4IZ440 Propojená data na webu

Organizační a „kontextový“ úvod

Vyučující: Doc. Ing. Vojtěch Svátek, Dr.

Zimní semestr 2017

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

Náplň předmětu

- Technologie **propojených dat** (linked data, LD) na (tzv. sémantickém) webu
- Cílem je naučit se data na sémantickém webu
  - **vytvářet a vystavovat**, včetně
    - návrhu datových schémat a doménových slovníků
    - extrakce dat, která nativně v RDF nejsou
  - **zpracovávat** (čistit, transformovat a propojovat)
  - **využívat** (v jednoduchých aplikacích)
- Důraz je na aktivní přístup a interakci
  - Když vám něco není jasné nebo připadá divné, ptejte se!

Probíraná témata

- **RDF**: jazyk pro reprezentaci „propojitelných“ dat na webu v obecné grafové struktuře
- **SPARQL**: dotazovací jazyk pro RDF
- **Datové slovníky** (ontologie): popis RDF dat existujícími slovníky, i tvorba nových slovníků
- Proces **zpracování dat** („ETL“) vedoucí k jejich vystavení v RDF: extrakce, transformace (včetně čištění a propojování) a výsledné publikování
- Využívání RDF dat v **softwarových aplikacích**
- **Byznys modely** podporující využití RDF dat na webu

Co předmět je a není?

- Nejde o desítky let ustálenou disciplínu, pro kterou by existovaly „nepohnutelné“ osnovy a učebnice (i když dobré knihy a příručky už vyšly!)
- Obsahem výuky není jedna přesně vymezená technologie (jazyk, metodika, software…)

ALE

- Jde o relativně volné propojení jazykových standardů, softwarových nástrojů, výzkumných projektů a iniciativ…
- „Správné odpovědi“ na mnohé klíčové otázky zatím nikdo na 100% nezná, obor se rok od roku rychle vyvíjí
- Vzhledem k tomu je u studentů velmi vítána vlastní iniciativa, zvědavost, zkoumání toho, co se nově objevuje jak na akademických konferencích, tak i např. v blogpostech

Zdroje pro studium

- Základní literatura:
  - kniha: ***Linked Data: Evolving the Web into a Global Data Space*** od T. Heatha a C. Bizera, online **http://linkeddatabook.com**
  - Slidy k přednáškám, budou postupně vystavovány na webu
  - Tutoriál L. Feigenbaum, E. Prud'hommeaux: ***SPARQL by Example***, online http://www.cambridgesemantics.com/semantic-university/sparql-by-example
  - Kapitoly z knihy **Umělá inteligence 6**, Academia (knihovna/studovna/zakoupit?)
  - učební text ***Aktuální problémy a perspektivy sémantického webu*** přístupný z webu předmětu
- Doplňková literatura:
  - Viz web; zvl. B. DuCharme: ***Learning SPARQL*** (knihovna/studovna)
  - Specifikace jazyků a formátů (RDF, Turtle, SPARQL), dostupné na webu
  - Tutoriály k nástroji LP-ETL přístupné z https://etl.linkedpipes.com/
  - Další tutoriály ze Semantic University na http://www.cambridgesemantics.com
  - V případě zájmu mnoho další literatury u vyučujícího, články na WWW a v digitálních knihovnách

Kontrolní testy

- Teoretické znalosti budou ověřeny třemi písemnými testy (předběžně 13+14+13 = 40 bodů)
  - psaní „kódu“, grafy, odpovědi volným textem, zaškrtávačky
- Bodové minimum z každého testu je 5; v jednom testu lze minimum nesplnit bez dalších sankcí
- V případě nedosažení minima ve dvou testech lze o absolvování předmětu uvažovat jen pokud budou splněny všechny ostatní povinnosti a součet bodů dosáhne 60
  - následuje ústní přezkoušení, a pokud je úspěšné, student získá známku „vyhověl“ bez ohledu na celkový počet bodů
- *Testy budou zřejmě v 5., 8. a 13. týdnu semestru, a to na přednáškách*

