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

Extrakcia dát z webu

[WebExtraction]

Modul Project management

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: 1.2
Dátum poslednej zmeny: 15.05.2017

1 Úvod

Vo väčšine prípadov extrahovania dát je požadovaným výsledkom set dát, ktoré spolu určitým spôsobom súvisia. Preto hovoríme o určitej doméne, v ktorej dáta extrahujeme. Doménu v našom systéme reprezentujú projekty, ktoré zahŕňajú dátové polia spolu so skriptami.

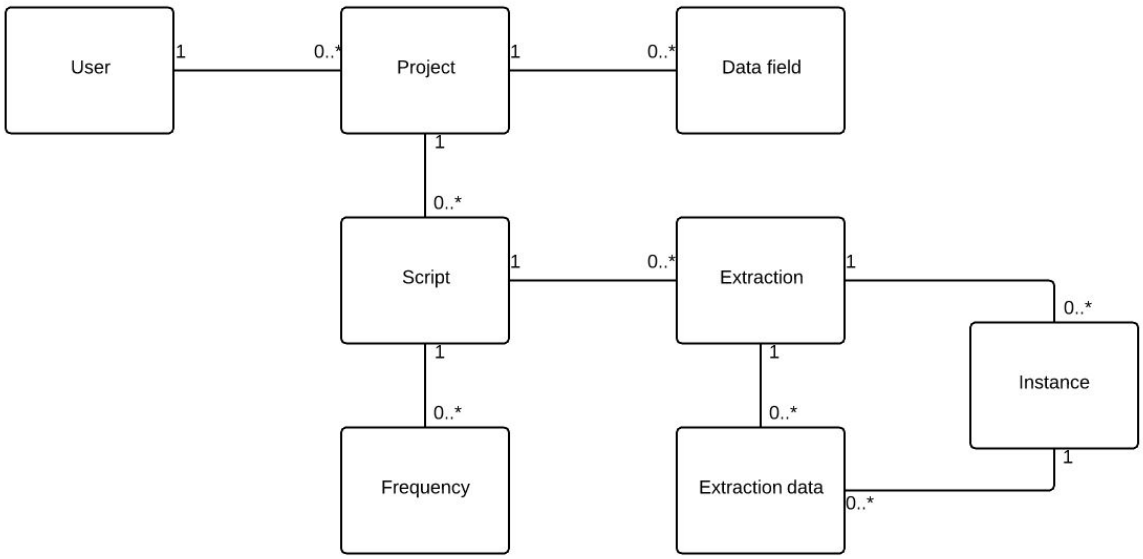
2 Project management

2.1 Analýza

Princípom celej webovej extrakcie je sťahovanie dát z webu. Keďže chceme aby mali požívatelia pri určovaní dát na sťahovanie poriadok, bolo nutné vytvoriť miesto kde sa bude definovať každé extrahovanie z webu.

2.2 Návrh

Z výsledku analýzy sme dospeli k vytvoreniu 3 modelov a to projekt, dátová schéma a skript. Platí, že používateľ má mnoho projektov a jeden projekt patrí jednému používateľovi, projekt má mnoho skriptov a jeden skript patrí jednému projektu, projekt má mnoho dátových schém a jedna dátová schéma patrí jednému projektu. Celá situácia je zobrazená na obrázku nižšie. V tejto etape riešenie nebude limitovaný počet projektov skriptov a atribútov, ktoré si môže používateľ vytvoriť.



