4IZ440 Reprezentace a zpracování znalostí na WWW

Seznámení s předmětem

Doc. Ing. Vojtěch Svátek, Dr.

Zimní semestr 2011

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

Hlavní otázky, které by měl předmět zodpovědět

- Jak je možné využít strukturu WWW ke sdílení a propojování strukturovaných dat a znalostí mezi lidmi a organizacemi?
- Jak taková strukturovaná data/znalosti vytvářet?
- Jak je možné v takto propojených datech a znalostech vyhledávat, případně nad nimi automaticky odvozovat pomocí logických postupů, a prezentovat výsledek uživateli?
- Jak je možné získávat strukturované znalosti z nestrukturovaných textů, zejména na webu?
- Co z výše uvedeného se v současnosti nebo v blízké budoucnosti uplatní v praxi, a jak?

Motivační scénáře

- Digitální knihovna, ve které lze pro daný dotaz obdržet nejen dokumenty obsahující zadaná klíčová slova, ale i jiné dokumenty řešící stejnou nebo podobnou problematiku
- Repozitář obrázků nebo videí, ve kterém lze vyhledávat podle různých obsahových hledisek
- Portál integrující informace o osobách, organizacích, místech apod., získané z různých, nepředddefinovaných zdrojů
- Webový portál, který automaticky mění svou strukturu na základě dat, které do něj poskytovatelé dodávají
- Agregátor blogů využívající souvislosti mezi jednotlivými vyslovenými tvrzeními
- Aplikace, která za běhu vyhledává adekvátní webové služby a kombinuje jejich výstupy za účelem vyhovění požadavku uživalele (např. zabezpečení zahraniční cesty se vším všudy – doprava, hotel, pojištění…)

- Některé ze scénářů využití znalostních technologií na webu lze již dnes využívat v rámci veřejně přístupných služeb, některé se testují neveřejně v rámci projektů pro komerční sféru, a některé jsou zatím jen na úrovni omezených, čistě výzkumných prototypů

Co předmět je a není?

- Nejde o ustálenou disciplínu, pro kterou by existovaly „tradiční“ osnovy a „závazné“ učebnice
- Obsahem výuky není jedna přesně vymezená technologie (jazyk, metodika, software…)

ALE

- Jde o velmi volné propojení výzkumných iniciativ, softwarových nástrojů, jazykových standardů…
- „Správné odpovědi“ na klíčové otázky v oboru zatím nikdo na 100% nezná
- Vzhledem k tomu je u studentů velmi vítána vlastní iniciativa, zvědavost, zkoumání toho, co se nově objevuje na webu, na konferencích apod.!

Klíčové pojmy

- Sémantický web
- Propojená data (Linked Data)
  - Linked Open Data, Linked Government Data, Linked Open Commerce, Linked Life Data, …
- RDF (Resource Description Framework)
  - Úložiště, SPARQL, koncové body
  - RDFa, mikroformáty, mikrodata
- Topic Maps (mapy témat)
- Ontologické inženýrství
  - OWL
  - Deskripční logika
  - Ontologické návrhové vzory
- Sémantické anotování
  - Extrakce informací

Klíčové pojmy (zjednodušeně)

- Sémantický web: zhruba vymezuje náplň předmětu – jde o využití webové infrastruktury pro zpřístupnění „strojům srozumitelných“ informací (sémantika = význam, smysl)
  - Termín zaveden kolem r.2000 pro oblast vzniklou spojením
    - nástrojů a standardů sítě WWW
    - technologie reprezentace a zpracování znalostí, zejména modelování znalostí (ontologického inženýrství) a formální logiky (deskripční, event. Hornova logiky)
  - Později se zapojily i další komunity: zpracování přirozeného jazyka, text/web mining, databáze, filosofie, zpracování neurčitosti, ...

Klíčové pojmy (zjednodušeně)

- Linked Data: soubor principů pro propojování dat v RDF, pocházejících z různých zdrojů, a množina takto již přizpůsobených zdrojů
- Mikroformáty, mikrodata: prostředek pro jednoduché zachycení sémantiky v HTML
- RDF (Resource Description Framework): hlavní jazyk pro reprezentaci dat v sémantickém webu, standard W3C; spíše „databázově“ orientovaný
  - Úložiště RDF (RDF store): „databáze“ pro data ve struktuře RDF
  - SPARQL: dotazovací jazyk analogický SQL
  - Koncové body (SPARQL endpoints): webová rozhraní pro dotazování do (centralizovaných nebo distribuovaných) úložišť
  - RDFa: formát pro vkládání dat RDF do stránek v HTML, „vylepšení “ mikroformátů
- Topic Maps (TM): alternativní jazyk pro reprezentaci sémantických dat, standard ISO; spíše „dokumentově“ orientovaný, omezené možnosti odvozování

Klíčové pojmy (zjednodušeně)

