Extrakce a zpracování dat z ne-RDF zdrojů

Doc. Ing. Vojtěch Svátek, Dr.

4IZ440, Zimní semestr 2014

http://nb.vse.cz/~svatek/rzzw.html

Motivace

- Naprostá většina dat se do formátu RDF převádí z jiných (nativních) formátů

Odkud lze extrahovat data

- Tabulková data
  - RDBMS, CSV, HTML, XLS…
- Hierarchická data
  - XML…
- Síťová ne-RDF data
  - většinou ale uložená v tabulkách
- Volný text

Extrakce z relačních DB

- Mapování z RDB do RDF se definuje
  - pomocí jazyka R2RML
    - W3C doporučení ze září 2012, http://www.w3.org/TR/r2rml/
    - Umožňuje definovat mapovací pravidla, optimalizující strukturu dat pro využití jako LD
    - Podporuje různé způsoby vystavení RDB dat jako LD: endpoint, export dumpu, negociace obsahu
  - formou tzv. Direct Mapping
    - Struktura RDF doslova kopírující strukturu RDB
    - http://www.w3.org/TR/2012/PR-rdb-direct-mapping-20120814/
- Prototypové implementace: Oracle, PostgreSQL, MySQL…
    - viz http://www.w3.org/TR/rdb2rdf-implementations/

Extrakce z XML a CSV

- Pro XML jde zpravidla o XSLT šablony, někdy s nadstavbami
  - Např. Valiant, https://github.com/bertvannuffelen/valiant
- Pro CSV – řada nástrojů
  - Např. TARQL, https://github.com/cygri/tarql

Extrakce z volného textu

- Nejnáročnější a nejméně spolehlivá, využívá lingvistických technik
- Spíše identifikace entit a jednotlivých trojic než vyčerpávající převod obsahu dat
- Klíčová iniciativa: http://nlp2rdf.org
  - Interoperabilita lingvistických nástrojů pomocí mezijazyka NIF (NLP Interchange Format) založeného na RDF

Ruční extrakce z tabulek

- Tabulky (HTML, XLS, CSV apod.) menšího rozsahu lze převádět ručně
- Nejpopulárnější nástroj: OpenRefine
  - http://openrefine.org/
  - Původně GoogleRefine
- S pluginy pro LD je označován jako LODRefine
  - http://code.zemanta.com/sparkica/
  - Vyvinuto v projektu LOD2, ve spolupráci slovinské firmy Zemanta, DERI, a dalších partnerů

OpenRefine

- Import dat ze souboru nebo přes clipboard
- Čištění a transformace dat v tabulkovém zobrazení
- Propojení dat na webové služby
- Linkování na entity (z Freebase)
- Využívání skriptovacího jazyka GREL (podobného JavaScriptu)
  - Využívání regexp… analogicky k FILTER v SPARQL

LODRefine navíc nabízí

- Mapování na DBpedii, ev. další endpointy
  - Rekonciliace – přiřazení hodnot k URI
  - Augmentace – „dotažení“ dalších dat o entitě
- Extrakci entit z volného textu v polích tabulky
- Napojení na CrowdFlower – evaluace rekonciliace pomocí crowdsourcingu
- Export do RDF
  - Interaktivní návrh kostry RDF grafu

VIZ CVIČENÍ…

též tutoriál na https://www.youtube.com/watch?v=4Ve93C238gI