VJEŽBA 2: Web roboti / Web pauci

Web roboti ili **Web pauci** (engl. *Web robots, bots, Web spiders, Web crawlers, ants, automatic indexers*) su računalni programi koji na neki sustavan i unaprijed definiran način pretražuju Internet.

Gdje se koriste web roboti?

- U Web tražilicama (engl. *search engines*) za stvaranje kopija posjećenih Web stranica, njihovo indeksiranje[1] i ubrzavanje budućih pretraga Interneta.
- Prilikom izrade programa koji parsiraju Web stranice, na njima pronalaze e-mail adrese, i onda na njih šalju različite reklame i neželjene e-mail poruke (engl. *spam*).

Kada se Web roboti koriste u svrhu automatskog pretraživanja Interneta, općenito se realiziraju na slijedeći način:

1. korisnik određuje informaciju koju želi potražiti na Internetu (npr. želi pronaći sve e-mail adrese ili brojeve telefona) i formalno je prikazuje (npr. u obliku regularnog izraza);
2. korisnik Web robotu zadaje početnu Web stranicu (engl. *seed*) odakle će on započeti svoju pretragu;
3. Web robot pronalazi sve link-ove unutar te stranice i zapisuje ih u listu link-ova koje mora posjetiti (engl. *crawl frontier*);
4. Web robot posjećuje i pretražuje sve link-ove sa liste, te cijelo vrijeme u listu dodaje nove link-ove koje je pronašao na novim Web stranicama (ovaj se korak ponavlja sve dok lista ne ostane prazna ili dok se ne dosegne maksimalni broj iteracija izvršavanja algoritma kojega je unaprijed definirao korisnik);
5. Web robot rezultate pretrage prikazuje korisniku.

Pretraživanje Interneta pomoću Web robota se u računarstvu naziva (u nedostatku općenito prihvaćenog hrvatskog prijevoda) **Web crawling** ili **Web spidering**.

Zadatak ove laboratorijske vježbe jest isprogramirati vlastitog Web robota koji bi sa link-ova pronađenih na početnoj Web stranici pronašao određene informacije. Detaljnije upute vezane za funkciju Web robota će biti dane na laboratorijskim vježbama. Biblioteke C/C++ funkcija koje bi vam mogle biti korisne prilikom izrade Web robota navedene su i objašnjene u nastavku teksta, a preporučeni programski jezik za programiranje Web robota je C++.

1 Indeksiranje web stranice je proces kojim se ta stranica predstavlja na nekakav formaliziran način. Jedan od mogućih načina indeksiranja jest pomoću matrice koja prikazuje koliko se puta određena riječ iz nekog (unaprijed definiranog) rječnika pojavljuje u toj Web stranici. Indeksiranje možete zamisliti kao kratak sadržaj koji se nalazi na poleđini gotovo svake knjige - ukratko vam opiše radnju knjige prije no što odlučite želite li ju pročitati ili ne. Na takav način rade i Web tražilice - svaku Web stranicu indeksiraju (tj. ukratko opišu), te na temelju tog indeksa odluče hoće li je prikazati korisniku kada on ili ona upiše neki upit u tražilicu.

Biblioteka libcurv

Da bi se Web robot isprogramirao u programskom jeziku C/C++, u zaglavlju .c/.cpp dokumenta potrebno je uključiti biblioteku **libcurl** (koja je dio cURL aplikacije) na slijedeći način: **#include <curl/curl.h>**. Funkcije definirane u biblioteci libcurl (pogledati *Tablicu 1*) omogućuju dohvaćanje HTML (engl. *HyperText Markup Language*) sadržaja Web stranice.

