.sightsmap.com/)



Najčastejšie klikanie na stránke [3]



Sledovanie kde sa užívateľ najčastejšie pozerá [4]

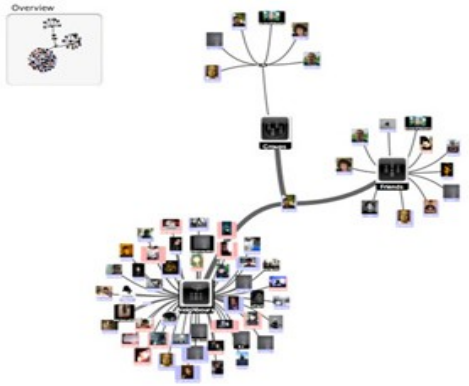
Web Structure Mining

- Dolovanie zo štruktúry prepojenia objektov na Internete
  - Tok informácií na Internete
  - Odkazovanie medzi stránkami
- Štruktúra dokumentov
  - Rozloženie a vzory HTML/XML tagov

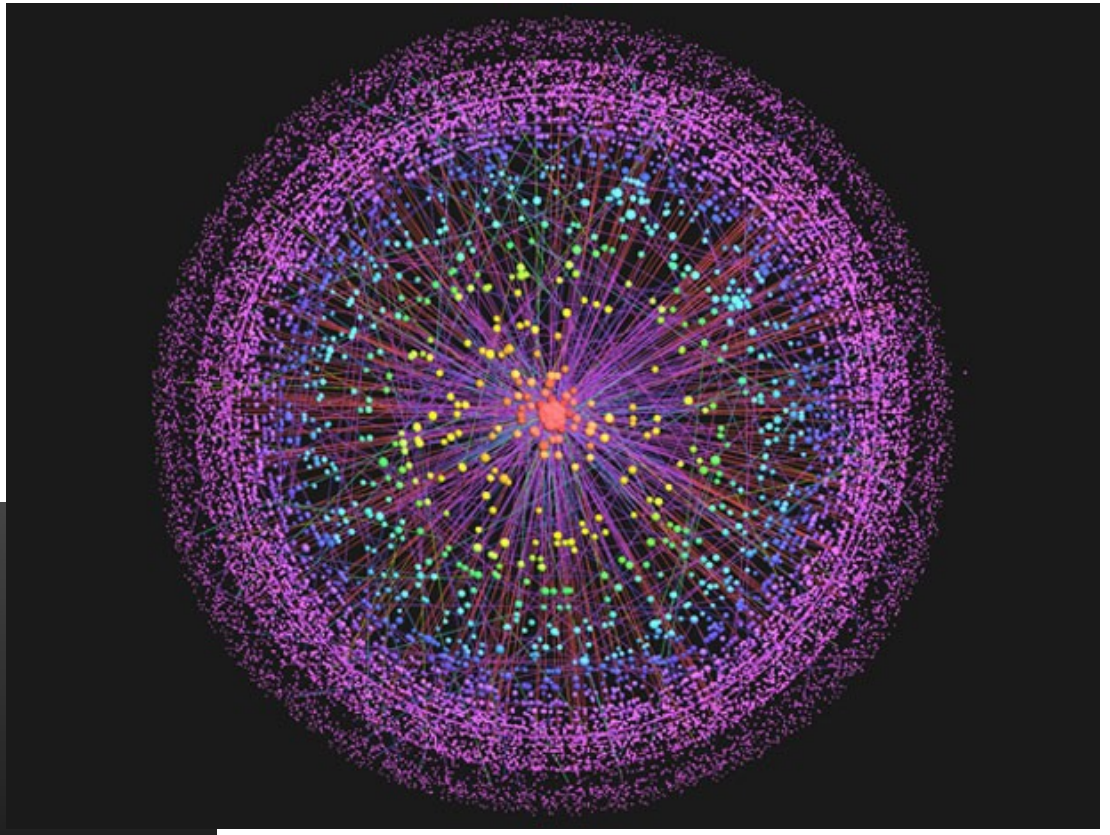
Využitie Structure Miningu

- Prepojenie dokumentov na internete
  - Prepojenie (sub)sietí: užívatelia, vedci, teroristi
  - Citačné siete
  - Autoritatívne servery / stránky
  - Šírenie informácií
- Identifikácia častí dokumentov
  - Nadpis, hlavička, pätička, hlavný obsah
  - Menu, reklama

