Strigil - systém pro získávání strukturovaných dat z webu

Návrh softwarového projektu

Vedoucí:
- Jakub Stárka (starka@ksi.mff.cuni.cz)
- Martin Nečaský (necasky@ksi.mff.cuni.cz)

Počet řešitelů
- 5 - 6

Motivace

V současné době existují v České republice desítky informačních systémů určených pro zadávání a zveřejňování veřejných zakázek. Tyto systémy se výrazně liší ve formátech výstupů i jejich strukturování. Některé používají jako výstup HTML, jiné provádějí export do různých tabulkových formátů, či např. pdf. Vzhledem k množství těchto systémů a různorodosti formátů dochází k problémům s agregací dat a jejich dalším strojovém zpracování. Z tohoto důvodu je také obtížné tato data jednoduchým způsobem automaticky získávat a navzájem propojovat do širších souvislostí.
Existuje velké množství nástrojů, které umožňují získávání dat z webu, např. [1,2]. Jedná se jednak o univerzální crawlery, sloužící k mapování webu a jeho indexování, a tudíž jsou neefektivní v získávání konkrétních dat. Dále jsou k dispozici systémy pro plánování spouštění vlastních stahovacích skriptů, např. [3]. Kromě toho existuje několik komerčních aplikací [4,5], které umožňují uživatelům specifikovat konkrétní části HTML stránky a dále zpracovat tato extrahovaná data. Problém těchto řešení je jednak v jejich uzavřenosti, která znemožňuje jejich další rozšiřování na další formáty (např. pdf, doc, …) a především neumožňuje provázání s vyšší úrovní abstrakce (např. jasně definovanou ontologií).

Cíl projektu

Cílem projektu je navrhnout a implementovat SW systém, který umožní svým uživatelům pomocí definovaných pravidel získávat strukturovaná data z nestrukturovaných zdrojů (webové stránky, office dokumenty, …). Konkrétně umožní uživateli pomocí přehledného rozhraní určit části dokumentu, které odpovídají definované ontologii a případně další vlastnosti procesu stahování (periodické opakování, omezení na počet přístupů, atd.). Tato získaná data nadále exportuje do formátu rdf nebo posílá dalším aplikacím pomocí předem definovaného rozhraní.

Hlavní cíle projektu tedy jsou:
- Návrh a implementace nástroje umožňujícího automatické stahování stránek dle předem definovaných šablon
- Získávání konkrétních dat ze stažených dokumentů
- Schopnost konfigurace, přidávání nových pravidel a ontologií za běhu

- Automatické upozorňování na podezřelé vstupy
- Export získaných dat do souboru, nebo do nějakého uložiště pomocí předem definovaného rozhraní
- Autorizace uživatelů
- Z pohledu uživatele bude systém umožňovat:
    - definici nových pravidel pro procházení webu a získávání dat
    - sledování neočekávaných výsledků extrakce (upozornění na změnu struktury dokumentu)
    - sledování statistik a stavu systému
- Systém bude umožňovat práci alespoň s následujícími dvěma typy dokumentů
    - HTML
    - xls

Nefunkční požadavky

Systém bude implementován v jazyce Java a bude provozován pod systémem MS Windows. Z důvodů snadné obsluhy nezávislé na platformě bude hlavní rozhraní systému implementování jako webová stránka s využitím technologie AJAX pro zvýšení komfortu obsluhy. Generování jednotlivých šablon bude probíhat v modulech přizpůsobených specifikům konkrétního formátu.
Pro větší univerzálnost bude klientská část komunikovat s jádrem pomocí jasně definovaného rozhraní, které umožní snadné rozšíření množiny cílových formátů.
Jako ukázku navrženého systému bude provedeno scrapování některých systému pro ukládání veřejných zakázek (např.: ezak[1], isvzus[2]).

Další požadavky na projektový tým

- K projektu bude vytvořena hned od počátku webová stránka, kde bude postupně vznikat informační, uživatelská, programátorská a instalační dokumentace k vytvářenému software
- Veškerá vytvářená dokumentace bude v anglickém jazyce

Předpokládaný průběh práce

V projektu bude implementován prototyp specifikovaného SW systému. Projekt je dimenzován na práci 6 členného týmu v průběhu 9 měsíců. Práce budou probíhat dle následujícího schématu:

1. Analýza existujících implementací a přístupů [M1-M2]
2. Podrobná specifikace konkrétních funkcí systému, architektury a rozhraní mezi jednotlivými moduly [M2-M3]
3. Implementace prototypu [M3-M7]
4. Testy na reálných datech, ladění [M7-M9]
5. Dokumentace (programátorská, uživatelská, instalační) [M7-M9]

Předpokládané rozdělení práce

[1] http://www.ezak.cz/
[2] http://www.isvzus.cz/

- 2 lidé - implementace jádra systému (plánování úkolů, správa stahovacích plánů, správa ontologií, export strukturovaných dat, …)
- 1 - 2 lidé - vytvoření intuitivního uživatelského rozhraní pro specifikaci hledaných dat pro jednotlivé podporované formáty
- 2 - 3 lidé - návrh a implementace metod pro získávání dat z HTML, XLS(X), ...

Reference

1. A. Dasgupta, A. Ghosh, R. Kumar, C. Olston, S. Pandey, and A. Tomkins. The discoverability of the Web. In Proc. 16th Int'l Conf. on World Wide Web, pages 421--430, 2007.
2. C. Olston and M. Najork. Web crawling. Found. Trends Inf. Retr., 4(3):175--246, 2010.
3. ScraperWiki
4. http://www.screen-scraper.com/download/choose_version.php
5. http://www.visualwebripper.com/
6. Ricardo A. Baeza-Yates , Berthier Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1999
   ftp://mail.im.tku.edu.tw/seke/slide/baeza-yates/chap10_user_interfaces_and_visualization-modern_ir.pdf