Praktické úkoly

- Zpracovat samostatně semestrální projekt (max. 40 bodů)
  - Průběžná prezentace na posledním cvičení 11.12. – ústní, plus odevzdání jednostránkového konceptu
  - Finální odevzdání ve zkouškovém období (termín bude upřesněn)
- Zpracovat samostatně ústní a písemný minireferát (max. 15 bodů)
  - Ústní prezentace 27.11. nebo 4.12.
  - Z odborného článku z hlavní vědecké konference v oboru - ISWC 2017 (vysoce doporučeno, po dohodě lze i z jiné tematicky vhodné konference, případně časopisu), přístup k článkům bude zajištěn
- Průběžné úkoly na cvičeních (celkem až 15 bodů), předběžně
  - Převody RDF mezi různými způsoby vyjádření
  - Tvorba dotazů SPARQL na různých úrovních náročnosti (dotazování, kontrola a transformace dat, propojování…) a pro různé typy dat
  - Návrh jednoduchého datového slovníku
  - **Týmová** případová studie zahrnující extrakci, čištění, propojení a využití (v primitivní aplikaci) **fiskálních** dat

Dimenze „uvažování o tématu“

- Životní cyklus LD
  - zdrojová data – RDF schémata a slovníky – ETL – dotazování pomocí SPARQL – využití v aplikaci
- Věcné domény dat
  - encyklopedická; statisticko-ekonomická (jiné: biomedicína, knihovnictví, produkty/služby…)
- Myšlenková prostředí
  - Věda: výzkumné projekty, experimenty a publikace
  - Inženýrství: reprezentační jazyky a SW nástroje
  - Byznys: modely výnosů a nákladů

A k čemu Vám to dále může být?

- Pro studenty bezprostředně směřující do praxe
  - Nový pohled na techniky, které se v praxi běžně využívají (integrace datových zdrojů, grafové DB, správa metadat, podnikové taxonomie, webová API, textová a webová analytika, architektury orientované na služby, konceptuální modelování, …)
  - Připravenost na novinky, které do praxe zřejmě proniknou v příštích 5-10 letech
- Pro studenty se zájmem o prozkoumávání neprozkoumaného
  - Přehled významné části výzkumných témat, která jsou na KIZI řešena a mohou být předmětem doktorských disertací, ale i diplomových prací

Tým vyučujících

- **Doc. Ing. Vojtěch Svátek, Dr.:** většina přednášek i cvičení
- **Ing. Jakub Klímek, Ph.D.** (FIT ČVUT, MFF UK): technologie ETL pro RDF – *přednáška i cvičení*
- **Ing. Jan Kučera, Ph.D.** (KIT VŠE): případová studie ETL pro RDF (ČSSZ) - *přednáška*
- **Ing. Marek Dudáš:** tvorba slovníků pro RDF; využití RDF dat v aplikaci – *oboje cvičení*
- **Ing. Václav Zeman:** ETL pro encyklopedická data – *cvičení*
- **Mgr. Jindřich Mynarz:** konzultační podpora ohledně jazyka SPARQL a technologií propojených dat obecně

Související předměty

- 5FI430 – Znalosti a ontologické inženýrství
  - M. Vacura, KFIL, povinný pro KI
  - Problematika ontologických modelů jako bohatší varianty datových slovníků – formální odvozování nad koncepty, filozofické ukotvení atd.
- 4IZ470 – Dolování znalostí z webu
  - V. Svátek, povinný pro ZWT (zpravidla již absolvovali…)
  - Součástí „web miningu“ je extrakce informací z webových textů – komplementární k převodu již strukturovaných dat do RDF
- 4IZ530 – Logické programování
  - Š. Sem, oborově volitelný pro ZWT a KI
  - Programování v jazyce Prolog; má řadu společných rysů s psaním dotazů ve SPARQL, lze v něm psát sémantické aplikace

Pro hlubší zájemce

- Pracovní skupina KIZI **SWOE**
  - „Semantic web and ontological engineering“, viz http://kizi.vse.cz/swoe
  - účastní se i studenti

…weeding the semantic web garden

- Neformální výzkumný seminář **KEG**

  „Knowledge Engineering Group“
  - některé (1./2.) čtvrtky v semestru od 16 hodin zpravidla na 473NB (zasedačka FIS)
  - možno přijít bez předchozího přihlášení
  - viz http://keg.vse.cz/seminars
  - kdo chcete dostávat oznámení, napište!

- Aktivity mezi-institucionální iniciativy **OpenData.cz**: http://opendata.cz
  - vč. společného týmu KIT a KIZI, http://opendata.vse.cz/