- Ontologie: formální konceptuální model určité problémové oblasti; pojmy a vztahy z modelu slouží k dodání sémantiky datům jak pro RDF, tak i pro TM
  - Ontologické inženýrství:
  - OWL: jazyk pro reprezentaci ontologií, zpravidla používaný ve spojení s RDF
  - Deskripční logika: logický kalkul umožňující odvozování nad výroky v RDF/OWL
  - Ontologické návrhové vzory: opakovaně použitelné fragmenty ontologií, zpravidla opatřené vysvětlujícím komentářem, příklady apod.
- Sémantické anotování: přiřazování významu (tj. pojmů z ontologie) částem textu – může být ruční nebo automatické
  - Extrakce informací: automatické sémantické anotování textu, směřující ke zpřístupnění nalezených informací ve strukturované podobě

Co se od Vás očekává?

- Získat rámcový přehled o oblasti
  - Základní literatura:
    - slidy k přednáškám, budou postupně vystavovány na webu (od 2. týdne)
    - kniha: ***Linked Data: Evolving the Web into a Global Data Space*** od T. Heatha a C. Bizera, online **http://linkeddatabook.com**
  - Doplňková literatura:
    - *A Semantic Web Primer* od G. Antoniou a F. van Harmelena
    - *Semantic Web for the Working Ontologist* od D. Allemanga a J. Hendlera
    - případně *XML Topic Maps* od J. Parka a S. Huntinga
    - dostupné ve studovně a ve skladu
  - Definice používaných jazyků jsou dostupné na WWW
  - V případě zájmu mnoho další tištěné literatury u vyučujícího, články na WWW a v digitálních knihovnách
- Bude ověřeno dvěma písemnými testy (40% hodnocení), kombinujícími požadavky na konkrétní znalosti se schopností samostatně uvažovat a formulovat teze
  - Při nedosažení minimálního počtu bodů (5 resp. 10) z některého z testů lze o absolvování předmětu uvažovat jen pokud budou splněné ostatní povinnosti, a součet bodů dosáhne 60 – pak následuje ústní přezkoušení, a pokud je úspěšné, student získá známku „vyhověl“ bez ohledu na celkový počet bodů
- *Testy jsou předběžně v 7. a 12. týdnu semestru*

Ale také a především…

- Zpracovat samostatně nebo (v případě rozsáhlejšího projektu) ve dvojici semestrální projekt (36% hodnocení)
- Některé možné varianty projektu (detaily budou upřesněny):
  - nápaditý sémantický mash-up („mesh-up“)
  - rozsáhlé testy hotové volně šířeného sémantického nástroje
  - znalostní model (ontologie, mapa témat) s dedikovaným rozhraním
  - doménová ontologie bohatě a konzistentně opatřená axiomy
  - sada formálních gramatik pro extrakci informací z textu
  - podrobná specifikace obecného nástroje (typicky ve vazbě na plánovanou DP)
  - přehledová komparativní studie určitého tématu zpracovaná na základě nejnovější literatury *(nelze za ni ale získat plný počet bodů…)*
- Kromě posledních dvou variant se požaduje i textový dokument rozebírající postup práce, provedená rozhodnutí při realizaci atp.
- *Termíny odevzdání budou upřesněny*

A dále

- Podílet se na kolaborativní tvorbě RDF slovníku z určeného oboru, a na tento slovník navázat bázi dat RDF převedenou z nativního XML (12% hodnocení)
- *Oborem tvorby slovníku jsou v tomto semestru „zařízení na zpracování obnovitelných zdrojů energie“*
- *Práce proběhnou v 7. a 8. týdnu*

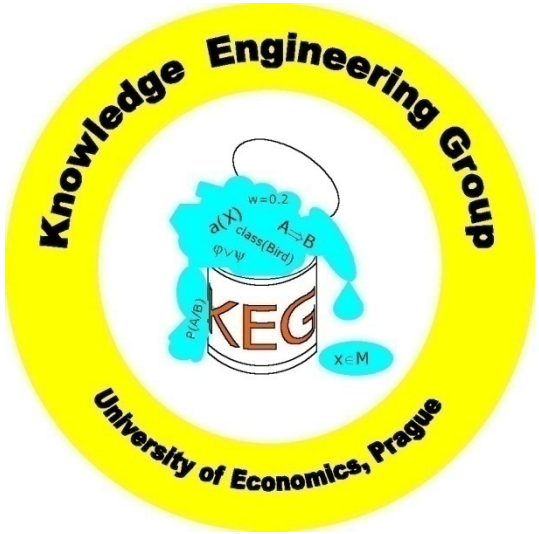
- Zpracovat samostatně ústní a písemný minireferát z odborného článku z hlavní konference v oboru (12% hodnocení)
- *Určenou konferencí je ISWC 2011, viz http://iswc2011.semanticweb.org/program/research-papers/*
- *V průběhu semestru (prezentace 11. týden)*

