UNIVERSITY OF TARTU
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
Institute of Computer Science
Computer Science Curriculum

Andres Viikmaa

Building product database for shopping search engine

Master's Thesis (30 ECTS)

Supervisor: Timo Petmanson, MSc

Tartu 2015

Building product databae for shopping searchengine

Abstract:

Abstract

Keywords:
List of keywords

Tootekataloogi loomine toodete otsingumootori jaoks

Lühikokkuvõte:
Abstract

Võtmesõnad:
List of keywords

Unsolved issues

Contents

1 Introduction

Creating product information database includes gathering product data from hundreds of thousand manufactures. Contacting each of them individually is time consuming and manual process requiring large amount of human labour.

2 Scraping the web

Kirjeldus, et mis on web scaping

2.1 Extracting data from web

HTML lehtede parsimine

css selektorid

regulaaravaldised

[KFT14]

2.2 Semantic web

https://courses.cs.ut.ee/2007/internet/Main/Web3

Üldine kirjeldus semantic veebist

Ajalugu

Tänapäev

Mis see pole piisav

2.3 Web crawlers

Käivad ainult mööda linke, ei vaata väga sisu

Lühike sissejuhatus ajaloost tänapäevani ja mõned näited

http://nutch.apache.org/

2.4 Web scrapers

Tegelevad siis ka andmete eraldamisega lisaks linkidel käimisele

Erinevad tasemed alates lihtsast skriptist lõpetades valmisprogrammidega

Andmete eraldamine tuleb ise teha

2.5 Web scraping as a service

Pole arvutusvõimsust -> kasuta teenust

Lihtne liides click-and-go

2.6 Automatic web scraping and data extraction

Senimaani, iga leht on eraldiseisev objekt

Enamasti paljud lehed sarnase struktuuriga, populaarsed veepipoe platformid, blogid, foorumid

Sarnast sisu esineb mitmel veebilehel - saab õppida sisu baasil lehe struktuuri

3 Conclusion

what did you do?

What are the results?

future work?

References

[KFT14] Kei Kanaoka, Yotaro Fujii, and Motomichi Toyama. Ducky: a data extraction system for various structured web documents, 2014.

Non-exclusive licence to reproduce thesis and make thesis public

I, Alice Cooper (date of birth: 4th of February 2048),

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1 reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2 make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

Type Inference for a Fourth Order Logic Formulae

supervised by Axel Rose and May Flower

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu/Tallinn/Narva/Pärnu/Viljandi, dd.mm.yyyy