

# Robotii Web

## Sabin-Corneliu Buraga

Articol aparut in *PC Report*, vol.9, nr.92, mai 2000

---

### 1. Prezentare generala

**Robotii Web**, regasiti si sub numele de paianjeni (*spiders*), reprezinta programe care traverseaza în mod automat structura hipertext a spatiului WWW, în scopuri de extragere a informatiilor folosind protocoalele Web standard.

Trebuie facuta o distinctie clara între robotii Web si navigatoarele Web care sunt aplicatii actionate de om sau între robotii Web si agentii Web care se bucura si de alte caracteristici, precum inteligenta, mobilitatea ori autonomia.

Activitatea unui robot Web este cea de a realiza o conexiune HTTP (HyperText Transfer Protocol) la un server Web continând un set de pagini, pornind de la un identificator uniform de resurse (URI), de a extrage informatiile dintr-un document HTML si din toate documentele desemnate de legaturile lui.

### 2. Utilizari

Robotii Web se pot utiliza în scopuri multiple, dintre care amintim:

- **analiza statistica**

Prin traversarea unui întreg site Web, un robot poate furniza date privind media documentelor stocate pe un server, procentul de documente de un anumit tip, marimea medie a unei pagini, gradul de interconectare cu alte documente, locale sau la distanta etc. În fapt, primul robot implementat a fost conceput cu scopul de a contoriza serverele Web din lume, pentru a se cerceta cât de întins este spatiul hipertext.

- **mentinere**

În prezent, este foarte important a se mentine în bune conditiuni starea hiperlegaturilor dintre documentele Web. Un robot poate ajuta la descoperirea si rezolvarea asa-numitelor "legaturi moarte" care pointeaza spre resurse inexistente. Desi serverele Web pot înregistra toate cererile care nu pot fi satisfacute din cauza specificarii adreselor invalide, administratorul unui site Web de proportii poate recurge la asistarea din partea unui robot (ca **MOMSpider**) pentru descoperirea automata a

legaturilor eronate.

Robotii pot verifica si structura documentelor HTML, semnalînd erorile de design si de stil ale acestora.

- **oglundire**

Tehnica oglindirii ( *mirroring* ) este preluata de la arhivele FTP, reprezentînd copierea la alta locatie a întregii structuri arborescente, în mod recursiv, a fisierelor unui site si reactualizarea periodica a acestora. Aceasta asigura fiabilitate, viteza mai mare de transfer, evitarea încarcarii traficului de retea sau acces neconectat (*off-line*).

Pentru Web, oglindirea poate fi realizata de un robot, care trebuie sa aiba grija de rescrierea referintelor la alte documente, la pastrarea integritatii hipertextului si la actualizarea regulata a paginilor WWW.

Oglindirea poate fi evitata, folosind în mod inteligent memoria *cache* a serverelor *proxy* (intermediare), care ofera posibilitati de actualizare selectiva si de organizare a resurselor.

- **descoperirea resurselor**

Probabil cea mai spectaculoasa si importanta aplicatie a robotilor Web este utilizarea lor la descoperirea resurselor. Cresterea progresiva a volumului de informatii a dus la necesitatea conceperii de aplicatii pentru sumarizarea, indexarea, supraveghierea modificarilor informatiilor de pe Web.

Astfel, fiecare motor de cautare, continînd baze de date privind localizarea si tipul de informatii dorite de utilizatori, apeleaza la serviciul robotilor Web pentru descoperirea resurselor Internet.

Un avantaj suplimentar este cel dat de monitorizarea modificarilor survenite în cadrul paginilor Web (servicii "Mind-It" sau "What's new").

- **utilizari combinate**

Desigur, robotii Web pot îndeplini sarcini multiple, ca de exemplu descoperirea resurselor si realizarea de statistici Web sau mentinerea integritatii legaturilor si, concomitent, detectarea schimbarilor documentelor HTML.