Tablica 1.Funkcije biblioteke libcurl

Ime funkcije	**Opis**	**Argumenti**
curl_easy_init	Funkcija koja se prva poziva. S njom se inicijalizira cURL *session* i dobije se cURL *handle*.	CURL *curl_easy_init( ); Funkcija vraća: • CURL *handle* koji se koristi kao ulaz za sve ostale funkcije koje u nazivu imaju 'easy'[2]; • NULL u slučaju da je došlo do greške.
curl_easy_setopt	Funkcija s kojom se namještaju parametri za cURL *handle*.	CURLcode curl_easy_setopt (CURL *handle, CURLoption option, parameter); Funkcija prima: • CURL *handle*; • CURLoption opciju[3] kojom se specificira način na koji će se libcurl ponašati (opcijom **CURLOPT_URL** se dohvaća HTML sadržaj Web stranica, a opcijom **CURLOPT_WRITEDATA** se taj sadržaj zapisuje u neki tekstualni dokument, strukturu ili varijablu); • parametar koji ovisi o tome što CURLoption opcija očekuje (za opciju CURLOPT_URL parametar bi bio ime Web stranice i oznaka protokola koji se koristi da bi joj se pristupilo). Funkcija vraća: • CURLcode je cURL kod[4] koji označava tip greške do koje je došlo prilikom izvršavanja funkcije. Ako nije došlo do nekakve greške, CURLcode je CURLE_OK.

2 Naziv 'easy' se nalazi u imenima funkcija koje se koriste u jednostavnom (tzv. 'easy') cURL sučelju.

3 Popis svih CURLoption opcija možete pronaći na slijedećem link-u: http://curl.haxx.se/libcurl/c/curl_easy_setopt.html#CURLOPTURL

4 Popis svih opcija koje CURLcode može poprimiti možete pronaći na slijedećem link-u: http://curl.haxx.se/libcurl/c/libcurl-errors.html

curl_easy_perform	S ovom se funkcijom izvršava prijenos podataka.	CURLcode curl_easy_perform (CURL * handle); Funkcija prima: • cURL *handle*; Funkcija vraća: • cURL kod.
curl_easy_cleanup	Funkcija koja se posljednja poziva, i s kojom se zatvara cURL *session*.	void curl_easy_cleanup (CURL * handle); Funkcija prima: • cURL *handle*;

Biblioteka libcurl je instalirana na adrii, ali ako je želite instalirati na nekom drugom računalu sa instaliranim Ubuntu Linux operacijskim sustavom, potrebno je izvršiti slijedeće korake:

```
sudo su          // za instalaciju su vam potrebna administratorska prava
apt-get install curl                    // instalacija cURL aplikacije
apt-get install libcurl4-gnutls-dev     // instalacija developerskih dokumenata
```

Aplikaciju cURL možete koristiti direktno iz komandne linije na slijedeći način:

```
curl www.fesb.hr
```

gdje www.fesb.hr možete zamijeniti sa adresom Web stranice čiji HTML sadržaj želite dohvatiti. Nakon izvršenja ove naredbe, u terminalu (tj. konzoli) će vam se prikazati HTML sadržaj odabrane Web stranice.

Programi koji koriste libcurv biblioteku se kompajliraju i pozivaju na slijedeći način:

```
gcc -Wall -lcurl -o ime_exe_datoteke ime_datoteke.c       // C program
g++ -Wall -lcurl -o ime_exe_datoteke ime_datoteke.cpp     // C++ program
./ime_exe_datoteke
```

Dodatak: regularni izrazi

Da bi se u HTML sadržaju Web stranice pronašli svi link-ovi, taj je sadržaj najprije potrebno parsirati.

Parsiranje dokumenta se može obaviti na više načina, ali najlakše se obavlja pomoću **regularnih izraza**. Da bi se u C++ programu mogle koristiti funkcije za pisanje regularnih izraza, u zaglavlju programa potrebno je napisati: **#include <boost/regex.hpp>** (naziv **regex** dolazi od engl. ***reg**ular **ex**pressions*). Za C programe koristi se #include <regex.h>, ali sučelje za C++ je mnogo jednostavnije i ono se preporuča.

Programi koji koriste C++ biblioteku za regularne izraze se kompajliraju i pozivaju na slijedeći način:

```
g++ -Wall -lboost_regex -o ime_exe_datoteke ime_datoteke.cpp
./ime_exe_datoteke
```