Extrakce z nestrukturovaných dat

Ing. Ivo Lašek

(upravil doc. Ing. Vojtěch Svátek, Dr.)

Zimní semestr 2012

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

Extrakce pojmenovaných entit

- Extrakce informací ze nestrukturovaných textů
- Rozpoznávání objektů, které mají zpravidla nějaké jméno (města, lidé, firmy, produkty)
- Někdy rozšířeno i na problém extrakce relací
  - Např. Firma A *koupila* Firmu B – rozpoznání, že se jedná o akvizici
- Často také označováno jako Rozpoznávání pojmenovaných entit – Named Entity Recognition (NER)

Podproblémy NER

- Rozpoznávání pojmenovaných entit – identifikace v textu, tj. určení, kde přesně se pojmenovaná entita nachází.
  - Např. „United States of Tara“
    - Jde o dvě entity (United States a Tara)?
    - Nebo máme na mysli název seriálu United States of Tara?
- Dezambiguace
  - Rozpoznání typu pojmenované entity (Paris – město, nebo osoba)?
  - Mapování na konkrétní entitu v rámci typu (Paris – o kterou z desítek Paříží jde? Wikipedia jich zná 26 jenom v USA)

Přístupy k NER

- Založené na slovnících (gazetteers)
- Ručně vytvořená pravidla
- Strojové učení
  - Hidden Markov Models (HMM)
  - Conditional Random Fields (CRFs)
  - Neuronové sítě
  - k Nearest Neighbors (kNN)
  - Shlukování
- Ensemble Learning
  - Bagging, Boosting
  - Neuronové sítě

Vybrané nástroje

- Extraktory pojmenovaných entit
  - OpenCalais- http://viewer.opencalais.com/
  - Zemanta - http://www.zemanta.com/demo/
- Integrované platformy
  - GATE - http://gate.ac.uk/
  - UIMA - http://uima.apache.org/
  - NLP2RDF – http://nlp2rdf.org/

Extrakce relací – přístupy

- Vzory M. Hearstové (Hearst patterns)
  - Vylepšené regulární výrazy
  - Identifikace obvyklých vzorců vyskytujících se v textu, z nichž je možné dovodit daný predikát, typicky pro vztah nadřazeného a podřazeného termínu
    např.: `such NP as { NP, }* {(or|and)} NP`
- Využití redundance dat; iterativní proces nastartovaný přes „seed“ data
  - DIPRE (Dual Iterative Pattern Relation Expansion/Extraction)
    - Autor S. Brin (Google)
    - Primárně pro semi-strukturovaná data; samostatný proces pro zvolenou relaci
    - Dány vzory instancí, které jsou spolu v daném vztahu (např. Hillary, Bill -> manželé)
    - Prohledávání velkého korpusu textových dat a vyhledávání společného výskytu obou (resp. všech) entit, s důrazem na přesnost
    - Pokud jsou nalezeny, je sestaven regulární výraz, na jehož základě se v korpusu vyhledávají další entity ve stejném vztahu
    - Pracuje s pozitivními i negativními příklady
  - NELL (Never-Ending Language Learner) - CMU
    - Populování pevně dané ontologie („seed“ jsou instance tříd a relací); využívá disjunktnost
    - http://rtw.ml.cmu.edu/rtw/
  - ReVerb – U. Washington
    - Otevřená extrakce (OpenIE); časté souvýskyty sloves s koncepty vedou na vznik relací
    - http://www.cs.washington.edu/research/knowitall

Dezambiguace – přístupy

- Indexování
  - V podstatě sestavení slovníku obsahujícího možné typy pro každou entitu
- Dezambiguace typů
  - Pokud mám více možností (např. Paris – osoba, Paris – místo)
  - Na základě kontextu se určuje správný typ entity

DBpedia.org

- Extrakce strukturovaných informací z Wikipedia.org
- Informace jsou reprezentovány jako RDF
- SPARQL endpoint: http://dbpedia.org/sparql/

- Např.: najdi všechny filmy, které natočil režisér filmu Tokyo Mew Mew

```
PREFIX dbprop: <http://dbpedia.org/property/>
PREFIX db: <http://dbpedia.org/resource/>
SELECT ?who ?work ?genre WHERE { db:Tokyo_Mew_Mew
   dbprop:director ?who .
   ?work dbprop:director ?who .
   OPTIONAL { ?work dbprop:genre ?genre }
}
```

About: Prague

An Entity of Type : populated place, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