### 3. Pericole si costuri de utilizare ale robotilor

Prin traversarea unui numar mare de hiperlegaturi, robotii necesita o largime buna de banda, deoarece ei pot opera continuu perioade lungi de timp (saptamîni sau chiar luni). Pentru a accelera aceste operatii, multi roboti au implementate tehnici de extragere paralela a datelor, metoda denumita *operare în foc rapid* ( *rapid fire* ), rezultînd un trafic considerabil (o încetinire temporara a transferului de date). Mai mult, serverele Web pot fi supraîncarcate de cereri multiple de accesare venite din

partea robotilor în detrimentul cererilor agentilor utilizator. Asa sa / book / web / Implementarea robotilor permitînd foc rapid trebuie evitata.

Un alt aspect care trebuie luat în considerare este timpul de actualizare a bazelor de date ale motoarelor de cautare folosind pentru descoperirea resurselor robotii Web. Robotii de cautare a informatiilor vor trebui sa decida care informatii sunt importante a fi transmise programelor de indexare.

Un alt pericol deriva din exploatarea necontrolata a robotilor Web de catre utilizatorii finali care pot folosi optiuni inadecvate sau pot rula mai multe instante de program, conducînd la abuzuri nedorite.

Robotii Web, în special cei netestati îndeajuns, pot fi ineficienti sau pot poseda vicii de arhitectura si astfel sa dauneze traficului de informatii pe Internet, mai ales daca sunt folositi de persoane neavizate ori de neprofesionisti. Implementari eronate pot determina robotii sa intre în arii aproape infinite denumite *gauri negre* (atunci cînd de exemplu un document are o legatura care se refera la el insusi, iar programul nu detecteaza acest aspect). De asemeni, robotii nu trebuie sa acceseze tipuri de date fara relevanta, avînd dimensiuni considerabile (e.g. arhive, fisiere executabile, fisiere multimedia etc.).

## 4. Tipuri de roboti Web

Criteriile de clasificare a robotilor sunt multiple, vom încerca sa prezentam robotii Web dupa activitatile care pot sa le realizeze.

- a. **roboti academici** - sunt acei roboti disponibili în medii academice, avînd ca scop anumite activitati de colectare a datelor dintr-o universitate sau de mentinere a integritatii legaturilor dintr-un site academic.
- b. **roboti de proiectare** - poseda capabilitati de proiectare si de asistare în activitatile de design a paginilor Web sau de proiectare a altor tipuri de roboti.
- c. **roboti conversationali** - ofera un partener virtual de discutii în Internet, fiind de obicei integrati serviciilor de teleconferinte pe Web. Ca exemplu, putem mentiona **Eliza**.
- d. **roboti de comert** - sunt roboti înlesnind activitatile de comert electronic, licitatii pe Web, bursa etc.
- e. **roboti distractivi** - ofera diverse facilitati de amuzament (jocuri, predictii, recomandari de adrese interesante etc.) pentru utilizatorii care parcurg paginile Web.
- f. **roboti guvernamentali** - sunt acei roboti vizînd servere continînd informatii de interes guvernamental sau diplomatic ori cu caracter secret.
- g. **roboti inteligenti** - manipuleaza informatii, posedînd inteligenta artificiala, fiind utilizati pentru explorarea inteligenta a resurselor Web (e.g. **Harvest** sau **W3QS**).
- h. **roboti de stiri** - monitorizeaza grupurile de stiri de pe Internet, modificarile din cadrul site-urilor mass-media (ziare electronice, posturi radio sau de televiziune prezente pe Web etc.), schimbarile de adrese si altele.
- i. **roboti de cautare** - sunt robotii utilizati de motoarele de cautare (ca de

exemplu **WebCrawler**).

- j. **roboti de actualizare** - se folosesc pentru actualizarea automata a hiperlegaturilor si pentru detectia schimbarii adreselor Web.

## 5. Catalogarea informatiilor utilizând roboti

În prezent asistam la o dezvoltare exploziva a spatiului WWW, într-o faza de existenta a prea multor informatii, cu un continut prea dinamic, conducând la generarea unui haos în Internet.

