RDFa Crawler
Webcrawler inklusive RDFa parsing funktion

STEFAN ACHMÜLLER, ROLAND GRITZER, MATHIAS GSCHWANDTNER
Universität Innsbruck
February 20, 2017

Zusammenfassung

Enabling to scrap the world wide web for RDFa data. Give a URL to start from, furthermore black- and whitelisting is possible. DRFa data, integrated within html documents, is generated by applying rdfa.info sequence. RDFa data is annotated by the RDF n-triple format.

Schlüsselwörter: webcrawler, parser, rdfa

Inhaltsangabe

1 Problem Definition

Mit der historischen Entwicklung des Internets wurden immer wieder Techniken eingeführt, die den Datenaustausch zwichen Rechnern vereinfachen. Während im klassischen Internet, auch Web 1.0 (URL, HTTP, HTML, etc.) genannt, der Fokus auf dem Aufbau und Transport der Daten liegt, betrachtet der Ansatz "Semantic Web" mögliche Interpretationen der Daten. Dies wird auch Web 3.0 genannt und ermÃűglicht eine vereinfachte Abarbeitung von Aufgaben, basierend auf Internetdaten. So kann zum Beispiel unterschieden werden, ob das Wort "Bremen" auf einer bestimmten Webseite sich entweder auf eine deutsche Stadt, einen Familiennamen oder einem sonstigen Namen bezieht. Die Kerneigenschaft des "Semantic Web" stellt die Universalität der Relationten dar. Dies bedeutet, dass prinzipiell alle Informationsobjekte miteinander verknüpft werden kÃűnnen um Wissen zu repräsentieren [1].

Um Daten mit diesen Metainformationen anzureichern, wurden diverse Annotationstypen eingeführt. Neben JSON-LD und Microdata, bietet sich hierfÃijr das "Resource Description Frameworks" (RDF) an. Ziel dieser Arbeit ist die Erstellung einer Programmierbibliothek zum automatischen extrahieren von semantischen Informationen, annotiert mittels RDFa (W3C Standard Annotation für RDF), welche in HTML Webseiten eingebettet sind [2].

1.1 RDFa Information

RDFa lite und die sequenzen [3].

1.2 Webcrawler

Webcrawler - finden was man sucht [4].

2 Methodologie

Do research: javascript, node, schema.org, etc. Split work into parser and crawler. Set up front end (testing and presentation). Start implement. Put together and write documentation.

2.1 NodeJS und npm

dfsflkj

2.2 Modul 1: RDFa Parser

dsflkj

2.3 Modul 2: Webcrawler

dklfja

2.4 Zusammenführung

dsfjd

3 Ergebnis

1 package to download via node npm. 2 separate functionalities within package: crawler (without parser), and parser (without crawler).

3.1 Beispielanwendung

dasflkj

Referenzen

1. Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.
2. Wolfgang Halb, Yves Raimond, and Michael Hausenblas. Building linked data for both humans and machines. In *LDOW*, 2008.
3. Ben Adida, Mark Birbeck, Shane McCarron, and Steven Pemberton. Rdfa in xhtml: Syntax and processing. *Recommendation, W3C*, 7, 2008.
4. Brian Pinkerton. *Webcrawler: Finding what people want*. Citeseer, 2000.