Sémantický web a extrakce informací

Martin Kavalec

kavalec@vse.cz

Katedra informačního a znalostního inženýrství FIS VŠE

Přehled témat

- Vize sémantického webu
- Extrakce informací pro sémantický web
- Ontologie jako spojovací článek

Motivace pro sémantický web

- Stávající web je určen pouze pro vnímání lidmi, strojové zpracování je možné jen ve velmi omezené míře
- Úkoly, které vyžadují integraci informací z různých zdrojů je nutné provádět ručně:

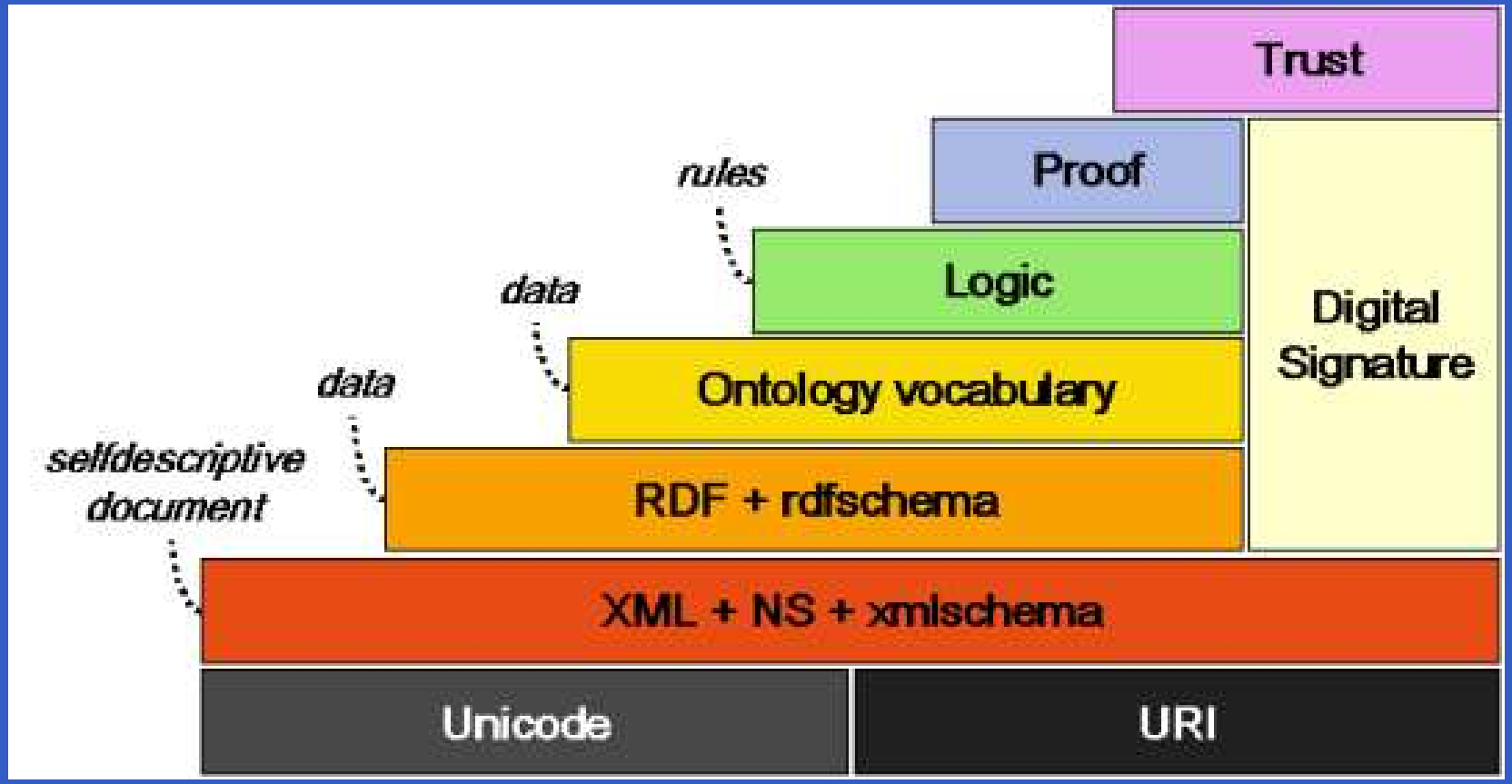
  Př.: Které obchody v Praze prodávají kolo *Apache Tomahawk*, za kolik, jaký je jejich telefon a adresa?

Co je sémantický web

Sémantický web je rozšíření stávajícího webu, ve kterém každá informace má dobře definovaný význam a tím umožňuje lepší spolupráci lidí a počítačů při jejich zpracování.

