SERVICII DE STRUCTURARE A CONŢINUTULUI PAGINILOR WEB

Viorel CĂRBUNE, Andrei NICU, Dorin TURTUREAN, Irina MOROI, Liudmila CAZAC

Universitatea Tehnică a Moldovei

Abstract: *În acest articol s-a descris o metodologie de structurare a conţinutului paginilor Web care oferă o gamă largă de posibilităţi pentru Web developeri de a grupa conţinutul din pagină în dependenţă de tipul acestuia, cu scopul de a optimiza indexarea paginilor Web de către motoarele de căutare. Pe lîngă aceasta s-au identificat potenţialele avantaje ale aplicării acestei metodologii la trecerea de la Web-ul clasic la cel semantic. Deasemeni s-a argumentat creşterea productivităţii procesului de comunicare dintre server şi motoarele de căutare.*

Cuvinte cheie: *pagină Web, motor de căutare, Web semantic, Web developer, crawler, html, server.*

Odată cu apariţia Internetului a apărut şi posibilitatea schimbului de date dintre utilizatorii acestuia. Apariţia limbajului HTML a fost ca urmare a necesităţii prezentării adecvate a aceluia-şi conţinut pe diferite maşini de calcul. Astfel informaţia postată de către autor putea fi disponibilă cititorilor din întreaga lume indiferent de configuraţia maşinilor de calcul de care dispuneau aceştia. Limbajul HTML – Hyper Text Markup Language drept răspuns la cerinţele clienţilor începe să evolueze şi pe lîngă text apare posibilitatea de a plasa pe pagini Web imagini, video, audio, forme şi multe altele, astfel abătîndu-se de la sarcina iniţială pentru care a fost conceput – de a marca textul, şi capătă ca urmare tot mai multe funcţionalităţi secundare. Din aceste considerente s-a ajuns la etapa în care conţinutul fişierelor HTML a devenit neomogen. Lucrul Web developerilor cu astfel de fişiere a devenit mult mai anevoios. O tentativă de separare funcţională a fost divizarea stilurilor de restul conţinutului paginii prin introducerea fişierelor CSS în arhitectura paginilor Web. Această structură arhitecturală oferă posibilitatea de a schimba modul de afişare a conţinutului fără a modifica însuşi conţinutul. Necătînd la aceste posibilităţi limbajul HTML îşi păstrează imunitatea în ceea ce priveşte conţinutul.

Odată cu dezvoltarea tehnologiilor Web au apărut şi alte tipuri de utilizatori decît cei clasici şi anume roboţii de căutare – utilizatori intermediari. Necătînd la faptul că fiecare tip de utilizator are nevoie de tipuri de date diferite, Web serverele nu ţin cont de acest lucru şi asigură transferul integral de date prin protocolul HTTP livrînd tot conţinutul paginii Web către client. Din aceste considerente ar fi mai eficient ca Web serverul să fie capabil de a oferi doar informaţia necesară sau în cel mai rău caz – întreaga categorie din care face pate informaţia cerută de către client . Pentru îndeplinirea acestui scop se recomandă gruparea conţinutului paginii Web după careva categorii **Figura 1**.

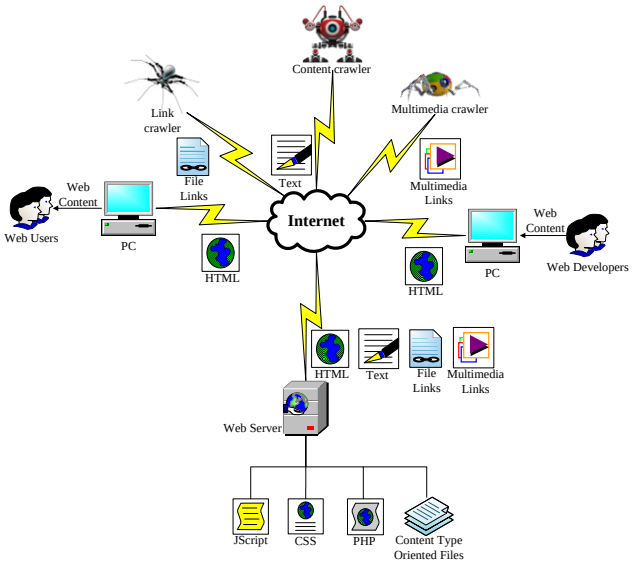

Fig. 1: Traficul tiputilor de conţinut în internet.

