

Web-rozhranie ako prostriedok uľahčenia dolovania dát pre bežných užívateľov

Web-UI = Data-Mining for Normal Users Made Easy

Rastislav Mucha — Zlatica Ivanová-Muchová

Abstract

This contribution illustrates some possibilities of making the access of normal users to searching for informations and data-mining easier by providing shortcuts (caned search queries), by integrating the front ends of search engines into the existing local web, and by interconnecting the powerful environment for data analysis R with Zope/Plone site using an interface to the Python programming language.

Keywords

knowledge discovery, data-mining, web searching, R, Python, Zope, CMS Plone

Úvod

Zámerom tohto príspevku je poukázať na niektoré možnosti, ako uľahčiť bežným používateľom samostatné získavanie údajov a („dolovanie“) relevantných informácií z nich a to nielen v oblastiach týkajúcich sa informačných a komunikačných technológií (IKT).

Skúsenosti zo sledovania prístupov na fakultný internet/intranet Fakulty záhradníctva a krajinného inžinierstva (FZKI), ktorý je postavený na systéme správy obsahu (CMS) Plone (Plone Foundation, 2005; Zope Community, 2005), poukazujú na skutočnosť, že len malá časť návštevníkov (tak z internetu, ako aj intranetu) dokáže efektívne vyhľadávať informácie. Podobná situácia vládne aj pri vyhľadávaní na internete. Jednou z možností ako užívateľom pomôcť, je vytvoriť „skratky“ k niektorým informačným zdrojom.

Získanie (identifikácia) zaujímavých vzťahov a súvislostí z dostupných údajov aplikáciou vhodných metód môže byť tiež uľahčená, aspoň v úvodnej fáze, sprístupnením niektorých postupov z oblasti označovanej ako knowledge discovery (KD), resp. data-mining prostredníctvom jednoduchého rozhrania.

Vyhľadávanie

Vhodnou skratkou k vybraným údajom uloženým lokálne je poskytnúť užívateľom vopred pripravené vyhľadávacie otázky k dostupnému vyhľadávaciemu mechanizmu, v prípade CMS Plone môžeme využiť štandardný typ Topic, Téma (obr. 1).

Integrácia vyhľadávacích rozhraní niektorých služieb, archívov a dátových repozitórií dostupných na internete do lokálneho webu sprístupňuje návštevníkom ich služby priamo a to aj v prípade, že o nich predtým nevedeli, resp. ich nevyužívali (obr. 2).

Data-Mining

Zisťovanie súvislostí a zaujímavých vzťahov v údajoch si vyžaduje určité povedomie o dostupných a vhodných metódach, prístup k nim a vhodnú reprezentáciu získaných výsledkov. Obr. 3—5 ilustruje náš pokus o sprístupnenie možností poskytovaných prostredím R (R Core Development Team, 2004—2005) prepojením so Zope a CMS Plone prostredníctvom rozhrania v programovacom jazyku Python (Python Software Foundation, 2004—2005). Zatiaľ dostupné rozhrania dovoľujú využitie niektorých metód exploračnej analýzy dát, klasifikačných a regresných stromov, Bayesovskej analýzy, k-NN, pričom ponúkané možnosti, podľa našich poznatkov, presahujú štandardný záber metód využívaných väčšinou užívateľov v našom širšom okolí.