Obr. 1 - Diagram navrhovaných závislostí medzi entitami

Najskôr si používateľ musí vytvoriť projekt, pre každý jeho prípad použitia. V prípade ak by chcel používateľ sťahovať rôzne typy dát, pre každý typ bude musieť vytvoriť samostatný projekt. K projektu teda logicky prislúcha dátová schéma. V dátovej schéme si

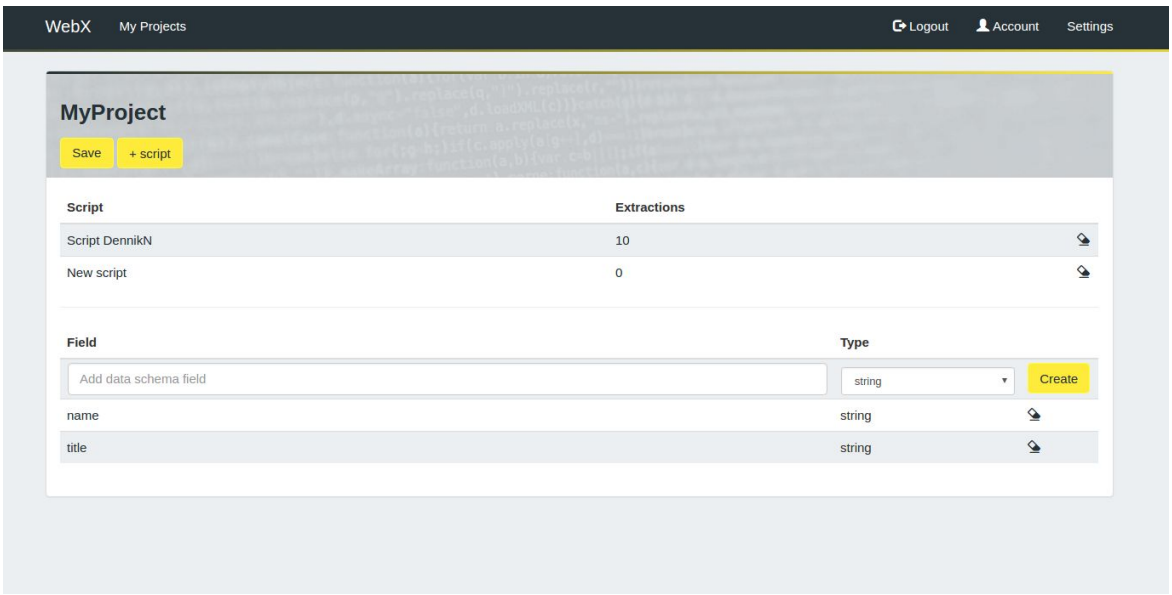
používateľ definuje atribúty a ich typy. Po definovaní dátovej schémy si používateľ môže vytvoriť skripty pre všetky weby z ktorých chce extrahovať.

Používateľ môže mať napríklad 2 projekty pre extrahovanie dát z eshopu s notebookmi a pre extrahovanie dát z inzercie na bazári. V projekte extrahovania dát z elektronického bazáru má používateľ definované atribúty cena, popis lokalita a telefónne číslo. Následne si vytvorí skripty pre každú bazárovú stránku o ktorú má záujem.