Procesul de regasire a informatiilor se bazeaza pe faptul ca din întreg volumul de date numai o fractiune reprezinta documente relevante pentru utilizator. Cea mai populara tehnica este cea a indexarii documentelor pe baza **cuvintelor cheie** furnizate fie explicit de creatorul acestor documente, fie în urma unei catalogari automate realizate de un robot. Cautarea se realizeaza folosind algoritmi de parcurgere locala de tip DFS sau BFS sau prin procesarea într-o ordine inteligenta a legaturilor spre alte documente.

Întrebarile care se pun sunt:

- i. Cît de relevante sunt activitatile de indexare si de sumarizare automate?
- ii. Documentelor HTML le pot fi atasate anumite attribute care sa le descrie continutul?

Raspunsul la ultima întrebare este dat de urmatoarele aspecte:

- Standardul HTML permite autorilor de pagini WWW sa enumere cuvintele cheie care sa descrie cel mai adecvat continutul lor informational, prin folosirea în antet a tag-ului **META**. Iata un exemplu:

```
<meta name="description" -- descriere succinta --
      content="Manipularea informatiilor multimedia">
<meta name="keywords"    -- cuvinte cheie --
      content="Internet, multimedia, hypertext, document, robots,
agents, logic, real-time, protocol, WWW, Web">
<meta name="author"     -- autor --
      content="Sabin Corneliu Buraga <busaco@infoiasi.ro>">
<meta name="owner"     -- proprietar --
      content="Sabin Corneliu Buraga <busaco@infoiasi.ro>">
```

- O metoda complementara este utilizarea de tag-uri ascunse care sa fie exploatare de diversi roboti (de pilda tag-urile speciale ale programului Teleport).
- Descrierea inteligenta a resurselor poate fi realizata cel mai bine cu **RDF (Resource Description Framework)**. Deja exista tehnici de generare automata a metadatelor RDF pentru descoperirea resurselor Web, pornind de la titluri, cuvinte cheie, descrieri, data crearii, numarul total de cuvinte dintr-un document etc.
- Pot fi folosite diverse alte metode (euristici): explorare structurala, utilizarea retelelor neuronale sau a algoritmilor genetici etc.

Pentru excluderea robotilor din zone Web lipsite de relevanta, continând date temporare ori private, s-a adoptat un **standard pentru excluderea robotilor**. Acest standard se bazeaza pe accesarea unui fisier text `robots.txt` (stocat pe serverul

Web) de catre un robot de explorare, fisierul `specifinfoias.ro/paisaco/robots.txt` este de la parcurgerea automata (pentru evitarea gaurilor negre sau din alte ratiuni).

Un exemplu de astfel de fisier este cel de mai jos:

```
#/robots.txt pentru http://www.infoiasi.ro
User-agent: *                # toti robotii
Disallow: /tmp/              # date temporare
Disallow: /busaco/work/     # spatiu privat
```

În vederea evitarii indexarii continutului unei pagini Web se poate scrie în antetul ei:

```
<meta name="robots" content="noindex">
```

În activitatea de catalogare a informatiilor, de multe ori intervine ierarhizarea datelor în functie de subiectul pe care-l trateaza, aceasta clasificare ducând la aparitia *serviciilor director* (de genul GENVL). Clasificarea dupa subiect este similara retelei lingvistice WordNet.

Un robot de indexare a informatiilor poate sa se confrunte cu diverse probleme precum mutarea URI-urilor, cautarea într-o oglindire si nu într-o locatie originala, duplicarea legaturilor si altele.

Un utilizator poate apela la un serviciu de înregistrare automata la o suita de motoare de cautare, de obicei gratuit, ca de exemplu [AddMe!](#).

## 6. Sfaturi în conceperea unui robot Web

Înainte de a purcede la conceperea unui robot, folosind sau nu metode de calcul al hiperinformatiei, trebuiesc avute în vedere urmatoarele:

- Chiar avem nevoie de un alt robot sau adoptam nevoilor noastre unul deja existent?
- Robotul trebuie identificat usor de administratorul Web si autorul acelui robot trebuie sa fie contactat facil.
- Robotul va trebui sa fie mai întâi testat pe date locale, într-o retea proprie, înainte de a fi disponibil în Internet.
- Robotul va fi moderat în ceea ce priveste resursele: prevenirea focului rapid, eliminarea cautarilor redundante si inutile.
- Robotul se va conforma standardului de excludere a robotilor.
- Autorul robotului va analiza continuu activitatile robotului propriu.
- Rezultatele furnizate de robot (diversele statistici sau alte date) vor putea fi facute disponibile spre consultare celor interesati.