Plone

Zdroje k Plone Resources

názov (titul)	typ	posledná zmena	popis
"Oficiálny" preklad CMF/CMS Plone 2 do Slovenčiny	News Item	2005-02-03 02:40:17	"Oficiálny" preklad CMF/CMS Plone 2 do Slovenčiny nájdete v sekcii "Na stiahnutie" (Downloads) v časti venovanej IKT (ICT) / Slovak translation for CMF/CMS Plone 2 (ICT/Downloads)
Prehľad o využívaní systémov na správu obsahu (CMS) na FZKI - súčasný stav, pripravované projekty	File	2005-02-05 03:02:11	Prehľad o využívaní systémov na správu obsahu (CMS) na FZKI - súčasný stav, pripravované projekty (plný text príspevku) / Content Management Systems (CMS) Utilization at HLEF Review - Current State, New Projects (full contribution), Plone
CMS Plone na FZKI	Folder	2005-02-05 03:04:22	Využívanie systémov správy obsahu (CMS) na FZKI, Plone
plone-fzki-sk.pot.gz	Link	2005-02-03 02:40:35	Tento súbor takmer určite nebudete potrebovať. Lokalizačný po(t)-súbor CMF/CMS Plone 2 pre Slovenčinu (gzp-ovaný), ale bez vykonania msgmerge a msgattrib (ak neviete o čom je reč stiahnite si radšej plone-fzki-sk.po.gz) doplnený o preložené reťazce (názvy) používané na FZKI SPU, ktoré zodpovedajú typickej štruktúre fakulty univerzity na Slovensku (samozrejme názvy konkrétnych pracovísk Vám asi budú na nič).
plone-sk.pot.gz	Link	2005-02-03 02:39:18	Lokalizačný po(t)-súbor CMF/CMS Plone 2 pre Slovenčinu (gzp-ovaný), ale bez vykonania msgmerge a msgattrib (ak neviete o čom je reč stiahnite si radšej plone-sk.po.gz)
plone-sk.po.gz	Link	2005-02-03 02:39:32	Lokalizačný po-súbor CMF/CMS Plone 2 pre Slovenčinu (gzp-ovaný)
Prehľad o využívaní systémov na správu obsahu (CMS) na FZKI - súčasný stav, pripravované projekty (úvod)	Document	2005-02-05 03:03:41	Prehľad o využívaní systémov na správu obsahu (CMS) na FZKI - súčasný stav, pripravované projekty (úvod) / Content Management Systems (CMS) Utilization at HLEF Review - Current State, New Projects (introduction), Plone
plone.pot.gz	Link	2005-02-03 02:39:02	Lokalizačný po(t)-súbor (gettext po template) pre CMF/CMS Plone 2 (gzp-ovaný), ktorého preklady sú tu dostupné (pre prípad, že by ste chceli skontrolovať/opraviť preklad). Zdroj: http://plone.org
Projekt - Portálové riešenie a CMS na SPU v Nitre	File	2005-02-05 03:06:44	Doplňujúce poznámky k CMS na SPU (CMS, CMS SPU, Správa obsahu na SPU, Systém správy obsahu, CMS Plone) 8.-9. augusta 2004 spracovali: Miloslav Mucha a Rastislav Mucha (12 strán, 13 obrázkov, 1642426 bytes, pdf)

Obr. 1: Na tomto obrázku sú predstavené výsledky získané vyhľadávaním termínu „Plone“ pomocou preddefinovanej, vopred pripravenej otázky do interného vyhľadávača webu postaveného na CMS Plone. Otázka bola administrátorom pripravená tak, aby výsledkom boli všetky relevantné interné zdroje.

Vyhľadanie v Google, RFC a FAQ

Vyhľadanie v Google, RFC a FAQ

Google

WWW uniaq.sk fzki.uniaq.sk

RFC.net

RFC index | STD index | BCP index | FYI index

RFC:

Internet FAQ Archives - www.faqs.org

Hľadať FAQs - Plný text

Obr. 2: Rozhranie prezentované na tomto obrázku umožňuje priame (jednoduché) vyhľadanie v Google (google.com), dokumentoch o protokoloch a štandardoch Internetu na RFC.net a v archívoch „často kladených otázok“ FAQ (www.faqs.org).

