Sem vložte zadání Vaší práce.

České vysoké učení technické v Praze
Fakulta informačních technologií
Katedra ... (softwarového inženýrství)

Bakalářská práce

InfoWeb - Nástroj získávání informací z webů

Vedoucí práce: Ing. Jiří Hunka

1. dubna 2017

Poděkování

Chtěl bych poděkovat za trpělivost vedoucímu, Ing. Jiřímu Hunkovi.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval(a) samostatně a že jsem uvedl(a) veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů. V souladu s ust. § 46 odst. 6 tohoto zákona tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, jež jsou její součástí či přílohou, a veškeré jejich dokumentace (dále souhrnně jen „Dílo“), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli způsobem, který nesnižuje hodnotu Díla, a za jakýmkoli účelem (včetně užití k výdělečným účelům). Toto oprávnění je časově, teritoriálně i množstevně neomezené. Každá osoba, která využije výše uvedenou licenci, se však zavazuje udělit ke každému dílu, které vznikne (byť jen zčásti) na základě Díla, úpravou Díla, spojením Díla s jiným dílem, zařazením Díla do díla souborného či zpracováním Díla (včetně překladu), licenci alespoň ve výše uvedeném rozsahu a zároveň zpřístupnit zdrojový kód takového díla alespoň srovnatelným způsobem a ve srovnatelném rozsahu, jako je zpřístupněn zdrojový kód Díla.

V Praze dne 1. dubna 2017 .....

Odkaz na tuto práci

Tuček, Jakub. *InfoWeb - Nástroj získávání informací z webů* . Bakalářská práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2017.

Abstrakt

V několika větách shrňte obsah a přínos této práce v češtině. Po přečtení abstraktu by se čtenář měl mít čtenář dost informací pro rozhodnutí, zda chce Vaši práci číst.

Klíčová slova Nahraďte seznamem klíčových slov v češtině oddělených čárkou.

Abstract

Sem doplňte ekvivalent abstraktu Vaší práce v angličtině.

Keywords Nahraďte seznamem klíčových slov v angličtině oddělených čárkou.

Obsah

Seznam obrázků

Úvod

V předmětech BI-SP1 a BI-SP2 v prostředí FIT ČVUT byl realizován týmový projekt pro získávání informací z webů s primárním zaměřením na potřeby obchodů. Projekt řešil problém automatizace získávání dat z webů, jelikož stávající služby neposkytují veřejné rozhraní nebo mají velkou chybovost dat.

Požadavky internetových obchodů jsou především tvořeny nutností držet krok s trhem a tedy sledovat vývoj cen prodávaných produktů u konkurence. Teprve na základě těchto dat je možné reagovat na trh a měnit vlastní cenu produktů.

Cílem této práce je popsat požadavky internetových obchodů, stávající stav a možná řešení. Dále na základě těchto poznatků zhodnotit vytvořené řešení a včetně korektnosti zvolených postupů navrhnout vylepšení. Ty implementovat, řádně otestovat vylepšení a zhodnotit výsledný stav projektu.

Kapitola 1

Cíl práce

1.1 Analytické cíle

1. Rešerše aktuálního stavu získávání dat pro potřeby internetových obchodů
2. Analýza vzniklého řešení týmového projektu vzniklého v prostředí ČVUT FIT, včetně důrazu na použité postupy při softwarovém vývoji
3. Návrh a zhodnocení implementovaných vylepšení

1.2 Praktické cíle

1. Implementace vylepšení systému

Kapitola 2

Analýza a návrh

V této kapitole se budu nejprve zabývat samotnou problematikou získávání informací z webů s důrazem na internetové obchody. Jelikož je tato problematika již řešena existujícími službami, je nutné je zhodnotit a popsat jejích chování. Dále zhodnotím současný stav projektu, zvolené postupy při vývoji a výslednou funkcionalitu. Nakonec navrhnu vylepšení vzniklého systému, ty nejdůležitější implementuji a výsledný stav zhodnotím.

2.1 Získávání informací z webů

2.1.1 Problematika

Získávání informací z webů je efektivní možnost jak získat databázi informací, které se na internetu vyskytují. Tato činnost však stojí na problematice data získávat a uchovávat v potřebné struktuře, jelikož jinak z dat nejsme schopni vyčíst potřebné informace. Vzhledem k specificitě dat, které jsou v kontextu činnosti zajímavá a dále kvůli unikátnosti webových stránek není možné jednoznačně určit jak data získat v požadovaném formátu.

2.1.2 Výběr dat

Nejčastější řešení a jediné řešení je kombinace lidské inteligence a automatizování co nejvíce činností. To je obvykle dosaženo roboty, která stahují data a lidské práce určující jaké informace nás zajímají.

Získávání dat ze stažených stránek lze poté zjednodušit na problematiku lokace elementů v HTML, které jsou pro nás zajímavé. Lokaci elementu v HTML se kterým je potřeba pracovat lze poté jednoznačně určit pomocí dvou možnosti:

1. XPath
2. CSS Selector

2.1.3 XML Path Language

XML Path Language[1] nazývaný zkráceně XPath je jazyk, který slouží k výběru elementu v dokumentu ve formátu XML.

XML chápeme jako jazyk popisující strukturu dat, které jsou strojově i lidsky čitelné. HTML je speciální typ XML, který popisuje obsah dat pro prezentaci ve webovém prohlížeči pomocí předem definované struktury, které prohlížeče rozumí.[2] Díky této vlastnosti lze tedy použít XPath pro definování cesty k prvku (a jeho obsahu), který uchovává potřebnou informaci na webové stránce.

