Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ
Študijný odbor: INFORMATIKA

Ľubomír Chamraz, Ivan Kišac, Ján Krausko,
Michal Kurťák, Marián Šimko, Michal Šimún

TVORBA OBAĽOVAČOV NA ZÍSKAVANIE INFORMÁCIÍ Z WEBU

Tímový projekt
(ponuka)

Vedúci tímového projektu: Mgr. György Frivolt
1. október 2006

1 Úvod

Predkladaný dokument obsahuje ponuku tímu týkajúcu sa témy „Tvorba obaľovačov na získavanie informácií z webu“. Dokument je súčasťou dokumentácie tímového projektu v predmete „Tvorba softvérového systému v tíme“ a poskytuje zadávateľovi témy všetky potrebné informácie na to, aby sa mohol rozhodnúť o prideľovaní témy.

2 Ponuka

2.1 Kto sme

Všetci šiesti členovia tímu sú študentmi prvého ročníka inžinierskeho štúdia na FIIT STU v Bratislave v odbore softvérové inžinierstvo. Ide o komplexne vyprofilovanú skupinu študentov, pre ktorých nie je cudzia spolupráca v tíme, zodpovedný prístup a schopnosť dotiahnuť začaté projekty do konca. Žiaden z nich neprekročil nominálnu dĺžku štúdia, čo svedčí o ich schopnosti plnej adaptácie na akékoľvek nové technológie v oblasti informatiky a informačných technológií. Široký rozptyl ich špecializácií zabezpečuje tímu potrebnú flexibilitu a vytvára všetky predpoklady na úspešnú tvorbu softvérového systému v tíme.

Bc. Ľubomír Chamraz

Skúsenosti s programovaním získava už od roku 1997. Začínal s programovacím jazykom Turbo Pascal a neskôr sa začal venovať programovaniu v jazyku Borland C. V súčasnosti pracuje pre softvérovú firmu, kde sa zdokonaľuje v prostredí jazyka Visual C++ a v technológii MFC. Vo voľnom čase rád experimentuje s novými technológiami, akými sú napríklad HTML, JavaScript, PHP, MySQL, ale aj platforma .NET. Zo školských lavíc získal znalosti v oblastiach týkajúcich sa mobilných zariadení (platforma Java 2 Micro Edition), umelej inteligencie a sieťových technológií Cisco Systems (4 semestre Cisco Networking Academy).

Bc. Ivan Kišac

Absolvoval s vyznamenaním bakalárske štúdium na Fakulte informatiky a informačných technológií Slovenskej technickej univerzity v BA. Téma jeho záverečnej práce znela Zisťovanie charakteristík pripojenia v rámci SR na základe IP adresy, pri riešení ktorej získal skúsenosti s využívaním informácií z rôznych databáz. Má skúsenosti s programovaním vo viacerých programovacích jazykoch (Prolog, Lisp, Pascal). V poslednom období sa venuje programovaniu v Delphi s orientáciou na prácu s databázami. Primárnym programovacím jazykom je C/C++ a prostredím je Microsoft Visual Studio. Získal certifikáty z prvých dvoch semestrov Regionálnej Cisco akadémie. Počas štúdia získal pri riešení projektov skúsenosti s prácou v tíme.

Bc. Ján Krausko

Je členom skupiny PeWe (Personalized Web) group v rámci ktorej pracoval na ročníkovom projekte „Vyhľadávanie v prostredí webu so sémantikou“, pričom získal skúsenosti s tímovou prácou a zdokonalil svoje znalosti z oblasti uchovávania a interaktívnej prezentácie údajov na webe. Bližšie sa zaoberal semantickým webom a dopytovaním ontológií. Ovláda viacero programovacích jazykov (C/C++, Java) má prehľad v ontologických jazykoch (RDF, OWL), skriptovacích a značkovacích jazykoch (JavaScript, HTML, XML). Popri štúdiu softvérového inžinierstva sa venuje aj networkingu, ukončil 3 semestre Cisco Academy a v tomto roku plánuje zavŕšiť celý program CCNA (Cisco Certified Network Associate). V súčasnosti pracuje ako sieťový operátor

vo firme Soitron a.s.

Bc. Michal Kurťák

Je absolventom bakalárskeho štúdia na FIIT STU v odbore Informatika, špecializácia Softvérové inžinierstvo. Téma jeho záverečného projektu bola „Katalóg návrhových vzorov“. Za vynikajúce výsledky počas štúdia a vynikajúco vypracovanú záverečnú prácu bol ocenený cenou dekana FIIT STU. Počas štúdia pracoval jeden rok na pozícii C/C++ vývojár na platforme GNU/Linux. Pri tejto práci sa stretol aj so skriptovacími jazykmi Python a Bash. Neskôr sa začal venovať programovaciemu jazyku Java a technológiám pod platformou J2EE. Absolvoval päťmesačné školenie venované práve platforme J2EE a objektovo-orientovanej analýze a návrhu. Na týchto školeniach získal znalosti v technológiách JSP/Servlet, SQL a JDBC, Hibernate, EJB, Web Services, XML, UML, RUP a o novom komponentovo-orientovanom web-framework-u Wicket. Momentálne popri štúdiu pracuje na pozícii Java vývojár.

Bc. Marián Šimko

Má široký rozhľad v problematike IT. Ovláda viacero programovacích jazykov, z ktorých najviac používa C/C++. V priebehu riešenia záverečného projektu bakalárskeho štúdia získal rozsiahle skúsenosti s jazykmi XML a XSL (XSLT, XSL-FO) a ich aplikáciami v praxi. Poznatky získané v študijnom programe Softvérové inžinierstvo na FIIT STU rozšíril o štúdium Cisco Networking Academy, kde aktuálne študuje 4. semester a v blízkej dobe očakáva zisk certifikátu CCNA (Cisco Certified Network Associate). Dôkazom toho je aj jeho práca sieťového operátora v súkromnej firme. Tím ocení jeho nekonfliktnú osobnosť a schopnosť urovnávania problémov.