Majoritatea robotilor Web actuali respecta recomandarile de mai sus.

## 7. Exemple

### 7.1 *DataBots* - un robot manipulând informatii

Creatie a companiei Imagination Engines Incorporated, **DataBots** este un robot utilizat în descoperirea informatiilor pe Web, folosind tabelele Excel pentru generarea unor retele neuronale menite a analiza datele luate din Internet. Paradigma utilizata este denumita *Creativity Machine*, permitînd luarea decizilor asistata de calculator pe baza informatiilor extrase de robotul Web.

## 7.2 *Wisebot* - utilizat pentru administrarea site-urilor Web

**Wisebot** este o aplicatie a companiei Tetranet, putînd fi utilizat la organizarea unui site Web. Robotul efectueaza o analiza a tuturor paginilor Web existente pe un server, retinînd într-o baza de date cele mai importante informatii despre ele (titlu, cuvinte-cheie, timpul ultimei actualizari etc.). Robotul poate automat genera cuvintele cheie aferente unei pagini, utilizînd tehnologia Extractor, prin contorizarea celor mai frecvente cuvinte si luînd în considerare pozitia lor în cadrul paginii. Astfel, se construiesc un index pentru întreg continutul hipertext al serverului, care va fi sursa pentru harta de navigare în cadrul site-ului (actualizata la momente regulate de timp în mod automat).

Acest robot este disponibil în mediile Windows. Pentru platformele UNIX, exista o multitudine de roboti similari, dintre care se pot mentiona **tkWWW** (scris în Tcl/Tk) sau **RBSE (Repository Based Software Engineering)**.

## 7.3 *Inktomi* - statistici Web

Robotul de cautare **Inktomi** a fost conceput în cadrul unui proiect de cercetare condus de Eric Brewer si Paul Gautier de la Universitatea Berkeley, cu scopul de a utiliza tehnicile de procesare paralela pentru indexarea, cautarea si analiza paginilor Web. Primele seturi de date au fost preluate în perioada *iulie-octombrie 1995* colectîndu-se 1,3 milioane de documente HTML unice, urmate în *noiembrie 1996* de 2,6 milioane de documente HTML.

### 7.3.1 Experimentul

În cadrul prelucrării datelor colectate de robot, s-au folosit urmatoarele aplicatii:

- **libink** este o biblioteca de componente pentru extragerea si manipularea datelor hipertext, constînd din patru module principale:
  - **analizorul HTML** este un scanner lexical inspirat din **flex**, configurabil si rapid;

- *analizorul URI* este un analizor al identificatorilor unice de resurse (URI);
- *translatorul DNS* converteste adresele simbolice ale serviciului numelor de domenii (DNS) în adrese IP numerice, utilizînd o memorie *cache* suplimentara;
- *serviciile generale pentru tabele hash* sunt folosite în cadrul procesului de prelucrare a datelor, implementînd tabele *hash* distribuite.
- *style* este un program standard UNIX raportînd diverse proprietati statistice (lungimea medie a unei fraze, numarul mediu de propozitii dintr-o fraza complexa, numarul total de cuvinte etc.) utile analizei documentelor din perspectiva limbajului natural. În cadrul experimentului au fost considerate doar documente scrise în limba engleza.
- *weblint* este un analizator structural de marcaje pentru documentele HTML, inspirat din utilitarul UNIX *lint*.