Prague is the capital and largest city of the Czech Republic. Situated in the north-west of the country on the Vltava river, the city is home to about 1.3 million people, while its metropolitan area is estimated to have a population of over 2.3 million. The city has a temperate oceanic climate with warm summers and chilly winters. Prague has been a political, cultural and economic centre of Europe and particularly central Europe during its 1,100 year existence.

dbpedia-owl:areaTotal	496000000.000000 (xsd:double)
dbpedia-owl:country	dbpedia:Czech_Republic
dbpedia-owl:leaderName	dbpedia:Bohuslav_Svoboda
dbpedia-owl:leaderParty	dbpedia:Civic_Democratic_Party_(Czech_Republic)
dbpedia-owl:leaderTitle	Mayor
dbpedia-owl:maximumElevation	399.000000 (xsd:double)
dbpedia-owl:motto	(Prague, Head of the State; Latin)
dbpedia-owl:populationAsOf	2011-01-14 (xsd:date)
dbpedia-owl:populationMetro	2300000 (xsd:integer)
dbpedia-owl:populationTotal	1290846 (xsd:integer)
dbpedia-owl:postalCode	1xx xx
dbpedia-owl:thumbnail	http://upload.wikimedia.org/wikipedia/commons/thumb
dbpedia-owl:timeZone	dbpedia:Central_European_Time

Kde se berou informace na DBpedii?

- Wikipedia dává pravidelně k dispozici exporty publikovaných článků ve zdrojovém kódu (Wiki Markup)

```
'''Prague''' ({{IPAc-en|icon|□|p|r|ɑː|g}}; {{lang-
cs|Praha}} {{IPA-cs|□praɦa|pron|Cs-Praha.ogg}}) is the
capital and [[List of cities in the Czech
Republic|largest city]] of the [[Czech
Republic]].<ref>{{cite
web|url=http://worldinfozone.com/facts.php?country=Czec
hRepublic |title=Czech Republic Facts |publisher=World
InfoZone |accessdate=14 April 2011}}</ref>
```

DBpedia Information Extraction Framework

- Nejprve dojde ke stažení dumpů se zdrojovými kódy jednotlivých článků
- Wikiparser přeloží staženou stránku na svou vnitřní reprezentaci (Abstract Syntax Tree = AST)
- Extraktor provádí převod takto zpracované stránky na grafovou reprezentaci
- Následně je výsledek extrakce uložen do cílového úložiště (v případě anglické DBpedie jde o Virtuoso)

DBpedia extraktor

- Celá škála extraktorů
- Většinou na bázi regulárních výrazů
- Příklady nejdůležitějších:
  - Label Extractor – extrahuje názvy článků
  - Mapping Extractor – extrahuje informace z infoboxů a mapuje je na společnou ontologii
  - Infobox Extractor – také extrahuje informace z infoboxů, ale neprovádí mapování
  - Page Links Extractor – extrahuje odkazy na související články na Wikipedii
- K dispozici je mnoho dalších extraktorů – možnost vytvářet nové – projekt je OpenSource

Prague

From Wikipedia, the free encyclopedia

Coordinates: 50°05′N 14°25′E

This article is about the capital of the Czech Republic. For other uses, see Prague (disambiguation).

Prague (/ˈprɑːg/; Czech: *Praha* pronounced [ˈpraɦa] (listen)) is the capital and largest city of the Czech Republic.[4] Situated in the north-west of the country on the Vltava river, the city is home to about 1.3 million people, while its metropolitan area is estimated to have a population of over 2.3 million.[5] The city has a temperate oceanic climate with warm summers and chilly winters.

Prague has been a political, cultural and economic centre of Europe[*citation needed*] and particularly central Europe[*citation needed*] during its 1,100 year existence. For centuries, during the Gothic and Renaissance eras, Prague was the permanent seat of two Holy Roman Emperors and thus was also the capital of the Holy Roman Empire.[*citation needed*] Later it was an important city in the Habsburg Monarchy and the Austro-Hungarian Empire,[*citation needed*] and after World War I became the capital of Czechoslovakia. The city played major roles in the Protestant Reformation, the Thirty Years' War, and in 20th-century history, during both World Wars and the post-war Communist era.

Prague is home to a number of famous cultural attractions, many of which survived the violence and destruction of twentieth century Europe. Main attractions include the following: Prague Castle, the Charles Bridge, Old Town Square, the Jewish Quarter, the Lennon Wall, and Petřín hill. Since 1992, the extensive historic centre of Prague has been included in the UNESCO list of World Heritage Sites.

