Masarykova univerzita
Fakulta informatiky

DBpedia

PV070 Digitální knihovny

Tomáš Effenberger
3. ročník, UČO 410350

Brno, 16. listopadu 2014

Úvod

DBpedia[1] je báze znalostí získaná extrakcí strukturovaných informací z Wikipedie. Fakta jsou uložena ve formě RDF[2] grafu, což je standardní způsob reprezentace znalostí, který umožňuje kompaktní uložení a snadné dotazování. Celou bázi znalostí lze stáhnout a nahrát do vlastní databáze[3], data jsou dostupná pod licencemi *Creative Commons Attribution-ShareAlike License* a *GNU Free Documentation License*. Je také možná vyzkoušet dotazy přímo přes veřejný přístupový bod (*DBpedia SPARQL endpoint*[4]).

DBpedia umožňuje snadno najít odpověď na otázky, pro které textové vyhledávání vhodné není. Například nalezení všech českých spisovatelů, kteří se narodili v 18. století nebo všechny Evropské státy, které mají méně než milion obyvatel.

Projekt DBpedia začal v roce 2007 spoluprací společnosti *OpenLink Software* a lidí ze dvou německých univerzit (Svobodná univerzita Berlín a Lipská univerzita).

Extrakce informací z Wikipedie

DBpedia extrahuje z Wikipedie strukturovaná data, např. názvy článků, odkazy mezi články, odkazy na externí stránky, odkazy na obrázky, údaje o kategoriích a údaje z informačních panelů (*infoboxů*), které jsou součástí mnoha článků a podléhají vždy určité šabloně (např. šablona pro stát, osobu, spisovatele atp.)

Problémem s infoboxy je nejednoznačnost pojmenování sémanticky shodných položek, např. „*date of birth*“ vs. „*birthdate*“ (nebo dokonce vs. „*datum narození*“). Aby mohly být údaje popsané jednoznačně, bylo vytvořeno mapování označení na konzistentní *ontologii* (ontologií se rozumí popis vlastností, tříd a jejich hierarchie). Ontologie aktuálně (listopad 2014) obsahuje více než 685 tříd a 2795 vlastností. [1]

Pro lepší představu uvádíme začátek hierarchie tříd, celou ji pak najdete na `http://mappings.dbpedia.org/server/ontology/classes/`.

---

1. `http://dbpedia.org/`
2. *Resource Description Framework*
3. `http://wiki.dbpedia.org/Downloads`
4. `http://dbpedia.org/sparql`

```
- Thing
  - Activity
    - Game
    - Sport
  - Agent
    - Deity
    - Family
    - Organization
    - Person
      ...
    ...
```

Kromě vytvoření ontologie jako takové bylo také potřeba definovat mapování nejednoznačných položek infoboxů na tuto konzistentní ontologii. Těchto mapování již existuje 4339, z toho 586 pro angličtinu. [1]

Reprezentace znalostí

DBpedia ukládá znalosti ve formě *RDF grafu*, tj. každý fakt je reprezentován trojicí subjekt – predikát – objekt. [4, s. 63] Uveďme několik příkladů:

```
Alan Turing - type - person
Alan Turing - birthdate - 23 June 1912
Alan Turing - nationality - British
```

Ve skutečnosti je to trochu složitější než v tomto uvedeném příkladu. Protože chceme, aby subjekty i predikáty byly vyjádřeny jednoznačně, nebude nám stačit jenom řetězec (`"Alan_Turing"`), ale budeme vyžadovat URI[5] (`http://dbpedia.org/resource/Alan_Turing`), které můžeme rozdělit na prefix a jméno. Ukážeme na příkladu:

URI pro Alana Turinga	`http://dbpedia.org/resource/Alan_Turing`
prefix	`http://dbpedia.org/resource/`
jméno	`Alan_Turing`

Zatímco subjekty a predikáty musí být jednoznačná URI, objekty mohou být buď URI nebo pouhé literály, např. `"British"@en`.[6] Literály nemusí být nutně jenom řetězce, ale třeba také celé číslo, desetinné číslo, čas nebo datum (`1912-06-23+02:00`).