2.3 Implementácia

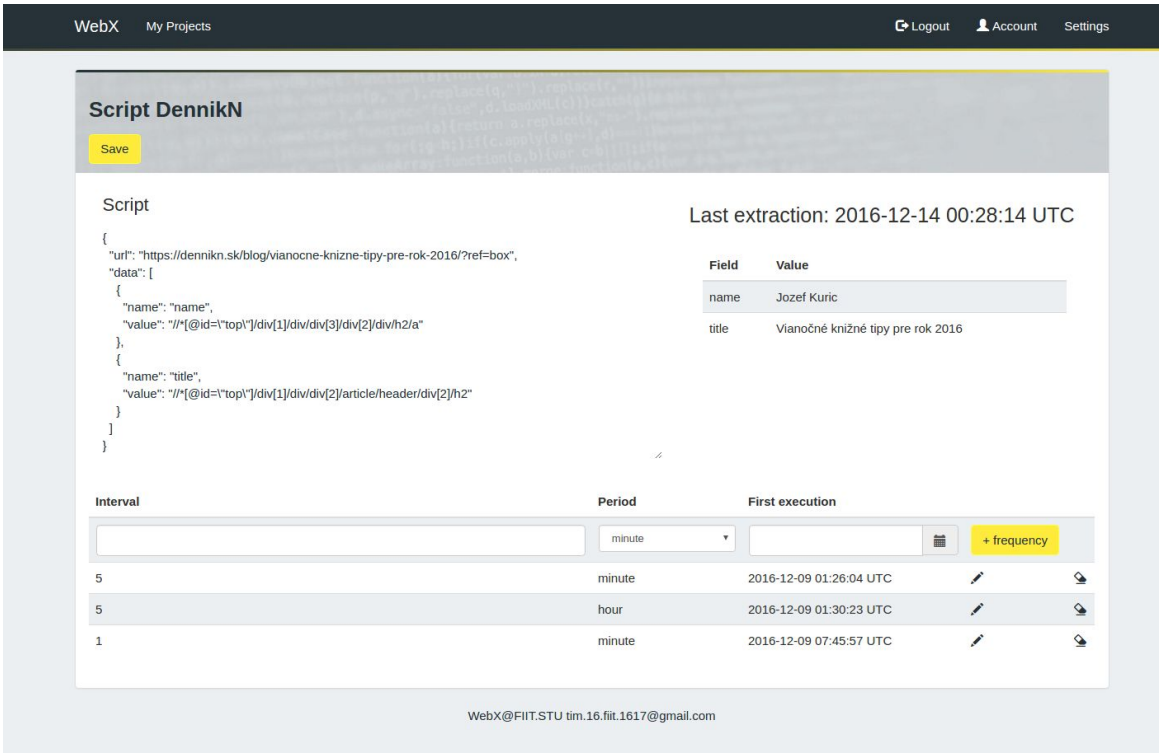
Pre projekty, skripty a dátové schémy sme najskôr vytvorili 3 modely spolu s prislúchajúcimi migráciami. V tejto fáze riešenia obsahovali projekty a skripty iba jeden atribút a to ich meno. Dátová schéma obsahovala okrem názvu aj typ, ktorý je implementovaný ako enum (vymenovaný typ) s hodnotami integer (celé číslo), float (desatinné číslo) a string (reťazec znakov). Zabezpečenie projektov pred neoprávneným čítaním a úpravou má na starosti gem ‘cancancan’.

View pre projekty sme robili v niekoľkých iteráciach. Zo začiatku sme mali pre prezeranie projektov a ich úpravu samostatné formuláre, po tímových konzultáciach sme sa zhodli na jednotnom view. Úprava atribútov projektu sa vykonáva priamo pri jeho prezeraní, t.j. “show” view. Túto funkcionalitu sme docielil využitím technológie AJAX a pokročilého CSS. Na tejto obrazovke (viď obrázok nižšie) je zobrazený zoznam skriptov a im prislúchajúcich exktrakcií. Pod ním sa nachádza zoznam dátových polí pre daný projekt, spolu so vstupným poľom pre vytvorenie nového.



Obr. 2 - View pre projekty

Z obrazovky projektu sa dá prekliknúť na jednotlivé skripty. Úprava jednotlivých atribútov funguje rovnako ako v projektoch. Rozdielom je hlavne text_area pre samotný skript v JSON formáte. Ako je vidieť na nasledujúcom obrázku:



Obr. 3 - Úprava skriptu

Používateľ má možnosť upraviť si skript priamo na tejto obrazovke alebo v rozšírení prehliadača. Vpravo od skriptu sa nachádza výstup poslednej extrakcie. Podobne ako dátové polia pre projekt, v skripte sme v tabuľke zobrazili frekvencie spúšťania skritpu spolu s poľom pre vytvorenie nových frekvencií.

2.4 Testovanie

Pre účely testovania sme najskôr vytvorili factories (továrne - hodnoty v databáze určené iba pre testovanie). Následne sme napísal niekoľko testov, ktoré skontrolujú či sa správne zobrazujú všetky potrebné informácie o projektoch, skriptoch a dátových schémach.