### 7.3.2 Rezultatele

Criteriile de analiza au fost urmatoarele:

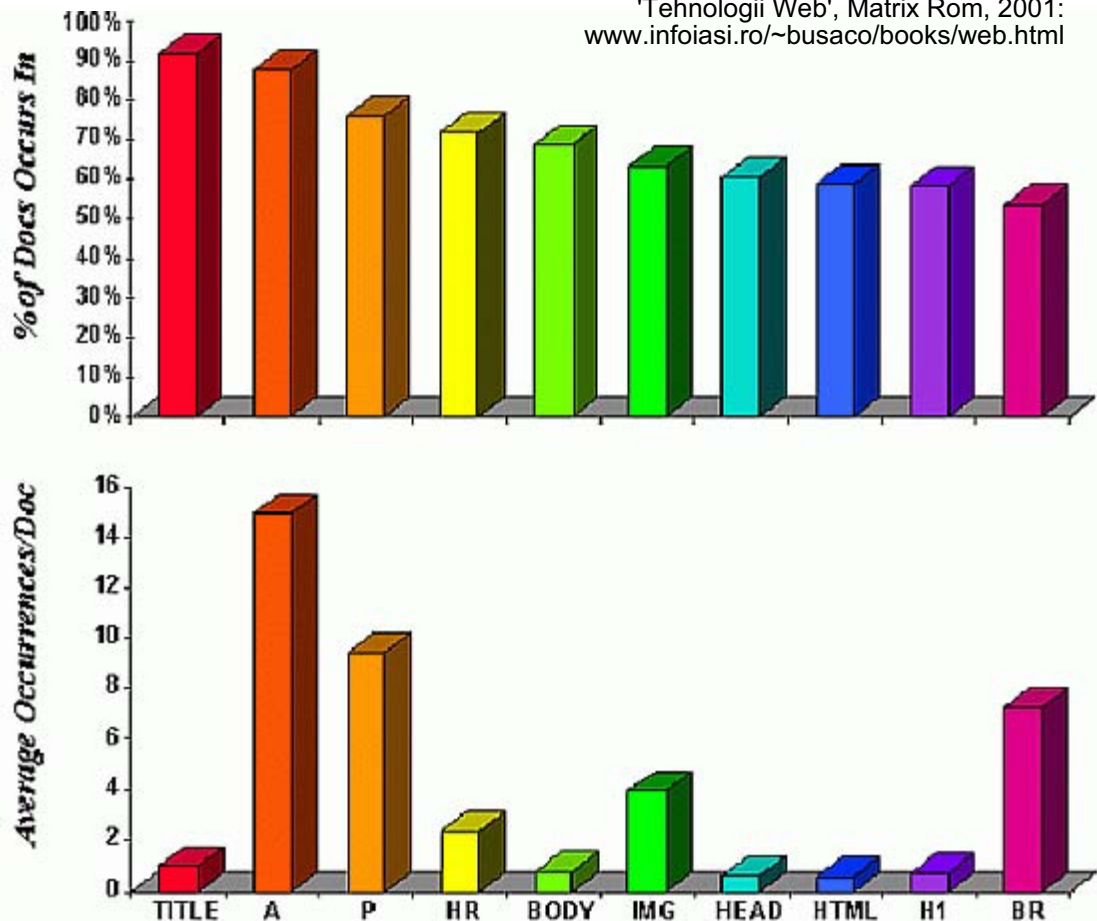
- lungimea documentelor
- media numar de marcatori/lungimea paginii
- utilizarea marcatorelor
- utilizarea atributelor
- utilizarea marcatorelor specifici unor navigatoare particulare
- utilizarea portului de conectare
- protocoalele în cadrul URI-urilor
- tipurile fisierelor specificate în componenta URI-urilor
- erorile de sintaxa

Iata cîteva dintre rezultatele obtinute.