5. *Uniform Resource Identifier*
6. `en` značí jazyk řetězce.

Kompletní příklad RDF trojic uvedených výše by pak vypadal následovně:

```
<http://dbpedia.org/resource/Alan_Turing>
    <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
    <http://xmlns.com/foaf/0.1/Person> .
<http://dbpedia.org/resource/Alan_Turing>
    <http://dbpedia.org/ontology/birthDate>
    1912-06-23+02:00
<http://dbpedia.org/resource/Alan_Turing>
   <http://dbpedia.org/property/nationality>
   "British"@en
```

Nebo s využitím odděleného definování URI prefixů (tzv. *Turtle syntax*):

```
@prefix rsrc: <http://dbpedia.org/resource/>
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix dbpedia-owl: <http://dbpedia.org/ontology/>
@prefix dbprop: <http://dbpedia.org/property/>
@prefix foaf: <http://xmlns.com/foaf/0.1/>

rsrc:Alan_Turing
  rdf:type foaf:person ;
  dbpedia-owl:birthDate 1912-06-23+02:00 ;
  dbprop:nationality "British"@en .
```

Dotazování

K dotazování nad RDF grafem slouží jazyk SPARQL.[7] Dotaz může obsahovat následující části:

PREFIX – deklarace prefixů URI (abychom mohli URI dále zkracovat)

FROM – určení datasetu (RDF grafu), nad kterým se má dotaz provést

SELECT – které informace mají být vráceny

WHERE – podmínky, které musí vrácená data splňovat

ORDER BY – řazení vrácených výsledků

LIMIT – omezení na počet vrácených výsledků

7. Název *SPARQL* je rekurzivní akronym *SPARQL Protocol and RDF Query Language*.

Lze také vybírat z různých výstupních formátů (XML, RDF, JSON, CSV, HTML a další).

Pro lepší představu uvedeme několik konkrétních příkladů.

Všechna fakta o Alanu Turingovi:

```
PREFIX rsrc: <http://dbpedia.org/resource/>
SELECT DISTINCT ?p ?o
WHERE {
  rsrc:Alan_Turing ?p ?o .
}
```

Datum a místo narození Alana Turinga:

```
PREFIX rsrc: <http://dbpedia.org/resource/>
PREFIX owl: <http://dbpedia.org/ontology/>
SELECT *
WHERE {
  rsrc:Alan_Turing owl:birthDate ?date ;
                   owl:birthPlace ?place .
}
```

Současný stav

Aktuální verze DBpedie (listopad 2014)[8] obsahuje přibližně 580 milionů faktů (záznamů ve formě RDF trojic) získaných z anglické verze Wikipedie. Tato fakta popisují celkem 4,58 milionu věcí, z toho je přibližně 1,4 milionu osob, 735 tisíc míst, 411 tisíc kreativních děl (alb, filmů, videoher atp.), 241 tisíc organizací, 251 tisíc biologických druhů a 6 tisíc chorob. [1]

Projekt DBpedia se však neomezuje jenom na angličtinu. Existují verze DBpedie pro 125 různých jazyků, včetně češtiny. Sečteno přes všechny lokalizované verze, DBpedia obsahuje 3 miliardy záznamů. Z toho je 25,2 milionu odkazů na obrázky a 29,8 milionů odkazů na externí stránky. [1]

DBpedia je také propojena s dalšími bázemi znalostí (např. *Wikidata*, *Freebase*, *Project Gutenberg*, *Geonames*[9]) pomocí 50 milionů odkazů. [1]

Protože článků na Wikipedii neustále přibývá (nemluvě o editacích již existujících článků), stanou se zveřejněná data DBpedie velmi rychle zastaralá.

