Slovenská technická univerzita
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ
Ilkovičova 3, 842 16 Bratislava 4

Tvorba obaľovačov na získavanie informácií z webu

Tím číslo 5: *Ľubomír Chamraz , Ivan Kišac , Ján Krausko , Michal Kurťák , Marián Šimko , Michal Šimún*
Vedúci tímu: *Mgr. György Frivolt*
Študijný odbor / program: Softvérové inžinierstvo / Softvérové inžinierstvo
Ročník, typ štúdia: 1, inžinierske štúdium
Dátum odovzdania: máj 2007

1 Zadanie

Obaľovač (angl. wrapper) je program, ktorý slúži na získavanie informácií z webových stránok. Jeho použitie ušetrí manuálnu prácu sledovania informácií na webových stránkach (resp. informácií zadaných v inom formáte). Cieľom projektu je identifikácia a stiahnutie relevantných informácií z neštruktúrovaného kódu HTML, ktorá je zameraná najmä na prezentáciu informácií pre človeka a nie na spracovateľnosť pre počítače.

Tento tímový projekt nadväzuje na prácu, ktorá bola vykonaná počas minulého akademického roka, keď bol zrealizovaný jednoduchý rámec na tvorbu predprogramovaných obaľovačov. Boli riešené nasledujúce problémy týkajúce sa prostredia obaľovača, niektoré problémy však nechajú priestor na ďalšie vylepšenia. Webové stránky sú zvyčajne nekorektné. Preto sa používajú parsre a iné nástroje, ktoré pracujú aj s nekorektným HTML kódom (mozilla, beutifulsoap, htmltidy atď.). Tu je možnosť vylepšenia parsera, ktorý je robustnejší voči chybám. Ponúka sa možnosť použitia Mozilla parsera pre tento účel. Nakoľko stránky sa môžu časom meniť, buď kvôli zmene rozhrania alebo kvôli meniacim sa HTML kódom, ktoré sú zadané manuálne ľuďmi v obaľovači je realizovaný kontrolný mechanizmus. Ak obaľovač nezvláda stránku, na ktorú je prispôsobený, pri kontrole korektnosti sa spustí skript zadaný vývojárom (napr. poslanie správy). V tomto projekte sa očakáva realizácia obaľovača, ktorý stavia na výsledok dosiahnutý minulý rok a rozširuje ju o ďalšie funkcionality:

- *Návrh a realizácia učenia obaľovača* - Cieľom je aby sa obaľovač dal implementovať čím jednoduchšie, podľa možnosti potreby zadať čím menej parametrov. Používateľ by mal mať možnosť vytvoriť obaľovač len pomocou zadania príkladov, čo chce zo stránky získať.
- *Tvorba obaľovača pomocou prehliadača* - Používateľ by mal mať možnosť tvoriť obaľovač pomocou prehliadača, prostredie na tvorbu obaľovača by mal akcie používateľa odchytať a spracovať.
- *Ľahko rozšíriteľný rámec na rozšírenie obaľovača o ďalšie spôsoby učenie vzorov* - Okrem jedného zrealizovaného učenia vzorov, treba mať na zreteli rozšíriteľnosť nástroja o ďalšie spôsoby učenia.
- *Vylepšenie navigačných možností* - Je potrebné implementovať ďalšie navigačné akcie na prekonanie stránok s formulármi, heslami.

2 Úvod

Väčšina webových informačných systémov v súčasnosti disponuje veľkým množstvom dát, ktoré ponúkajú používateľom s rôznymi záujmami, vedomosťami a inými personálnymi charakteristikami. V súvislosti s nárastom informačného priestoru existujúcich systémov je používateľom ponúknuté veľké množstvo dát, kde o väčšinu z nich nemajú záujem.

Jednou z možností ako eliminovať problém získavania požadovaných dát v informačnom priestore je vykonanie personalizácie systému, ktorú využívajú adaptívne hypermédia. Adaptívne systémy vykonávajú prispôsobenie systému na základe bázy znalostí (sémantiky) o doménovej oblasti a tiež o používateľovi systému. Druhou možnosťou ako používateľovi poskytnúť relevantné informácie je vykonanie procesu automatizovaného výberu a spracovania dát, ktorého funkcionalitu zabezpečujú obaľovače. Touto možnosťou sprístupnia informácií používateľom sme sa zaoberali v našom projekte.

Obaľovač (angl. wrapper) je program, ktorý vykonáva automatizovaný výber relevantných informácií z neštruktúrovaného kódu jazyka HTML zameraného najmä na prezentáciu obsahu webu pre človeka. Použitím obaľovačov možno získavať informácie z viacerých webových systémov a integrovať ich do štruktúrovanej formy s definovanou sémantikou dát (XML dokument).

Cieľom tohto projektu je navrhnúť koncepciu tvorby obaľovača na základe analýzy existujúcich nástrojov a metód v opísanej problémovej oblasti. Tímový projekt nadväzuje na prácu, ktorá bola vykonaná počas minulého akademického roka, keď bol zrealizovaný jednoduchý rámec na tvorbu predprogramovaných obaľovačov. Obaľovač by mal byť schopný identifikovať požadované dáta v dokumente len pomocou príkladov, ktoré používateľ zadá pri špecifikovaní častí dokumentu, ktoré má obaľovač extrahovať a ktoré by naopak nemal extrahovať.

Dokumentácia k projektu je rozdelená do dvoch častí. Prvá časť sa zaoberá procesom riadenia projektu počas oboch semestrov. Obsahuje najmä plánovanie projektu, rozdelenia úloh a rôzne dokumenty súvisiace s organizáciou práce v tíme. V druhej časti sa nachádza dokumentácia samotného softvérového systému. Sú tam obsiahnuté informácie týkajúce sa analýzy a návrhu obaľovača, ako aj opis prototypu, implementácia a overenie riešenia.