R + Zope + RPy

```

library(cluster)
library(tree)
library(deal)

Loading required package: dynamicGraph
Loading required package: toltk

#This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables
#sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of
#iris. The species are Iris setosa, versicolor, and virginica.
#Tento známy (Fisherov resp. Andersonov) súbor dát uvádza merania (v cm) dĺžky a šírky lupienkov a kališných
#listkov pre 50 kvetov z každého z 3 druhov kosatca (Iris setosa, versicolor, virginica).

#Sepal.Length = Dĺžka kališného listka
#Sepal.Width = Šírka kališného listka
#Petal.Length = Dĺžka lupienka
#Petal.Width = Šírka lupienka
#Species = Druh

data(iris)
names(iris)

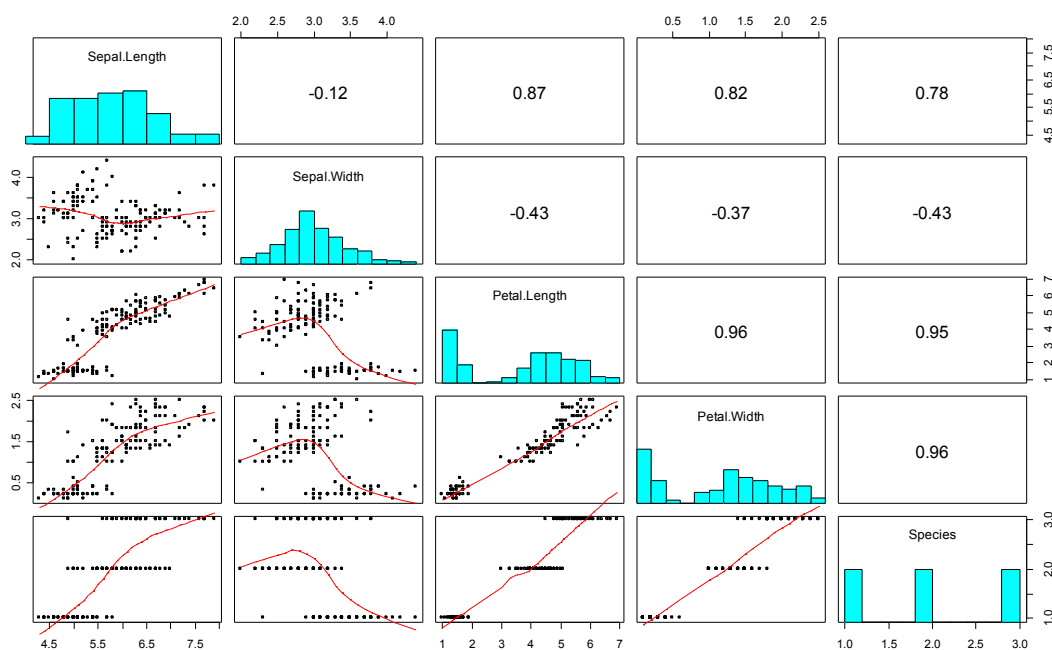
[1] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width" "Species"

iris.SLSWPLW <- iris[1:4]
summary(iris)

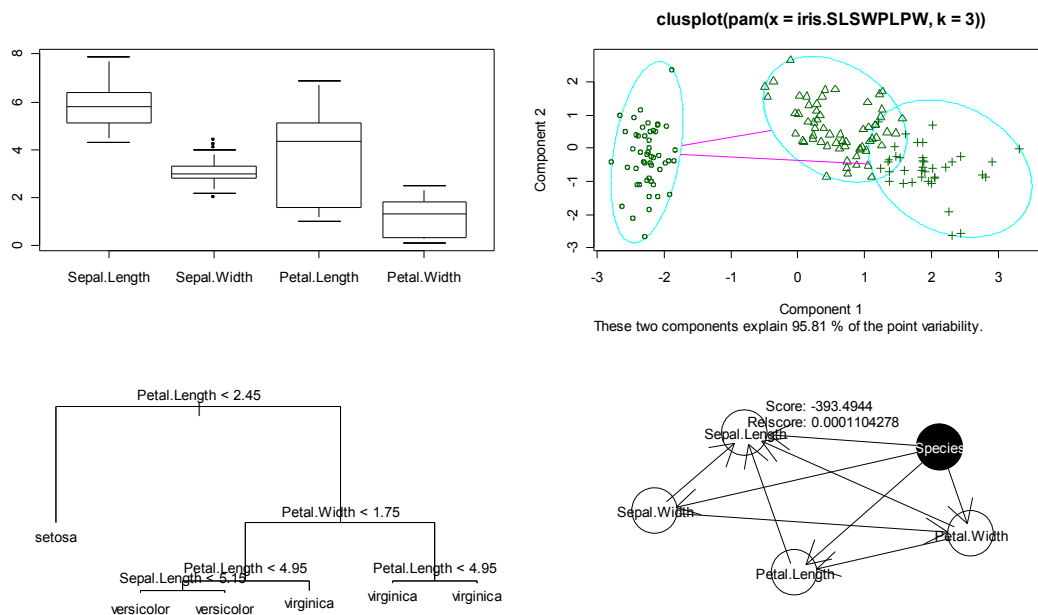
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
Min. :4.300	Min. :2.000	Min. :1.000	Min. :0.100	setosa :50
1st Qu.:5.100	1st Qu.:2.800	1st Qu.:1.600	1st Qu.:0.300	versicolor:50
Median :5.800	Median:3.000	Median:4.350	Median:1.300	virginica :50

Obr. 3: Príklad data-miningu v R cez Plone/Zope rozhranie. Výstup je koncipovaný tak, aby tu uvedenú textovú časť záznamu sedenia v R (R Session) mohol užívateľ použiť pri interaktívnej práci v systéme R. Ako vstupné dáta slúži jednoduchý textový súbor buď s pevnými oddeľovačmi alebo vo formáte CSV (ako oddeľovač slúži čiarka, alebo u nás častejšie bodkočiarka). Tento príklad využíva známy súbor tvorcovi mnohých štatistických metód R. A. Fishera s niektorými parametrami (dĺžka a šírka kališného listka, dĺžka a šírka lupienka) 3 druhov Kosatca.



Obr. 4: Maticový rozptylový diagram (pokračovanie výstupu z obr. 3) zobrazuje priebeh jednotlivých parametrov dátového súboru a súvislosti medzi nimi (v grafickej podobe — dole, histogramy na diagonále a korelácie — hore).



