Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií
Ilkovičova 2 , 842 16 Bratislava

Extrakcia dát z webu

[WebExtraction]

Dokumentácia k inžinierskemu dielu

Tím: č. 16, WebX
Vedúci tímu: Ivan Srba
Členovia tímu: Ján Brechtl, Tomáš Juhaniak, Martin Kalužník, Rastislav Krchňavý, Michal Kren, Martin Lacek, Andrej Vaculčiak
Akademický rok: 2016/2017
Autor: Ján Brechtl, Tomáš Juhaniak, Martin Kalužník, Rastislav Krchňavý, Michal Kren, Martin Lacek, Andrej Vaculčiak
Verzia číslo: 1.1
Dátum poslednej zmeny: 14.12.2016

1 Úvod

V dnešnom modernom svete plnom digitálnych technológií nepredstavuje až taký problém prístupnosť informácií, ako ich nekonzistentné uchovávanie, resp. ich reprezentácia. Poznáme rôzne druhy uchovávania informácií, či už pomocou textových alebo iných typov súborov. Keď z nich však chceme jednotlivé informácie získavať, nehovoriac o ich uchovávaní na jednom mieste, nastáva dosť ošemetná situácia, ktorá sa vyznačuje časovo náročným zberom dát z rôznych zdrojov a ich následná úprava do jednotného formátu, z ktorým sa následnej pracuje ďalej.

To isté platí aj pre webové služby. Aj keď aj v tejto oblasti existujú určité štandardy, stále to nie je len jeden, a teda je potrebné vedieť pracovať s viacerými alternatívami reprezentácie.

Tento problém otvára priestor pre realizáciu systému, ktorý bude vykonávať zber dát z rôznych webových serverov uchovávajúcich informácie v rôznej podobe. Jednou zo základných a veľmi užitočných vlastností by mala byť možnosť opakovaného zberu dát, nakoľko je dobré udržiavať dáta neustále aktuálne a kontrolovať ich manuálne je dosť nepraktické.

Obsahom dokumentu je okrem celkového pohľadu na systém aj detailnejší popis jednotlivých jeho modulov. Ku každému modulu sú uvedené informácie o analýze, návrhu a následnej implementácií, pričom sa nesmie vynechať testovanie, ktoré je ťažiskom pri tvorbe jednotlivých modulov.

2 Globálne ciele pre ZS

Keďže ide o tímový projekt, na ktorom spolupracujú ľudia, ktorí predtým spolu nepracovali, celý zimný semester, no najmä prvá polovica sa vyznačuje inicializáciou činností, dohadovaním a konfiguráciou mnohých komponentov a podporných nástrojov, no v neposlednom rade rozdelením si zodpovedností a začatím implementácie prvých verzií vybraných modulov systému.

Cieľom projektu je vytvoriť funkčný systém na pravidelnú extrakciu vybraných dát z vybraných webových stránok.

Používateľ sa má možnosť zaregistrovať a následne prihlásiť na server, ktorý mu poskytuje možnosť vytvorenia projektov, pričom v jednom projekte môže mať uložených viacero skriptov, každý pre rôznu stránku, ktoré definujú, aké dáta sa budú extrahovať.

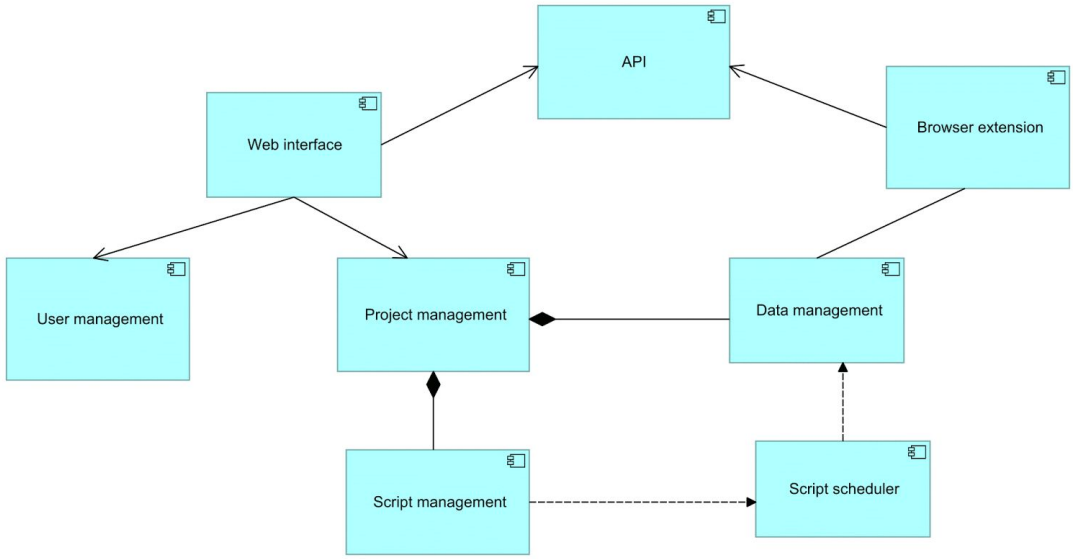
Okrem definovania extrahovaných dát systém umožní nastaviť dátum spustenia automatickej extrakcie a pravidelnosť jej spúšťania v rôznych intervaloch (minúty, hodiny, atď.).

Samozrejmosťou je prehľadný výpis používateľových projektov, skriptov, aj samotných extrakcií spolu s časom ich spustenia a vykonania.

Aby však používateľ mohol tieto extrakcie definovať, potrebuje na vybranej stránke nejakým spôsobom vybrať, ktoré dáta chce extrahovať. Túto funkcionalitu mu poskytne rozšírenie do prehliadača, pomocou ktorého sa prihlási na svoj účet a môže si dané skripty podrobnejšie definovať.

3 Celkový pohľad na systém

Podľa informácií od zákazníka, resp. vlastníka produktu je hlavnou funkciou systému automatizované zbieranie preddefinovaných dát z vybraných webových stránok. Používateľ si pomocou grafického rozhrania (webová stránka) môže spravovať vlastné projekty, skripty a dátové schémy, podľa ktorých sa vykonáva extrakcia. Anotácia dát na požadovanej stránke je realizované pomocou rozšírenia do webového prehliadača.



Obr. 1 - Zobrazenie modulov systému a ich vzťahy

Systém ako taký je tvorený dvoma základnými časťami. Prvou je webová aplikácia poskytujúca prístup k účtu používateľa, s možnosťou správy projektov a skriptov na zber dát. Tak isto je možné spravovať svoj vlastný účet (napr. zmeniť heslo).

Druhá časť je už spomínané rozšírenie do prehliadača, pomocou ktorého používateľ jednoducho definuje, aké dáta a z ktorej stránky sa majú extrahovať.

Dekompozícia celého systému na časti je znázornená na obr. 1.

4 Moduly systému

Pre lepšie pochopenie systému a určitú systematickú implementáciu boli definované moduly, ktoré logicky vychádzajú z diagramu v predchádzajúcej kapitole. Moduly prechádzajú postupnou implementáciou a každý z nich je podrobnejšie opísaný v samostatnom dokumente, ktorý sa venuje analýze, návrhu, implementácií a testovaniu konkrétneho modulu.

Systém teda pozostáva z týchto modulov:

1. User management
2. Project management
3. Browser extension