Structure mining – vizuálna analýza



Vizualizácia sociálnej siete [5]



Hierarchické zobrazenie smerovania v P2P sieťach [6]

3D zobrazenie štruktúry stránky [7]

Content mining[8]

- Dolovanie informácií z dokumentov
  - Textových, multimediálnych
  - Zhlukovanie / filtrovanie dokumentov
- Dolovanie z výsledkov hľadania
  - Information Retrieval Systems

Zber dát

- API, vyhľadávače, web spiders...
- Problémy pri získavaní dát
  - Formáty dokumentov
    - Web stránky, PS/PDF, (Open/Libre) Office, RTF, Flash
  - Jazyky dokumentov
  - Kódovanie dokumentov

Predspravovanie dokumentov

- Predspracovanie “za behu”
- Extrahovanie hlavných častí dokumentu
- Identifikácia Pomenovaných Entít (Named Entities)
- Stop list
- Bag of words
- Stemming

Techniky Web Miningu[8]

- Clustering
  - Zhlukovanie položiek s rovnakými vlastnosťami
- Klasifikácia
  - Príslušnosť prvku ku skupine na základe spoločných atribútov
- Asociačné pravidlá
  - Predpoveď závislosti atribútov (A a B => C)
- Path analysis
  - Ktoré kroky predchádzali k tomu, než sa užívateľ dostal na určitú stránku?
- Sekvenčné vzory
  - Vzory indikujúce určité správanie užívateľa
    (30% ľudí predým ako navštívi stránku firmy navštívil Yahoo! a vo vyhľadávanom reťazci sa vyskytovalo písmeno “w”)

Ďakujem za pozornosť!

Referencie

- [1] Brendan McGuigan. „*How Big is the Internet?“* [online]. 2005, [citované 27.11.2010]. < http://www.wisegeek.com/how-big-is-the-internet.htm >.
- [2] Bamshad Mobasher, “A Taxonomy of Web Mining,” 16-Jul-1997. [Online]. Available: http://maya.cs.depaul.edu/~mobasher/webminer/survey/node2.html. [Accessed: 11-May-2012].
- [3] Jeffrey Levy, “Home Page Heat Maps,” Home Page Heat Maps, 11-May-2011. [Online]. Available: http://levyj413.wordpress.com/2011/01/14/home-page-heat-maps/. [Accessed: 12-May-2012].
- [4] Yahoo!, “Eye-tracking: Where do readers look first? | Yahoo! Style Guide,” Eye-tracking: Where do readers look first?, 26-Jun-2010. [Online]. Available: http://styleguide.yahoo.com/writing/write-web/eye-tracking-where-do-readers-look-first. [Accessed: 12-May-2012].
- [5] last.fm, “Build Last.fm: Extend your Last.fm experience – Last.fm,” Build Last.fm: Extend your Last.fm experience – Last.fm, 12-May-2012. [Online]. Available: http://build.last.fm/item/42. [Accessed: 12-May-2012].
- [6] Duncan Graham-Rowe, “Mapping the Internet - Technology Review,” Mapping the Internet, 19-Jul-2007. [Online]. Available: http://www.technologyreview.com/read_article.aspx?id=18944. [Accessed: 12-May-2012].
- [7] Mozilla, “Firefox Beta with New Developer Tools and Add-on Sync is Ready for Testing | Future Releases,” Firefox Beta with New Developer Tools and Add-on Sync is Ready for Testing, 03-Feb-2012. [Online]. Available: http://blog.mozilla.org/futurereleases/2012/02/03/firefoxbeta11/. [Accessed: 12-May-2012].
- [8] Ahmed Rafea, “Web MINING Overview”, 05-Sep-2012. [Online] Available: http://www.cse.aucegypt.edu/~rafea/CSCE564/slides/WebMiningOverview.pdf. [Accessed: 12-May-2012].