Din **Figura 1** se observă faptul că Web Users sau utilizatorii clasici sunt deserviţi în continuare prin aceeaşi metodă, adică la cererile acestora serverul răspunde prin a oferi informaţia deplină conţinută pe pagina Web. Spre deosebire de utilizatorii clasici, utilizatorii de tip Link crawler sau roboţii de analiză a referinţelor nu au nevoie de conţinutul textual de pe pagină şi nici de codul HTML integral ci doar de referinţe. Pe utilizatorii de tip Content crawler i-ar interesa doar conţinutul textual. Grupul de utilizatorii Multimedia crawler înglobează în sine mai multe tipuri de roboţi: de căutare a imaginilor, a clipurilor video, a fişierelor audio etc.

Procesul de structurare a conţinutului este lăsată pe seama Web developerilor, însă se presupune că informaţia din aceeaşi categorie este amplasată în acelaşi fişier ce poartă denumirea paginii din care este extrasă iar extensia este definită de categoria din care face parte.De exemplu dacă din fişierul index.html se extrage întreg conţinutul textual, acesta este plasat în fişierul real sau virtual index.cnt. Analogic index.ref, index.img şi index.vid vor indica conţinutul de tip adrese de referinţe, căi către imagini, căi către fişierele video respectiv. Astfel rămîne la discreţia Web developerilor de a indica apartenenţa conţinutului la diverse categorii. Evident devine necesară adoptarea unui standard după care vor fi servite grupurile de Web clienţi atipici. Aceştia vor avea posibilitatea de a accesa direct sau indirect fişierele cu conţinutul necesar fără a încărca şi altă informaţie adăugătoare.

Pentru a separa conţinutul textual de codul HTML s-a apelat la facilitatea extinsă a browserelor moderne de a interpreta taguri introduse de utizator, ceea ce ne dă posibilitatea de a marca locul unde va fi situat careva tip de content pe pagină. Pentru a ţine cont de ordinea de amplasare a conţinutului s-a introdus restricţia ca fiecare linie din fişierele orientate pe conţinut să corespundă unei perechi de taguri din fişierul HTML. Prin urmare fişierul html în rezultat nu va mai conţine conţinut ci doar marcajul componentelor HCML – Hyper Component Markup Language .

HTML: 212 B

```
<html>
<head>
<title>HTML</title>
</head>

<body>
<p>Text1</p><br>
<p>Text2</p><br>
<a
href=”Link1”>LinkText1</a><br>
<a
href=”Link2”>LinkText2</a><br>
<img src=”Source1”>
</body>
```

HCML: 197 B

```
<html>
<head>
<title><ci></ci></title>
</head>

<body>
<p><ci></ci></p><br>
<p><ci></ci></p><br>
<a
href=””><ci></ci></a><br>
<a
href=””><ci></ci></a><br>
<img src=””><br>
</body>
```

index.cnt: 40 B

```
HTML
Text1
Text2
LinkText1
LinkText2
```

index.ref: 12 B

```
Link1
Link2
```

index.img: 7 B

```
Source1
```

Implementarea metodei de structurare a conţinutului paginilor Web devine posibilă odată cu implicarea directă a Web develorerilor prin respectarea unor cerinţe şi standarde comune în procesul de dezvoltarea a paginilor Web. Implementarea acestei metodologii la etapa actuală ar presupune modificări introduse la nivelul de servere Web şi anume implementarea posibilităţii de clasificare conform categoriilor a conţinutului paginii Web clasice pentru clienţii atipici prin adaugarea pe server a unui script care ar implementa această funcţionalitate. În momentul în care spre server se adresează clienţi atipici, aceştea pot cere acces către fişierele reale sau virtuale cu contentul necesar prin intermediul acestui script care va oferi acces direct către fişierul real sau îl va genera în cazul fişierului virtual.