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, textová a webová analytika, architektury orientované na služby, metamodelování IS, groupware, …)
  - Připravenost na novinky, které do praxe 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

Pro hlubší zájemce

- Neformální výzkumný seminář KIZI
  - resp. pracovní skupiny KEG (Knowledge Engineering Group)
- Téměř každý čtvrtek v semestru
  - od 10.30 do cca 12.00 na 413NB (zasedačka KIZI)
  - možno přijít bez předchozího přihlášení
- Viz http://keg.vse.cz/seminars



Pro hlubší zájemce

- Možnost zapojení do vědeckých projektů KIZI s touto problematikou
  - Vývojářské (Java, PHP aj.), ev. výzkumné činnosti vč. spoluúčasti na publikacích a cest na konference
  - Honorováno přes DPP/DPČ, ev. mimořádná stipendia
  - Možnost zpracovat DP na aktuální témata řešená světovým výzkumem
- Projekty FP7 EU **LOD2** a LinkedTV
- Dále:
  - Projekty GAČR PatOMat a Sémantizace webu
  - Projekty IGA VŠE Sewebar a Sémantické propojování dat ve veřejné správě

Tématické bloky

- Ukázky koncových aplikací
- Reprezentační jazyk RDF
- Dotazovací jazyk SPARQL
- Principy Linked Data
- Sémantika v HTML
  - Mikroformáty, mikrodata, RDFa, GRDDL
- Vizualizace, tvorba meshupů
- Vytváření RDF
  - Ručně, z RDB, XLS, XML atd.
- Odvozování nad RDF Schema
- Tvorba slovníků/ontologií
  - Kolaborace při tvorbě slovníků
- Oblasti Linked Data
  - Linked Data ve veřejné správě
  - Linked Data v eCommerce
- Propojování /deduplikace dat
  - mapování schémat/ontologií
- Deskripční logika
  - Formalismus, vztah k jiným logikám, tablové odvozování
- Revize a induktivní vytváření ontologií pro Linked Data
- Ontologické inženýrství
  - Ontologie v oblasti medicíny
  - Metodiky, návrhové vzory
- Topic Maps
- Sémantické anotování textů a extrakce informací
- Rozšiřující témata (stihne-li se)
  - Sém.webové služby
  - Sémantika multimédií
  - Pravidlové přístupy

Tématické bloky

Výukové týdny

1. Ukázky koncových aplikací
2. Reprezentační jazyk RDF
   - Dotazovací jazyk SPARQL
3. Principy Linked Data
   - Sémantika v HTML
4. – Mikroformáty, mikrodata, RDFa, GRDDL
   - Vizualizace, tvorba meshupů
5. Vytváření RDF
   – Ručně, z RDB, XLS, XML atd.
6. Odvozování nad RDF Schema
   - Tvorba slovníků/ontologií
7. – Kolaborace při tvorbě slovníků
   - Oblasti Linked Data
   – Linked Data ve veřejné správě
   – Linked Data v eCommerce
8. Propojování /deduplikace dat
   – mapování schémat/ontologií
9. Deskripční logika
   – Formalismus, vztah k jiným logikám, tablové odvozování
   - Revize a induktivní vytváření ontologií pro Linked Data
10. Ontologické inženýrství
    – Ontologie v oblasti medicíny
    – Metodiky, návrhové vzory
    - Topic Maps
12. Sémantické anotování textů a extrakce informací
    - Rozšiřující témata (stihne-li se)
    – Sém.webové služby
    – Sémantika multimédií
    – Pravidlové přístupy

Tématické bloky

Producent LD

- Ukázky koncových aplikací
- Reprezentační jazyk RDF
- Dotazovací jazyk SPARQL
- Principy Linked Data
- Sémantika v HTML
  - Mikroformáty, mikrodata, RDFa, GRDDL
- Vizualizace, tvorba meshupů
- Vytváření RDF
  - Ručně, z RDB, XLS, XML atd.
- Odvozování nad RDF Schema
- Tvorba slovníků/ontologií
  - Kolaborace při tvorbě slovníků
- Oblasti Linked Data
  - Linked Data ve veřejné správě
  - Linked Data v eCommerce

Konzument LD

- Propojování /deduplikace dat
  - mapování schémat/ontologií
- Deskripční logika
  - Formalismus, vztah k jiným logikám, tablové odvozování
- Revize a induktivní vytváření ontologií pro Linked Data
- Ontologické inženýrství
  - Ontologie v oblasti medicíny
  - Metodiky, návrhové vzory
- Topic Maps
- Sémantické anotování textů a extrakce informací
- Rozšiřující témata (stihne-li se)
  - Sém.webové služby
  - Sémantika multimédií
  - Pravidlové přístupy