ВИСОКА ПОСЛОВНА ШКОЛА
СТРУКОВНИХ СТУДИЈА – БЛАЦЕ
Часопис из области економије, менаџмента и информатике „БизИнфо“
Број 1/2010, стр. 63 – 69
Адреса: Краља Петра I, бр.70, 18420 Блаце

JEDNA REALIZACIJA WEB AGENTA

AN REALISATION OF THE WEB AGENT

Oliver Popović
Visoka poslovna škola strukovnih studija, Blace

Rezime: *U ovom radu je opisana i predstavljena programska realizacija Web agenta. To je kompjuterski program koji se koristi u pretraživanju World Wide Web-a po metodološkim, automatizovanim i jasno određenim pravilima*.
Ključne reči: *World Wide Web, Pretraživač, Crawler*

Abstract: *In this paper is described and presented a Web agent program implementation. This is a computer program used to search the World Wide Web by the methodological, automated and clearly defined rules*.
Keywords: *World Wide Web, Search engine, Crawler*

1.UVOD

Web agent (u daljem tekstu *crawler*) predstavlja program koji posećuje Web stranice i druge informacije kako bi kreirao indekse za web pretraživač. Svi pretraživači imaju takav program, koji je poznat i pod imenom „spider“ ili „bot“. *Crawleri* su obično programirani da posećuju stranice koje su označene kao nove ili nedavno ažurirane. *Crawler* je sposoban da indeksira kako celokupne sajtove, tako i pojedinačne web strane. On posećuje određene stranice, na njima pronalazi linkove ka drugim stranicama i zatim prelazi na indeksiranje reči na tim sledećim stranicama.

2. PROGRAMSKA REALIZACIJA

U okviru većeg projekta – pretraživača, praktično je realizovan jedan ovakav program koji ima sve osobine jednog *crawler*-a.
Crawler predstavlja posebni deo pretraživača, koji se sastoji iz više delova. Uzima se dodeljen fajl, skenira u potrazi za linkovima i upisuje linkove koji su pronađeni, a jos nisu upisani u bazu podataka.
Prvi fajl koji traži na nekom domenu jeste "index" sa nekim od ekstenzija: html, htm, php ili asp.
Ovo su pomoćne funkcije koje algoritam crawlera koristi u radu. Definisane su u jednom fajlu koji se uključuje u glavni program algoritma i koristi na određenim mestima.
Postoje tri funkcije koje rade na istom zadatku (*prefiks*, *sufiks* i *brojkoraka*).
Ukoliko se indeksira fajl u dubini folderske strukture, pa sadrži linkove sa relativnim adresiranjem koji vode na isti domen, relativne adrese moraju biti konvertovane u apsolutne.
Npr, ako imamo domen www.proba.com i stranicu www.proba.com/folder/podfolder/index.html koja u sebi sadrži link na ../ index.html, ovu relativnu adresu trebamo pretvoriti u apsolutnu.
Prvo funkcijom BrojKoraka(url) saznajemo koliko se koraka unazad vraćamo kroz hijerarhiju. Zatim funkcijom Otkini_prefiks "otkidamo" iz apsolutnog linka potreban broj nivoa, a funkcijom Otkini_sufiks iz tekućeg foldera "otkidamo" potreban broj nivoa, i tako dva dobijena stringa spojimo da čine apsolutnu adresu.
Funkcija uzima dva argumenta, relativnu URL adresu ($tekst) i broj nivoa koji se treba oduzeti sa desne strane relativne adrese ($brojac).
Promenljiva $tekst se razbija u niz stringova gde su graničnici znak "/". Prethodno se znak "//" konvertuje u "!!" kako bi se izbeglo tretiranje adresa koje počinju sa "http://". Na kraju funkcije se znak "!!" vraća nazad u "//".
Pošto funkcijom *explode* odvajamo reči u tekstu $tekst, funkcijom *sizeof* brojimo koliko smo elemenata dobili, zatim odbrojavamo onoliko elemenata niza koliko smo prosledili promenljivom $brojac i te članove nizova "uništimo" komandom *unset*.
Posle toga vraćamo za pretraživač interesantne elemente niza funkcijom *implode* u prvobitno stanje – spajamo ih znakom "/" i vraćamo rezultat.

```
function Otkini_sufiks($tekst, $brojac) {
	str_replace("//", "!!", $tekst);
	$put = explode("/", $tekst);
	$velicina = sizeof($put);
	for ($i = 0; $i <= $brojac; $i++){
	unset($put[$velicina-$i]);
	}
	$rezultat = implode("/", $put);
	str_replace("!!", "//", $rezultat);
	return $rezultat;
}
```