Variaţia capacităţii conţinutului interogărilor la trecerea de la tehnologia Web clasică la cea orientată pe tipologia utilizatorului este prezentată în **Tabelul 1**.

Tabelul 1. Capacitatea conţinutului interogărilor.

HTML	HCML	Content	Reference	Image
212	212	212	212	212
212	212	212	212	212
256	197	40	12	7
256	197	40	12	7

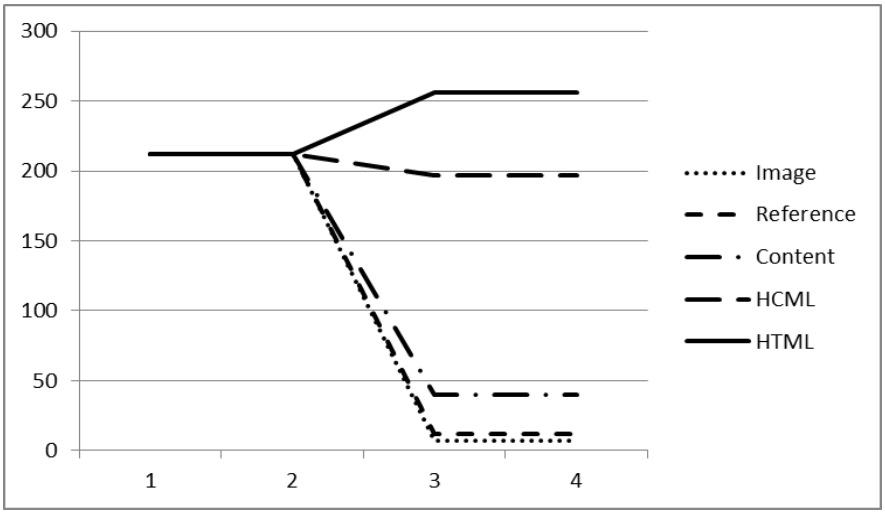

Fig. 2: Optimizarea traficului tiputilor de conţinut în internet.

Graficul din **Figura 2** marchează o descreştre semnificativă a traficului de date pentru utilizatorii atipici, indiferent de categoria de conţinut şi o creştere nesemnificativă a traficului de date pentru utilizatorii clasici. Această concluzie însă nu poate fi univocă deoarece în cazul paginilor Web orientate preponderent spre utilizatorii clasici, creşterea nesemnificativă a traficului este proporţională în timp cu numărul de utilizatori ce accesează pagina şi aplicarea acestei metodologii ar putea fi nerezonabilă din acest punct de vedere. Dacă însă prin aplicarea acestei metodologii se urmăreşte separarea conţinutului textual al paginii Web de restul conţinutului inclusiv de codul HTML sau gruparea conţinutului conform cărorva criterii specifice, atunci metoda propusă poate contribui cu succes la rezolvarea acestor probleme. În cazul unor servicii Web cum ar fi serviciile de noutăţi, care sunt orientate inclusiv în mare parte spre utilizatorii atipici reprezentaţi de alte pagini Web, implementarea acestei metodologii şi-ar găsi rostul. Astfel înainte de a implementa această metodă, mai întîi de toate ar fi recomnadabil de a analiza categoriile de utilizatori ţintă şi de a evalua raportul dintre aceştia pentru a putea estima avantajele şi dezavantajele implementării metodologiei expuse.

Implementarea metodologiei expuse în lucrare nu presupune neapărat utilizarea metodelor de implementare descrise anterior. Metodele de implementare utilizînd fişiere reale sau virtuale propuse anterior au ca scop final de a argumenta şi de a exemplifica doar posibilitatea reală de utilizare a metodologiei de structurare a conţinutului paginilor Web, şi pot fi substituite în dependenţă de specificul problemei prin metode cît mai optimal posibile.

Bibliografie

1. http://www.w3schools.com/
2. http://www.w3.org/
3. http://www.google.md/intl/ro/goodtoknow/web/101/
4. http://www.portalroman.com/articole/Internetul-140.html
5. http://invatza.info/paginiweb/introducere-paginiweb/5-ce-este-internetul.html