Ekstrakcija podatkov

Skupina **DOMACI-NJOKI**: Niki Bizjak, Bojan Vrangeloski, Uroš Škrjanc

***Povzetek*—Cilj prve seminarske naloge je bil napisati pajka, ki zna s spleta prenašati spletne vsebine in jih shranjevati v podatkovno bazo. Pred iskanjem podatkov po taki podatkovni bazi, pa je treba prenesene vsebine najprej prečistiti in iz njih izluščiti pomembne informacije. V drugem seminarju si bomo ogledali tri različne načine za ekstrakcijo podatkov iz HTML vsebin.**

I. Uvod

Druga seminarska naloga pri predmetu Iskanje in ekstrakcija podatkov s spleta je namenjena pridobivanju informacij iz vsebin, ki jih je s spleta prenesel pajek. V seminarju smo si ogledali tri različne načine ekstrakcije - dva taka, ki zahtevata veliko uporabnikovega poseganja in razumevanja strukture strani in tretjega, ki zna informacije prepoznati in izluščiti avtomatično, s primerjavo podobnih spletnih strani.

II. Regularni izrazi

Regularni izrazi nam omogočajo učinkovito iskanje informacij v nizih z uporabo končnih avtomatov.

III. XPath

XPath je poizvedovalni jezik, ki je bil razvit za namene ekstrakcije podatkov iz XML podatkovnih struktur.

IV. Avtomatična ekstrakcija podatkov

Na spletu se od uvedbe programskega jezika PHP naprej pojavlja dosti dinamičnih spletnih strani. Take strani hranijo podatke v podatkovni bazi in glede na dostopan URL naslov izvedejo poizvedbo v podatkovni bazi in na podlagi rezultata prikažejo podatke v uporabniku prijazni obliki. Pri avtomatični ekstrakciji podatkov, bi radi na podlagi dveh strani, ki imata enako strukturo, a drugačno vsebino, zgradili program, ki zna samodejno pridobiti podatke.

V našem primeru smo za avtomatično ekstrakcijo podatkov uporabili algoritem RoadRunner [1], [2]. Ta sicer v originalni implementaciji deluje na seznamu žetonov (angl. token), v našem primeru pa smo za predstavitev HTML vsebin uporabili kar DOM drevo.

Literatura

[1] V. Crescenzi, G. Mecca, and P. Merialdo, "The roadrunner project: Towards automatic extraction of web data," 01 2001.

[2] ——, "Automatic web information extraction in the road r unner system," in *International Conference on Conceptual Modeling*. Springer, 2001, pp. 264–277.