---

8. *DBpedia Version 2014* dostupná na adrese http://wiki.dbpedia.org/Downloads2014
9. Kompletní seznam: http://wiki.dbpedia.org/Interlinking

Proto vznikl systém *DBpedia live*,[10] který všechny změny Wikipedie okamžitě promítá do DBpedie, se zpožděním nejvýše v řádu minut. [3, s. 14]

Česká DBpedia

Existuje i česká verze DBpedie, ta aktuálně (listopad 2014) popisuje necelých 300 tisíc věcí pomocí 30 milionů faktů.[11]

Problémem je zatím nedostatečné mapování infoboxů české Wikipedie[12] a v důsledku nepříliš mnoho čistých a konzistentních faktů.

Situace se však postupně trochu zlepšuje a časem bude jistě i česká DBpedie použitelná pro zajímavé aplikace. V dubnu 2014 byl navíc po delší době obnoven veřejný přístupový bod české DBpedie, takže nyní je možné si vyzkoušet dotazy nad českou DBpedií na adrese `http://cs.dbpedia.org/sparql`.

Využití DBpedie

Již na začátku jsme uvedli možné využití DBpedie jako nástroj sofistikovaného vyhledávání ve Wikipedii. DBpedii však lze využít pro mnoho dalších úloh. Mít k dispozici takovou bázi znalostí je například velmi užitečné v oblasti zpracování přirozeného jazyka (např. pro desambiguaci nebo rozpoznávání pojmenovaných entit v textu). V oblasti digitálních knihoven může DBpedia nabídnout množství strukturovaných informací o autorech a jejich dílech. [3]

Klasickým využitím bází znalostí je také odpovídání na otázky (*Question answering*). Například Google využívá pro sémantické vyhledávání[13] vlastní bázi znalostí *Google Knowledge Graph*, jehož část je veřejně dostupná pod názvem *Freebase* (a je rozšiřována komunitou). DBpedia byla využita v rámci projektu *DeepQA project* pro vytvoření systému *IBM Watson*, který v roce 2011 vyhrál vědomostní hru *Jeopardy!* proti dvěma bývalým lidským vítězům. [2, s. 69]

---

10. Stránky projektu: `http://live.dbpedia.org/`
11. Aktuální informace: `http://wiki.dbpedia.org/Datasets/DatasetStatistics`
12. Statistiky mapování: `http://mappings.dbpedia.org/server/statistics/cs/`
13. *Sémantické vyhledávání* znamená pochopení dotazu a poskytnutí odpovědi přímo, místo odkazů na stránky, které možná odpověď obsahují.

Závěr

DBpedia přestavuje obrovský zdroj strukturovaných informací a lze očekávat, že s jejím kvantitativním a kvalitativním vylepšováním poroste také její využití. Není však jasné, jestli bude i v budoucnu hlavní bází znalostí, nebo tuto roli přebere např. Freebase. Největší odlišností DBpedie oproti Freebase je silná vazba na Wikipedii. Tato vazba na jednu stranu přináší neustálý přísun dat tím, jak uživatelé rozšiřují a doplňují Wikipedii. Zpočátku velké problémy špinavých a nekonzistentních dat se už z velké části podařilo vyřešit. Jako slabina se však může projevit, že DBpedie obsahuje *pouze* data z Wikipedie a jediná možnost, jak přidat něco do DBpedie je editovat Wikipedii. Naproti tomu Freebase obsahuje data z celé řady zdrojů a editovat ji lze přímo. Bude zajímavé sledovat, která z těchto bází znalostí získá navrch.

Bibliografie

[1] Christian Bizer, Volha Bryl a Daniel Fleischhacker. *DBpedia Version 2014 released*. 2014. URL: http://blog.dbpedia.org/2014/09/09/dbpedia-version-2014-released/ (cit. 16. 11. 2014).

[2] David Ferrucci et al. "Building Watson: An overview of the DeepQA project". In: *AI magazine* 31.3 (2010), s. 59–79.

[3] Jens Lehmann et al. "DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia". In: *Semantic Web Journal* (2014).

[4] Toby Segaran, Colin Evans a Jamie Taylor. *Programming the semantic web*. O'Reilly Media, Inc., 2009.