Prague boasts more than ten major museums, along with countless theatres, galleries, cinemas, and other historical exhibits. Also, Prague is home to a wide range of public and private schools, including the famous Charles University. Its rich history makes it a popular tourist destination, and the city receives more than 4.1 million international visitors annually, as of 2009.[6][7] Prague is classified as a global city.[*citation needed*]

A modern public transportation system connects the city. Prague is also accessible by road, train, and air.

Contents [hide]

t: Prague

y of Type : populated place, from Named Graph : dia.org, within Data Space : dbpedia.org

DBpedia

est city of the Czech Republic. Situated in the north-west of the city is home to about 1.3 million people, while its metropolitan area is estimate pulation of over 2.3 million. The city has a temperate oceanic climate with warm s chilly winters. Prague has been a political, cultural and economic centre o ope and particularly central Europe during its 1,100 year existence.

dbpedia-owl:areaTotal	496000000.000000 (xsd:double)
dbpedia-owl:country	dbpedia:Czech_Republic
dbpedia-owl:leaderName	dbpedia:Bohuslav_Svoboda
dbpedia-owl:leaderParty	dbpedia:Civic_Democratic_Party_(Czech_Republic)
dbpedia-owl:leaderTitle	Mayor
dbpedia-owl:maximumElevation	399.000000 (xsd:double)
dbpedia-owl:motto	(Prague, Head of the State; Latin)
dbpedia-owl:populationAsOf	2011-01-14 (xsd:date)
dbpedia-owl:populationMetro	2300000 (xsd:integer)
dbpedia-owl:populationTotal	1290846 (xsd:integer)
dbpedia-owl:postalCode	1xx xx
dbpedia-owl:thumbnail	http://upload.wikimedia.org/wikipedia/commons/thumb
dbpedia-owl:timeZone	dbpedia:Central_European_Time

Mapping Extractor

- Infobox extraktor extrahuje informace z takzvaných infoboxů (tabulky v pravé části článků na Wikipedii)
- **Problém:** Wikipedia je editována velkým množstvím různých lidí. Pojmenování jednotlivých vlastností je nekonzistentní. Např. místo narození, původ, rodné město
- **Řešení:** Ruční vytváření mapovací pravidel prostřednictvím Mappings Wiki (http://mappings.dbpedia.org)
- Mapping Extractor následně využívá tyto informace a mapuje informace z infoboxů na společnou DBpedia ontologii

Příklad mapovacího pravidla

Mapping cs:Infobox Hokejista

This is the mapping for the Wikipedia template Infobox_Hokejista.

Test this mapping with some example Wikipedia pages.

Read more about mapping Wikipedia templates.

Template Mapping (help)	
map to class	IceHockeyPlayer

Mappings

Property Mapping (help)	
template property	jméno hráče
ontology property	foaf:name

Property Mapping (help)	
template property	fotka
ontology property	picture

Leden 2011 – statistiky DBPedia

- 364 000 osob
- 462 000 míst
- 169 000 živočišných druhů
- 148 000 organizací
- 99 000 hudebních alb

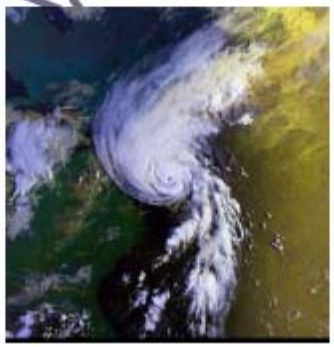
DBpedia Spotlight

- Nástroj pro rozpoznávání a disambiguaci pojmenovaných entit
- Vystavěn nad DBpedií a Wikipedií
- Mapuje pojmenované entity na konkrétní entity z DBpedie
- Využívá znalostní báze Wikipedie (wikilinky, přesměrování, rozcestníky/disambiguation pages)

Spotlight funkce

- Nalezení možných zmínek o dané entitě na DBpedii (podobná hodnota *rdfs:label*)
- V případě více možných kandidátů na základě kontextu ve větě odvoď nejpravděpodobnější kombinaci konkrétních entit (podle jejich vzájemného prolinkování na Wikipedii a podobnosti jejich popisků se zpracovávaným textem)

Mississippi, one of Bob's later songs, was first recorded by Sheryl on her album.



Ukázka extrakce z nestrukt. textu

- GATE
- Nástroj vyvíjený na VŠE: http://ner.vse.cz/thd/
  - V rámci projektu EU LinkedTV