Obr. 5: Pokračovanie grafického výstupu (z obr. 4). Boxploty, krabicové diagramy, ilustrujú rozptyl hodnôt meraných parametrov — vľavo hore. Diagram 3 zhhlukov získaných metódou delenia okolo centroidov ako priemet do roviny tvorenej prvými dvoma kľúčovými komponentami — vpravo hore. Klasifikačný strom — vľavo dole. Bayesovská sieť — vpravo dole.

Záver

Vyššie uvedené výstupy sú priamo využiteľné návštevníkmi fakultného webu FZKI (<http://fzki.uniag.sk>) na získanie nového pohľadu na nimi spracovávané dáta, rozšírenie repertoáru používaných metód a zdokonaľovanie zručností pri získavaní a analýze informácií. V rámci riešenia predstaveného projektu „Podpora užívateľov IKT využitím technológií data-miningu“ boli tiež preložené všetky správy stabilnej verzie CMS Plone a poskytnuté Internetovej komunite v rámci projektu Plone i18n.

Súhrn

Tento príspevok ilustruje niektoré možnosti zjednodušenia prístupu užívateľov k vyhľadávaniu informácií a technikám data-miningu prostredníctvom skratiek (vopred pripravené vyhľadávania), integráciou vstupných rozhraní vyhľadávačov do lokálneho webu a prepojením výkonného prostredia na analýzy dát R so Zope/Plone inštaláciou cez rozhranie v programovacom jazyku Python.

Kľúčové slová

objavovanie znalostí, data-mining, web-vyhľadávanie, R, Python, Zope, CMS Plone

Literatúra

- [URL1] Plone Foundation, 2005. Dostupné na Internete: <<http://plone.org>>.
- [URL2] Python Foundation, 2005. Dostupné na Internete: <<http://python.org>>.
- [URL3] R Core Team, 2005. Dostupné na Internete: <<http://www.r-project.org>>.
- [URL4] Zope Community, 2005. Dostupné na Internete: <<http://zope.org>>.

Kontakt

Ing. Zlatica Ivanová-Muchová, PhD., KPPU FZKI SPU, Hospodárska 7, 949 76 Nitra, Zlatica.Ivanova@uniag.sk