- **Lungimea documentelor**

Pentru cele 2,6 milioane de documente HTML colectate de Inktomi, dupa înlaturarea marcajelor s-a calculat lungimea fiecarui document. Lungimea minima gasita a fost de *4,4 KB*, lungimea maxima de *1,6 MB*, iar lungimea medie a fost de *2,0 KB*.
- **Utilizarea marcatorelor**

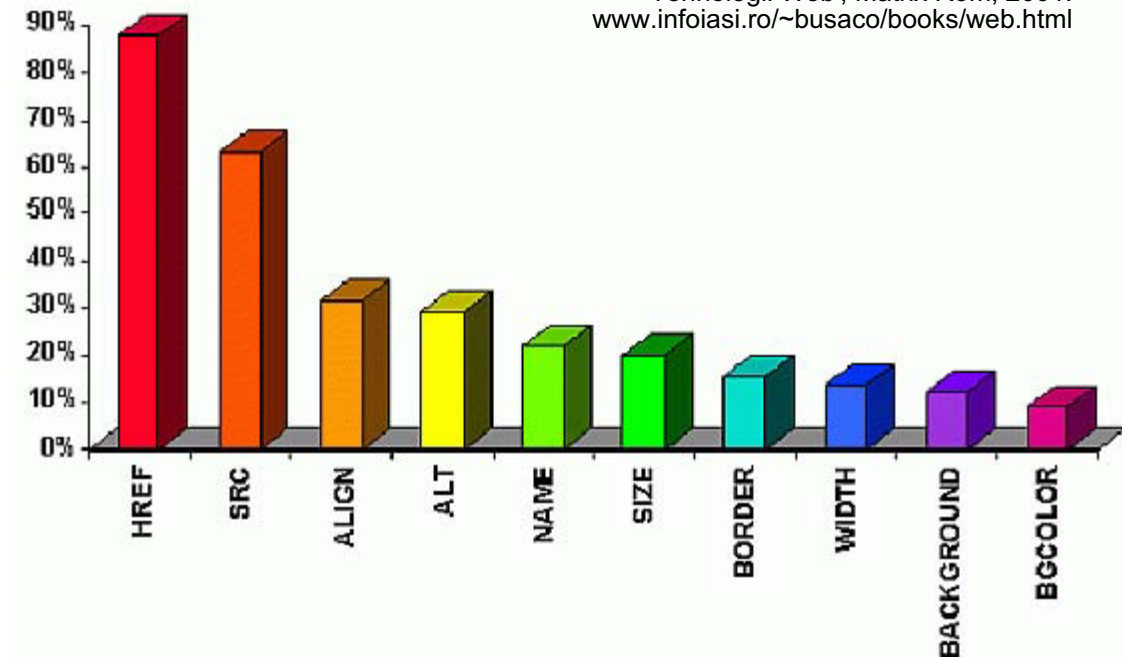
În ceea ce priveste distributia tag-urilor, numarul mediu de marcatori pe document a fost *71*, iar numarul de marcatori unici pe document a fost *11*. În figura de mai jos se poate remarca ponderea în procente a celor mai populare 10 tag-uri si numarul mediu de aparitii ale lor.



Cele mai utilizate 10 tag-uri

- o **Utilizarea atributelor**

Numarul mediu de attribute prezente într-un document a fost 29 , iar numarul mediu de attribute unice pe document a fost 4. În figura 9 se poate observa procentul de aparitie a primelor 10 attribute (firesc, atributul cel mai popular a fost [href](#), urmat de [src](#)).