2.1.4 CSS Selector

Jazyk CSS je používán pro vizuální popis prezentace webové stránky definované v HTML. Jazyk k určení prvků se kterými chce pracovat používá selektory, které označují prvek v HTML. Jako selektor může být použit jak samotný název prvku, tak vlastní definované třídy.[3]

Pomocí řetězení těchto selektorů jsme schopni jednoznačně získat element v HTML.

2.2 Současný stav řešení potřeb internetových obchodů

I v kontextu malého trhu jako Česká republika se lze bavit o velké konkurenci na poli maloobchodů prodávající své zboží na internetu. Internetové obchody potřebují monitorovat konkurenci a trh. Vzhledem k jejich zaměření je tedy nejvíce zajímají obchody prodávající stejné zboží. Potřebné informace o prodávaných produktů konkurencí se skládají z následujících hlavních atributů:

1. Cena
2. Inzerovaný název
3. Dostupnost

2.2.1 Srovnávače cen

Data lze získat pomocí srovnávačů cen jako *zbozi.cz*[4] nebo *heureka.cz*[5]. Problém u těchto služeb je však že jsou spíše určeny koncovým zákazníkům pro nalezení nejlepší ceny na trhu pro určitý produkt. S tím souvisí to, že největší srovnávače cen neposkytují data nebo rozhraní přes která by je bylo možné získat.

2.2.2 Existující služby

Problematiku sledování trhu s důrazem na firemní klientelu, řeší aktuálně několik existujících služeb.

Služby mají v zásadě velmi podobnou povahu služeb. Rámcově se jedná o porovnávání cen včetně historie na různých internetových obchodem či na srovnávačích cen. Uživatel si zadá okruh či seznam produktů, buďto formou manuální či vstupem ze souboru, případně přímých napojením na e-shop. Následně je možné konkrétní data zobrazit v grafech označující vývoj cen, trendů či náhlých změn. Dále umožňují externí výstup do souboru v dostupných formátech.

Největší rozdíl služeb je zda jsou data získávána přímo z obchodů nebo ze srovnávačích. Další odlišností je možnost zda služba dokáže sledovat i zahraniční trh.

Cena služeb se nejvíce odvíjí od počtu sledovaných produktů a četnosti aktualizací. Proto se měsíční platby mohou pohybovat od stovek korun po desítek tisíc korun.

2.2.2.1 Price checking

2.2.2.2 Pricing intelligence

2.2.2.3 Sledování trhu

2.2.2.4 Pricebot

2.2.2.5 Zahraniční nástroje

Tyto nástroje jsou obecněji zaměřené a obvykle požadují od uživatele techničtější zaměření, jelikož je nutné přesně specifikovat kde, co a jak chce sledovat. Vzhledem k tomuto omezení nejsou přímo pro provozovatele e-shopů vhodné kvůli nedostatečným technickým kapacitám a pro tuto práci důležité.

Bodový seznam zahraničních nástrojů:

1. Screen scraper [6]
    - Webová služba
    - procházení web skrz odkazy
    - potvrzování formulářů
    - využití interního vyhledávání
    - export do širokého množství formátu souborů
    - cena: $549 - $2,799 za měsíc
2. Web extractor [7]
    - Windows Aplikace

- procházení zadaných stránek
- hledání stránek pomocí klíčových slov
- export do csv formátu
- cena: $99 - $199 jednorázově

3. Web Scraper [8]

2.3 Zhodnocení současného stavu projektu

TODO

2.4 Návrh na vylepšení

TODO

2.5 Analýza nového řešení

TODO

Kapitola 3

Realizace

3.1 Způsob realizace stávajícího řešení

TODO

3.2 Implementace vylepšení

TODO

Závěr

Literatura

[1] XPath. In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2017-04-01]. Dostupné z: https://en.wikipedia.org/wiki/XPath

[2] PERUGINI, Saverio. HTML versus XML. In: Virginia Tech - College of engineering: Department of computer science [online]. Virginia Tech: Virginia Tech, 2002 [cit. 2017-04-01]. Dostupné z: http://courses.cs.vt.edu/~cs1204/XML/htmlVxml.html

[3] Cascading Style Sheets. In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2017-04-01]. Dostupné z: https://en.wikipedia.org/wiki/Cascading_Style_Sheets

[4] Heuréka [online]. [cit. 2017-04-01]. Dostupné z: http://www.heureka.cz/

[5] Zboží [online]. [cit. 2017-04-01]. Dostupné z: http://www.zbozi.cz/

[6] Screen scraper [online]. [cit. 2017-04-01]. Dostupné z: http://www.screen-scraper.com/

[7] Web extractor [online]. [cit. 2017-04-01]. Dostupné z: http://www.webextractor.com/

[8] Web Scraper [online]. [cit. 2017-04-01]. Dostupné z: http://http://webscraper.io/

Příloha **A**

Seznam použitých zkratek

XML Extensible markup language

HTML Hypertext Markup Language

CSS Cascading style sheets

Příloha **B**

Obsah přiloženého CD

- `readme.txt` ..... stručný popis obsahu CD
- `exe` ..... adresář se spustitelnou formou implementace
- `src`
  - `impl` ..... zdrojové kódy implementace
  - `thesis` ..... zdrojová forma práce ve formátu LaTeX
- `text` ..... text práce
  - `thesis.pdf` ..... text práce ve formátu PDF
  - `thesis.ps` ..... text práce ve formátu PS