– Tim Berners-Lee, James Hendler, Ora Lassila, The Semantic Web, Scientific American, May 2001

Vrstvy sémantického webu



RDF: datový model sémantického web

- Fakta ve tvaru [subjekt predikát objekt]
  [Rembrandt namaloval Aristoteles-s-bystou-Homérovou]
- subjekt a predikát jsou URI, objekt může být URI nebo literál
- URI jako nositel významu – díky tomu, že autor programu, uživatel programu i autoři publikovaných tvrzení chápou identifikiátory subjektů, predikátů i objektů stejně, může daný program pro uživatele udělat něco užitečného

Extrakce informací

Extrakcí informací se snažíme z volného nebo semi-strukturovaného textu automaticky identifikovat informace s určitým významem

Př. v souboru oznámení o pořádaných seminářích chceme identifikovat název vystoupení, jméno přednášejícího, místo, datum a čas konání.

- 2 podúlohy: Rozpoznání entit, extrakce relací mezi nimi
- Několik málo tříd entit (okolo 5)

Metody extrakce informací

- Wrappery
- Statistické modely (HMM)
- Extrakce pomocí pravidel
- Hybridní – LP^2

Extrakce informací z webu

Kromě textu jsou k dispozici další informace:

- struktura stránky a její formátování (vyznačení nadpisů, zvýraznění textu, uspořádání informací v seznamech a tabulkách)
- topologie webu, tj. která stránka se na kterou odkazuje a jaká slova k tomu používá
- metadata explicitně uvedená ve www stránce
- informace z analýzy struktury URL
- informace o obrázcích, jejich rozměrech a vlastnostech

Extrakce informací pro sémantický we

IE pro sémantický web se liší od situací, ve kterých bývá typicky využita:

- velký počet menších ontologií, které se průběžně vyvíjejí
- nutnost snadné adaptability metod na aplikační oblasti a na změny v ontologiích
- potřeba rozpoznávat větší počet tříd (řádově desítky)

Požadavky na metody extrakce

- možnost adaptace na omezeném vzorku trénovacích dat
- schopnost identifikace relací bez nutnosti hluboké syntaktické analýzy; systém by měl mít možnost využít lingvistické informace, pokud jsou k dispozici a jsou spolehlivé, v ostatních případech by měl využívat jednodušších metod
- možnost využít ontologické zdroje, pokud jsou k dispozici

Ontologie: zdroj i cíl extrakce

- Ontologie: říká, co chceme extrahovat
- významná součást ontologií: lexikální položky
- lexikon může posloužit jako zdroj pro automatické značkování trénovacích dat
- z takto značkovaných dat je možné získat extrakční vzory pro koncepty nebo relace
- tyto extrakční vzory mohou nacházet nové instance a lexikální položky pro ontologii
- možnost bootstrappingu – snažší adaptace na aplikační oblast

Učení ontologií

- učení konceptů
- učení taxonomických relací
- učení netaxonomických relací

Využití asociačních pravidel

Jak nalézt ze sady textů relace mezi koncepty:

1. Vyhledáme výskyty konceptů v textu (lexikální položky konceptů + jejich instancí)
2. Pokud se dva koncepty vyskytují blízko sebe, započteme je jako „transakci“
3. Na tyto transakce aplikujeme dolování asociačních pravidel
4. Získaná asociační pravidla představují dvojice konceptů, mezi nimiž lze hledat nějakou relaci

Implemetováno v modulu Text-To-Onto v KAONu

Jak spolu koncepty souvisí?

Problém: nemáme žádné vysvětlení, jak tyto koncepty spolu souvisí

Tato informace je v textech obsažena, avšak metoda zpracování ji nevyužívá

Cíl: Identifikovat tuto informaci v textu a přiřadit ji k získaným asociačním pravidlům

Relace mezi koncepty jsou často vyjádřeny slovesy, navíc tedy ke konceptům vyhledáváme slovesné fráze

Slovesné fráze

- identifikovány na základě POS-tagů
- POS-tagging je (ve srovnání s parsingem) relativně rychlý a robustní

Vyhledáváme fráze

- $V(C_1, C_2)$: vyskytující se podle vzoru C_1 near $verb$ near C_2
- $V(C)$: vyskytující se blízko konceptu C

K asociačnímu pravidlu s koncepty (C_1, C_2) zobrazíme $V(C_1, C_2)$ a $V(C_1) \cap V(C_2)$

Zajímavé slovesné fráze

Kvantifikátor pro ohodnocení VCC transakcí:

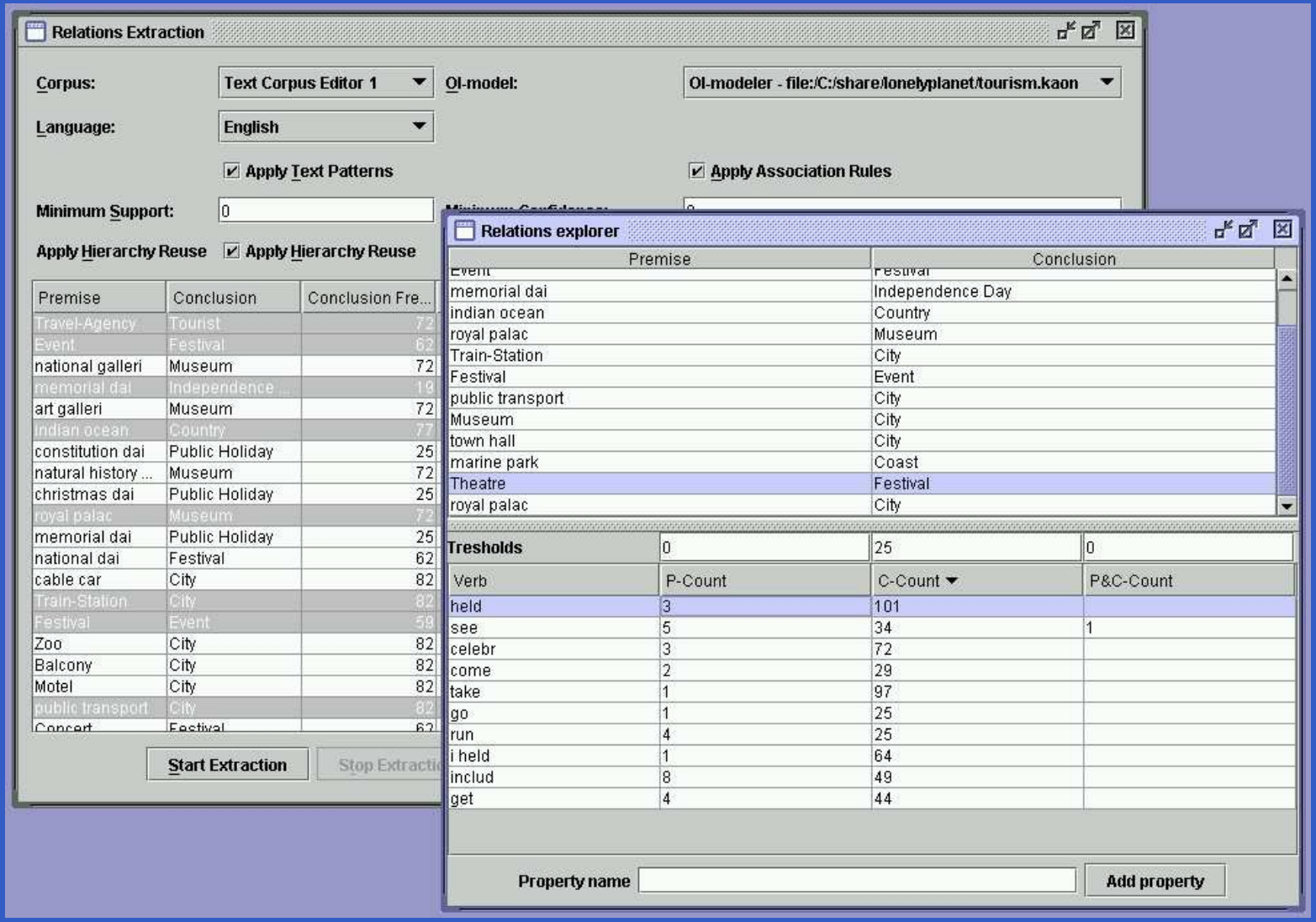
$$AE(c_1 \wedge c_2/v) = \frac{P(c_1 \wedge c_2/v)}{P(c_1/v).P(c_2/v)}$$

kde pravděpodobnosti jsou spočteny takto:

$$P(c_1 \wedge c_2/v) = \frac{|\{t_i|v, c_1, c_2 \in t_i\}|}{|\{t_i|v \in t_i\}|}$$

(ostatní pravděpodobnosti analogicky, t_i označuje jednotlivé VCC transakce)

Relation Explorer



Pracovní data

- texty: popisy zemí z www.lonelyplanet.com
- ontologie: část TAPu + rozšíření o termíny pro oblast turismu

Problémy působí bohatost a obraznost jazyka v použitých textech – tatáž informace je vyjádřena co nejvíce různými způsoby

Experiment s korpusem SemCor

- část korpusu z Brownovy univerzity
- koncepty i slovesa mapována na jejich synsety ve WordNetu
- použili jsme jen část SemCoru: novinové a odborné texty
- zvolili jsme synsety Person, Organization, Location (+jejich synsety, které na ně bylo možné zobecnit)
- subjektivně „čistější“ výsledky, zjevné rozdíly mezi novinovými a odbornými texty

Experiment s OpenDirectory

1. v sadě stránek odkazovaných z Business sekce Open Directory jsme našli výskyty termínů, pod nimiž byly stránky v katalogu zařazeny
2. nalezli jsme slovesa, která se k těmto termínům syntakticky váží častěji, než k jiným termínům – *indikátory důležité informace*, tj. informací o nabízených produktech

Extrakce informací pomocí indikátorů

3. pomocí několika nejlepších indikátorů jsme pak v testovací sadě stránek vyhledali věty, které by měly obsahovat informaci o produktech
4. v průměru 80 % vět tuto informaci opravdu obsahovalo

Podobným způsobem by bylo možné využít pro extrakci informací i lexikální položky relací v ontologii.

THE END

Děkuji za pozornost