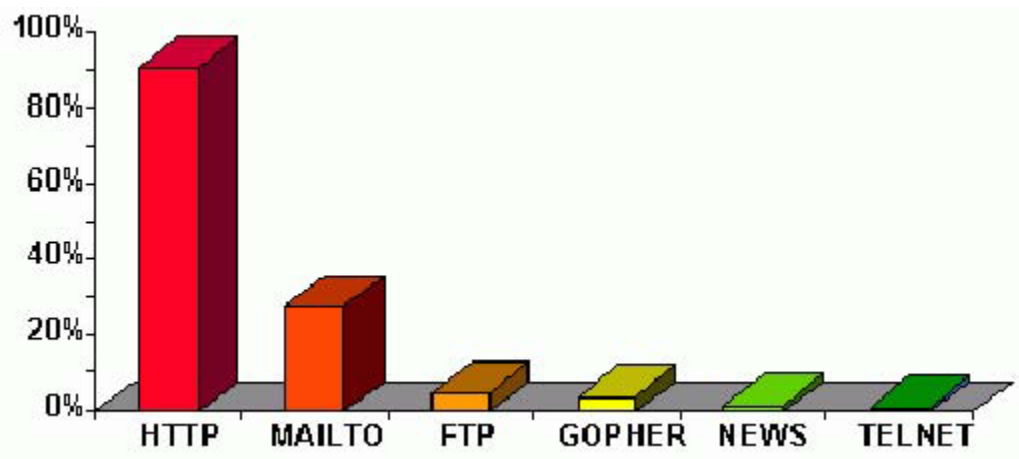
Cele mai utilizate 10 atribute

- **Utilizarea portului de conectare**

Protocolul de transfer HTTP uzual foloseste portul 80 pentru accesarea paginilor Web. Acest port este utilizat în proportie de 93,6%. Numarul de porturi unice specificate în documentele hipertext a fost 418 .

- **Protocoalele în cadrul URI-urilor**

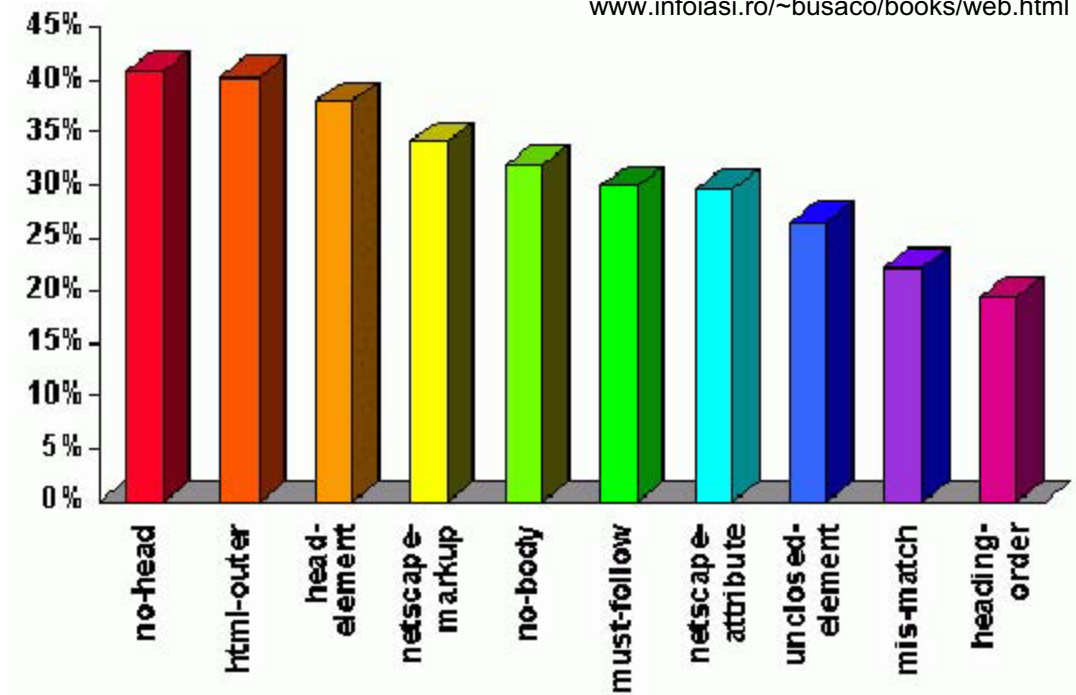
Extragînd URI-urile specificate în fiecare pagina Web, s-au putut calcula procentele de aparitie a celor mai utilizate protocoale: HTTP, SMTP (mailto), FTP, GOPHER, NNTP (news) si TELNET.



Frecventa de utilizare a protocoalelor

- **Erorile de sintaxa**

Programul [weblint](#) a gasit aproximativ 92000 (0,7%) de documente continînd erori sintactice. Figura urmatoare furnizeaza primele 10 cele mai comune erori.



