Web Analyzer

Uživatelská dokumentace

Jan Raszyk, 2007

Funkce programu

Program prochází webové stránky a analyzuje jimi používané technologie. Vstupem programu je množina počátečních stránek a parametry procházení (šířka, hloubka, maximální počet stránek a wildcard masky, které musí adresy procházených stránek splňovat). Výstupem programu je výpis údajů o zjištěných použitých webových technologiích a vyexportované kaskádové styly.

Úvodní obrazovka

Na úvodní obrazovce programu může uživatel nastavit parametry nového projektu analýzy (tlačítko New), nebo otevřít existující (nedokončený) projekt (tlačítko Open).

Parametry procházení

Crawl Depth určuje, do jaké hloubky je povoleno procházení odkazů. Počáteční stránky mají hloubku 0, stránky odkazované z této stránky budou mít hloubku 1, atd.

Crawl Width určuje, kolik nejvýše odkazů je možno navštívit z jedné stránky. Číslo 0 znamená neomezené množství odkazů.

Max Pages určuje maximální počet stránek, které má analýza projít.

Max Connections určuje nejvyšší počet možných síťových spojení.

Start Pages je seznam počátečních stránek, které budou jako první zařazeny do fronty k procházení. Další stránky k procházení se získají získáním odkazů vedoucích z těchto stránek.

Match Patters je seznam wildcard masek, které musí splňovat adresa URL k tomu, aby byla zařazena do procházení a analýzy. Tím lze procházení omezit například na adresy z jedné domény nebo jednoho serveru.

Po kliknutí na tlačítko Next je zobrazeno okno průběhu analýzy

Průběh stahování

Druhá obrazovka programu zobrazuje průběh procházení a analýzy stránek. Uživateli je zobrazeno, kolik stránek je zařazeno do fronty k procházení (položka Queued), kolik jich je aktuální stahováno a analyzováno (položka Processing), toto číslo většinou odpovídá uživatelskému nastavení Max Connections, a počet doposud navštívených stránek (položka Visited).

Poslední navštívená stránka ja zobrazena pod položkou Last URL.

Uživatel má možnost zastavit stahování tlačítkem Stop Download. V takovém případě dojde k dostahování a zanalyzování momentálně stahovaných stránek a k uložení souboru s projektem. Ten je také ukládán pravidelně během stahování. Po skončení stahování je možno aplikaci ukončit a při příštím spuštění načíst soubor analýzy a ve stahování pokračovat tam, kde bylo stahování zastaveno.

Výstup výsledných statistik bude proveden po kliknutí na tlačítko Export Results. Uživatel je pak požádán o zadání názvu souboru, do kterého budou v HTML formátu vypsány výsledky analýzy.

Výsledky analýzy

Jednotlivé položky výsledného dokumentu analýzy mají následující význam.

Total Pages
Celkový počet analyzovaných stránek

Pages with CSS
Počet stránek používajících kaskádové styly

Pages with JavaScript
Počet stránek používajících JavaScript

Pages with Flash
Počet stránek používajících technologii Flash

Pages with RSS
Počet stránek linkujících XML syndikační obsah

Pages with XML declaration
Počet stránek obsahujících XML deklaraci (XML hlavičku)

Pages with link to homepage
Počet stránek odkazujících na domovskou stránku stejného serveru

Non-semantic pages
Počet stránek s převažujícími formátovacími značkami (viz dále)

Semi-semantic pages
Počet stránek obsahujících jak formátovací, tak sémantické značky

Full-semantic pages
Počet stránek neobsahujících žádné formátovací značky

Doctype declarations
Seznam nalezených DOCTYPE deklarací a jejich počty

Image Types
Seznam nalezených typů obrázků (jpg, gif, png) a jejich počty

Formátovací a sémantické stránky

Moderní trend webdesignu je zbavovat stránky formátovacích elementů a nahrazovat je elementy se sémantickým významem a k formátování používat kaskádové styly. Mezi formátovací značky patří například: "b", "i", "u", "blink", "center", "font", "big", "small", "s", "basefont", ...

Mezi značky se sémantickým významem patří například: "ol", "ul", "dl", "dt", "dd", "abbr", "address", "blockquote", "q", "caption", "cite", "code", "del", "dfn", "em", "strong", "h1", "h2", "h3", "h4", "h5", "h6", "ins", "kbd", "samp", "var", …

Program sleduje množství těchto značek na stránce a podle toho stránku zařadí do jedné z kategorií Non-semantic, Semi-semantic a Full-semantic. Full-semantic stránky neobsahují žádné formátovací značky. Non-semantic stránky obsahují více než 20% formátovacích značek. Ostatní stránky jsou zařazeny do kategorie Semi-semantic.

Export kaskádových stylů

Program v průběhu stahování exportuje nalezené kaskádové styly z elementů <style> do souboru cssdump.css nacházejícího se v aktuálním pracovním adresáři.