Funkcija za otkidanje prefiksa radi analogno prethodnoj funkciji i dobija iste parametre. Parametar *brojac* se umanjuje za 1 jer se elementi niza broje od nule.

```
function Otkini_prefiks($tekst, $brojac) {
	$brojac--;
	$put = explode("/", $tekst);
	$velicina = sizeof($put);
	for ($i = -1; $i <= $brojac; $i++){
	unset($put[$i]);
	}
	$rezultat = implode("/", $put);
	return $rezultat;
}
```

Funkcija *Ekstenzija* uzima tekst linka i proverava da li ovaj vodi na neki tip sadržaja koji se može indeksirati. Ukoliko link sadrži interni link - sidrište "#" biće ignorisan.
Zatim, postoji mogućnost da link ima iza ekstenzije fajla i URL parametre iza znaka "?". Zbog toga se string razdvaja funkcijom explode po znaku "?" i uzima se samo njegov prvi deo za dalji rad.
Npr., razmotrićemo

```
proba/moja/index.html?vest=5
```

Ovako pripremljen string funkcijom *explode* razbijamo po parametru "." da bismo izdvojili ekstenziju fajla. Pošto nikad ne možemo da znamo na kojoj će se poziciji nalaziti ekstenzija, članove ovako dobijenog niza obrnemo, tako da ekstenzija dođe na poziciju 0.

Ukoliko je ekstenzija tipa html, htm, php ili asp, funkcija vraća prosleđen joj url što omogućava nastavak rada. U suprotnom, funkcija vraća nulu.

Ukoliko funkcija pronađe ekstenziju koja ima više od 4 znaka, postoji mogućnost da link uopšte ne vodi na konkretan fajl već da pokazuje u neki od foldera, gde bi trebalo potražiti neki od standardnih index fajlova. Funkcija će probati da u datom folderu pronađe index.html ili index.htm ili index.php ili index.asp, i ako nađe neki od ovih fajlova, funkcija će ga sasvim legalno vratiti kao rezultat.

```
function Ekstenzija($tekst) {

if(stristr($tekst, "#")) {
return 0;
}
	$ignorisanje_url_parametara = explode("?", $tekst);
	$put = explode(".", $ignorisanje_url_parametara[0]);
	$put = array_reverse($put);
	if (strlen($put[0])>=4) {
	if ($input = @file_get_contents($tekst. "/index.html"
))	{ $tekst = $tekst . "/index.html"; return $tekst; }
	if ($input = @file_get_contents($tekst . "/index.htm"
))	{ $tekst = $tekst . "/index.htm"; return $tekst; }
	if ($input = @file_get_contents($tekst . "/index.php"
))	{ $tekst = $tekst . "/index.php"; return $tekst; }
	if ($input = @file_get_contents($tekst . "/index.asp"
))	{ $tekst = $tekst . "/index.asp"; return $tekst; }
	}
	if	(($put[0]=='html')	||	($put[0]=='htm')	||
($put[0]=='php') || ($put[0]=='asp')) return $tekst;
	else return 0;
}
```

Funkcija BrojKoraka uzima jedan parametar i broji kod dostavljene relativne adrese koliko puta se pojavljuje izraz "../" tj. koliko foldera unazad kroz hijerarhiju relativna adresa vraća.
Dobijeni tekst razbijemo funkcijom explode u niz $put, inicijalizujemo brojač i prebrojimo elemente niza koji glase "..". Funkcija vraća broj takvih elemenata.