Bc. Michal Šimún

Spolupracoval s viacerými členmi zaoberajúcimi sa personalizáciou webu a webu so sémantikou (seminár PeWe). Riešením bakalárskeho projektu „Pravidlá pre prispôsobenie modelu používateľa“, za ktorého vynikajúce vypracovanie získal Cenu dekana, získal vedomosti z oblasti adaptívnych hypermédií. Počas bakalárskeho štúdia na Fakulte informatiky a informačných technológií v Bratislave, ktoré ukončil s vyznamenaním, nadobudol skúsenosti s programovacími jazykmi C/C++, Java a so skriptovacími jazykmi Perl, PHP, JavaScirpt. Má tiež skúsenosti s prácou v spoločnosti zameranej na mobilnú komunikáciu, kde nadobudol znalosti z oblasti podpory tvorby softvérových produktov (UML), značkovacích jazykov (HTML, XML), skriptovacieho jazyka PERL, vytvárania CGI skriptov a tiež databázových technológií (SQL, PostgreSQL, MySQL).

2.2 Prečo máme záujem o túto tému

Téma „Tvorba obaľovačov na získanie informácií z webu“ nás zaujala najmä svojou aktuálnosťou, keďže je potrebné z neštruktúrovaného webu dolovať informácie do podoby spracovateľnej v počítači. V súčasnej dobe existuje štandardný spôsob dolovania dát zo stránok prostredníctvom RSS kanálov, ten je však závislý od toho, či ho stránka vôbec podporuje.

Máme záujem aj o rozšírenie našich vedomostí v oblasti strojového učenia sa. Rozšírenie aktuálneho systému o túto funkcionalitu rapídne zvýši flexibilitu obaľovača. Silnou motiváciou je myšlienka sprístupnenia možnosti získavania informácií širokej verejnosti. Pokiaľ to bude možné, chceme nami vyvinuté a rozšírené nástroje implementovať aj do webového prehliadača, ktorý tvorí základné rozhranie medzi používateľom a webom.

2.3 Naša predstava o riešení

Keďže cieľom tímového projektu je nadviazať na prácu z minulého akademického roka, návrhu systému bude predchádzať dôsledná analýza už hotového systému. Na základe zistených vlastností prispôsobíme naše predstavy aktuálnemu stavu projektu a konkrétny návrh bude vychádzať z výsledkov analýzy.

Členovia nášho tímu majú skúsenosti s prácou v tíme, čo nám pomôže pri organizácii práce. Každý z nás má praktické i teoretické znalosti z viacerých oblastí, ktoré sa vzájomne dopĺňajú. Polovica členov tímu už má skúsenosti s tvorbou personalizovaného webu vďaka aktívnej účasti v skupine PeWe (Personalized Web) group. Ďalší členovia dokonale poznajú problematiku webu, publikovania na webe, použitia značkovacích jazykov v praxi, technológií súvisiacich s webovými transformáciami. Predpokladáme vývoj aplikácie v jazyku Java, kde má náš tím veľmi silné zázemie v podobe niekoľkých Java expertov, čo bude osožné aj pri implementácii aplikácie formou zásuvného modulu do prehliadača. Uvažujeme o rozšírení existujúceho obaľovača rôzne spôsoby učenia vzorov.

Návrh architektúry bude taktiež realizovaný až po podrobnom preštudovaní aktuálneho stavu projektu.

2.4 Požadované zdroje

Predpokladané nároky na softvér:

- JDK 1.5
- webový prehliadač (FireFox)
- CVS
- Eclipse

Za predpokladu, že bude zabezpečená internetová konektivita a okrem dostatočnej veľkosti operačnej pamäte (aspoň 1 GB) nemáme špeciálne nároky na hardvér.

3 Záver

Cieľom dokumentu bolo charakterizovať členov tímu č. 5 ako aj tím ako celok pre potreby zhodnotenia vhodnosti tímu pre riešenie zadanej témy. Obsahuje všetky potrebné informácie týkajúce sa skúseností jednotlivých členov tímu s danou problematikou, ich preferencií, ako aj elementárnych ideí návrhu systému.

Príloha A: Priority tém

1. Tvorba obaľovačov na získanie informácií z webu
2. Softvérová podpora životného cyklu študentského projektu
3. Znalostný manažment na báze technológie .NET
4. Tvorba textov s využitím LaTeXu

Príloha B: Tímový rozvrh

	1 07:20	2 08:15	3 09:15	4 10:10	5 11:10	6 12:05	7 13:05	8 14:00	9 15:00	10 15:55	11 16:55	12 17:50	13 18:50
Po						OOANS lch,jk,mk		ooans lch,jk,mk		TSST1 all			
Ut									PWI ik,ms		pwi ik, ms		
St	NS all		Preferované časy 1,2				PeWe ms	Preferovaný čas 3					
St		ns ik,ms,mk		ASS all			Pref. čas 4		MSI all		msi all		
Pi		ns lch,jk											

Poznámky:

lch – Ľubomír Chamraz

jk – Ján Krausko

ik – Ivan Kišac

mk – Michal Kurťák

ms – Marián Šimko + Michal Šimún

all – všetci členovia tímu

Mimo preferovaných časov má vždy aspoň jeden z členov tímu pracovné povinnosti.