Cele mai frecvente 10 erori detectate

*Legenda:*

- *html-outer* - nu exista tag-urile `<html>...</html>`
- *no-head* - lipseste elementul `<head>`
- *head-element* - tag-urile specifice antetului paginii Web (`<title>`, `<link>`, `<base>` sau `<base>`) apar în afara acestuia
- *no-body* - lipseste elementul `<body>`
- *must-follow* - lipsesc marcatorii obligatorii în cadrul altui tag
- *unclosed-element* - tag-urile de sfîrsit lipsesc
- *netscape-markup* - marcaje specifice Netscape (nu trebuie considerata neaparat o eroare, ci mai mult o abatere de la standardul HTML definit de Consorțiul Web)
- *empty-container* - elemente vide, fara continut
- *mis-match* - tag-uri nepotrivite (de exemplu `<h2>...</h3>`)
- *heading-order* - ordine inadecvata a elementelor de tip Hx

## 7.4 LiveAgent Pro - pentru realizarea oglindirilor Web

Dezvoltat de AgentSoft, robotul **LiveAgent Pro** este destinat efectuării automate a oglindirilor unui server Web la alta locație. Alte facilitati ale aplicatiei sunt completarea automata a formularelor electronice, posibilitatea de programare prin intermediul scripturilor sau utilizarea robotului în Intranet.

## 8. Referinte bibliografice

1. A.Ardö, S.Lundberg - "A regional distributed WWW search and indexing service - the DESIRE way", WWW7 Conference Proceedings, 1998

- 'Tehnologii Web', Matrix Rom, 2001:  
[www.informatica.usab.ro/WWW8/WWW8.html](http://www.informatica.usab.ro/WWW8/WWW8.html)
2. S.Brin, L.Page - "*The Anatomy of a Large-Scale Hypertextual Web Search Engine*", WWW8 Conference Proceedings, Canada, Elsevier Science, May 1999
  3. C.Gütl, K.Andrews, H.Maurer - "*Future Information Harvesting and Processing on the Web*", European Telematics: advancing the information society Proceedings, Barcelona, feb.1998
  4. F.Heylighen, J.Bollen - "*The World-Wide Web as a Super-Brain: from metaphor to model*", in R.Trapp (ed.) - "Cybernetics and Systems", World Science, Singapore, 1996
  5. C.Jenkins et al. - "*Automatic RDF Metadata Generation for Resource Discovery*", WWW8 Conference Proceedings, Canada, Elsevier Science, May 1999
  6. M.Koster - "*Robots in the Web: threat or treat?*", ConneXions, Volume 9, No. 4, April 1995 (1997: Updated links and addresses):  
<http://info.webcrawler.com/mak/projects/robots/threat-or-treat.html>
  7. M.Marchiori - "*The Quest for Correct Information on the Web: Hyper Search Engines*", WWW6 Conference Proceedings, France, Elsevier Science, 1997:  
<http://www.scope.gmd.de/info/www6/technical/paper222/paper222.html>
  8. A.Woodruff et al. - "*An Investigation of Documents from the World Wide Web*", Computer Science Division, University of California at Berkeley, 1997:  
<http://www.cs.berkeley.edu/~woodruff/inktomi/>
  9. \* \* \* - "*AddMe!*": <http://www.addme.com>
  10. \* \* \* - "*AltaVista*": <http://www.altavista.com>
  11. \* \* \* - "*Bot Spot*": <http://bots.internet.com>
  12. \* \* \* - "*Eliza*": <http://www-ai.ijs.si/eliza/eliza.html>
  13. \* \* \* - "*Excite*": <http://www.excite.com>
  14. \* \* \* - "*Harvest*": <http://harvest.transarc.com>
  15. \* \* \* - "*Inktomi*": <http://www.inktomi.com>
  16. \* \* \* - "*Lycos*": <http://www.lycos.com>
  17. \* \* \* - "*W3QS*": <http://www.cs.technion.ac.il/~konop/w3qs.html>
  18. \* \* \* - "*WebCrawler*": <http://www.webcrawler.com>
  19. \* \* \* - "*Webopedia*": <http://webopedia.internet.com/TERM/r/robot.html>
- 

Sabin-Corneliu Buraga este doctorand în *Computer Science* la Universitatea "A.I.Cuza" Iasi, putînd fi contactat la adresa e-mail [busaco@infoiasi.ro](mailto:busaco@infoiasi.ro).