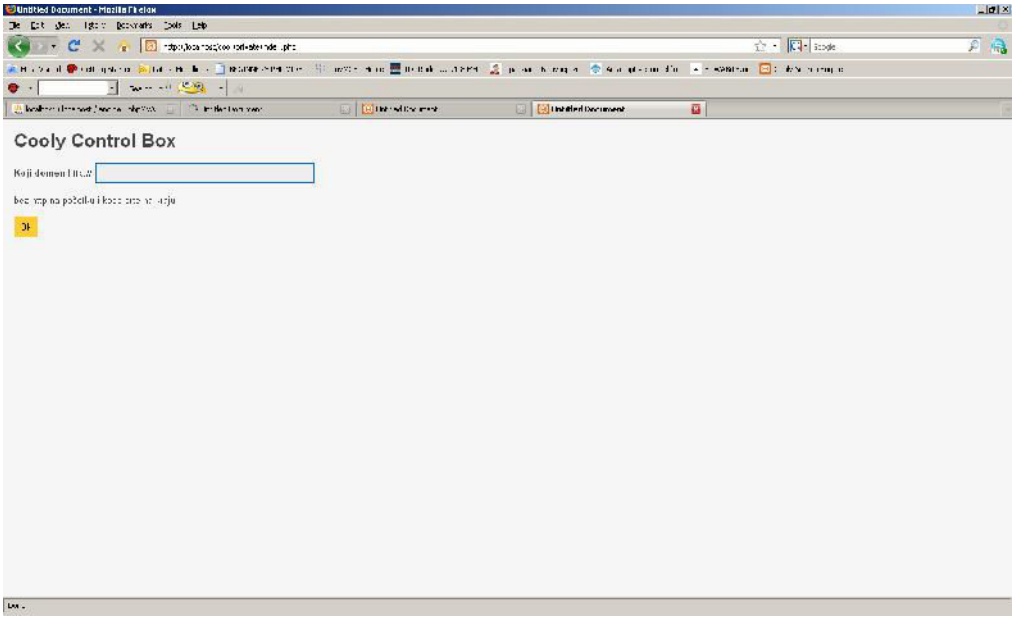
```
function BrojKoraka($tekst) {
	$put = explode("/", $tekst);
	$broj = 0;
	foreach ($put as $p) {
	if ($p == "..") $broj++;
	}
```

```
      return $broj;
}
?>
.
.
.
<?php
function indeksiraj($domen, $url){
echo "<p> ... indeksiram " .  $url . " ";
// ova funkcija odabranu stranicu iscitava u flat text i
title tag
$duz_domena = strlen($domen);
$obradjujem = Otkini_sufiks($url, 1);

$input = @file_get_contents($url);
$regexp              =              "<a\s[^>]*href=(\"??)([^\"
>]*?)\\1[^>]*>(.*)<\/a>";
eregi("<title>(.*)</title>", $input, $title);
$upit = "UPDATE search SET status = 2, title='" . $title[1] .
"',                          text='"                          .
mysql_real_escape_string(trim(strip_tags($input))) . "' WHERE
url = '". $url .  "' ";
.
.
.
.

}

function iniciraj($url, $domain){
                          if    (@file_get_contents($url    .
"/index.html")) $parametar = $url . "/index.html";
                          if    (@file_get_contents($url    .
"/index.asp")) $parametar = $url . "/index.asp";
                          if    (@file_get_contents($url    .
"/index.php")) $parametar = $url . "/index.php";
                          if    (@file_get_contents($url    .
"/index.htm")) $parametar = $url . "/index.htm";
                          $dodajURL  =  "INSERT  INTO  search
VALUES ('". $parametar .  "', '...', '...', 1)";
                          mysql_query($dodajURL);
                          echo  "<p>URL  Iniciran.  Sada  idemo
stranicu po stranicu...</p>";
}
?>
```



Slika 1. Unos domena koji će biti indeksiran

3.ZAKLJUČAK

U ovom radu je predstavljen jedan od načina realizacije agenta pretraživanja - crawlera. Agent pretraživanja se koristi u svakom Web pretraživaču, i to je jedan od njegovih najbitnijih delova. U zavisnosti od njegove programske i logičke strukture, zavisi i brzina i relevantnost pretrage Weba. Što je pretraga relevantnija, to je prikaz rezultata verodostojniji za krajnjeg korisnika.

LITERATURA

1. Baeza-Yates R., *Crawling a Country: Better Strategies than Breath-First for Web Page Ordering*, Proceedings of International World Wide Conference, Chiba, Japan, 2005
2. Bar-Yossef Z., Gurevich M., *Random Sampling from Search Engine's Index*, Proceedings of International World Wide Web Conference, Edinburgh, Scotland, 2006
3. Hawking D., *Web Search Engines: Part 2*, CSIRO ICT Centre http://www.computer.org/portal/site/computer/menuitem.5d61c1d591162e4b0ef1bd108bcd45f3/index.jsp?&pName=comput

er_level1_article&TheCat=1055&path=computer/homepage/0806&file=howthings.xml&xsl=article.xsl&, 2006
4. Manning D. C., Prabhakar Raghavan, Hinrich Schtze, *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, UK, 2008
5. Popović O., *Web pretraživači,* master rad, 2009
6. Zobel J., Moffat A., *Inverted Files for Text Search Engines*, ACM Computing Surveys, 2006
7. http://en.wikipedia.org/wiki/Web_crawler
8. http://searchsoa.techtarget.com/sDefinition/0,,sid26_gci211854,00.html