DBpedia 学习笔记

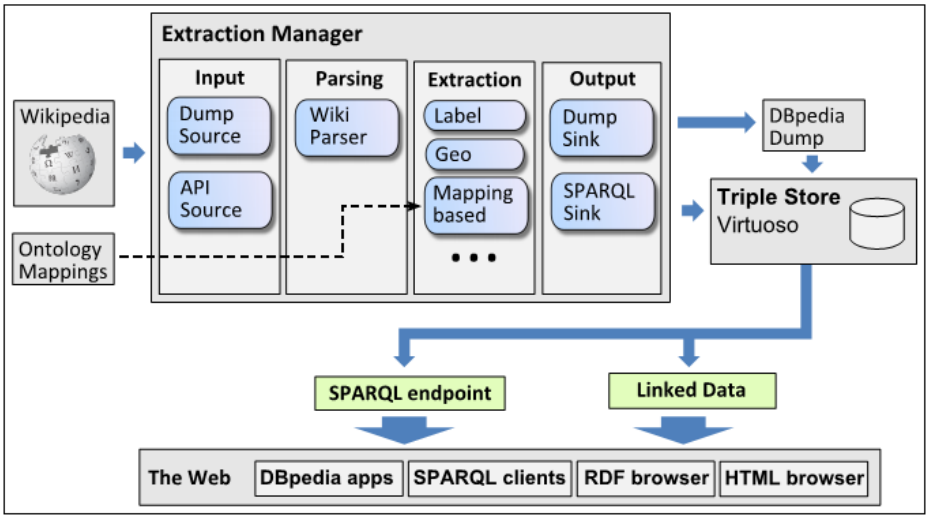
By Nancy Liu

一. 概述

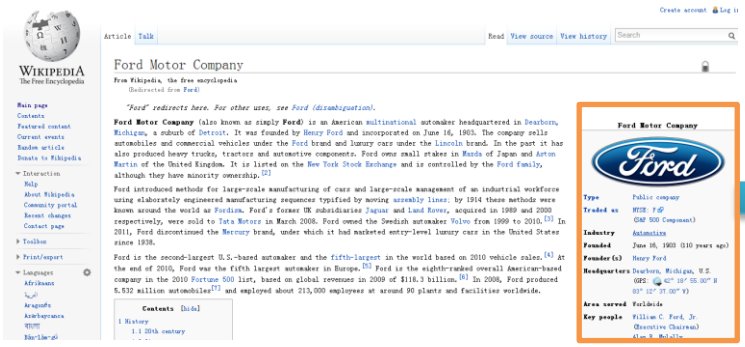
1. DBpedia 产生动机：由于 Wikipedia 固有结构的限制，使得一些查询无法实现，如“流过莱茵河的所有河”，“意大利 18 世纪之后的作曲家”。DBpedia 从 Wikipedia 中抽取结构化的信息，并建立语义网络。
2. DBpedia 项目开始时间，2006 年。
3. 本文将从 DBpedia 的知识抽取框架、DBpedia 本体、DBpedia 与 Linked Data 三个方面来介绍 DBpedia.

二. 抽取框架

1. 整体框架
   a) 输入：一种是 Wikipedia 仓库，一种是 MediaWiki 提供的 API。
   b) 解析：将 Wikipedia 中获取的源码转换为抽象语法树。
   c) 抽取：每一个 Wikipedia 的页面就是一个抽象语法树，对这个抽象语法树进行信息抽取，比如标签、摘要、地理坐标等。每一个抽取都利用一个抽象语法树，并生成一系列 RDF 声明。
   d) 输出：将得到的 RDF 声明写入槽。
   下图示意了 DBpedia 的抽取框架：



2. 抽取器
   DBpedia 抽取的内容主要来自 Wikipedia 的信息框以及页面中的标记。
   其中通常位于右上角的信息框又称为 DBpedia 最主要抽取的对象。
   如搜索”ford”.（如下图）这个信息框可以提供结构化的信息。



Ford Motor Company

Type	Public company
Traded as	NYSE: F (S&P 500 Component)
Industry	Automotive
Founded	June 16, 1903 (110 years ago)
Founder(s)	Henry Ford
Headquarters	Dearborn, Michigan, U.S. (GPS: 42° 18′ 55.00″ N 83° 12′ 37.00″ W)
Area served	Worldwide
Key people	William C. Ford, Jr. (Executive Chairman) Alan R. Mulally (President & CEO)
Products	Automobiles Automotive parts
Services	Automotive finance Vehicle leasing Vehicle service
Revenue	▲ US$136.26 billion (2011)[1]
Operating income	▲ US$8.681 billion (2011)[1]
Net income	▲ US$20.21 billion (2011)[1]
Total assets	▲ US$178.35 billion (2011)[1]
Total equity	▲ US$15.07 billion (2011)[1]
Employees	164,000 (2011)[1]
Divisions	Ford Lincoln Motorcraft
Subsidiaries	List [show]
Website	Ford.com

DBpedia 的抽取器可以分为以下四类：

a) 基于映射的信息框抽取器

解决的问题是从 infoBox 提取数据映射到 ontology.

infoBox 即 Wikipedia 中右上角的框，ontology 对应的是 DBpedia 中的本体。

b) 信息框“生”信息抽取

解决的问题是从 infoBox 提取数据直接存到 RDF.

即将 infoBox 中的属性-值对存放至 RDF 中，这种方式产生的数据语义并不丰富，质量还有待提高。

c) 特征提取

特征提取是利用一些抽取器将某一个确定的特征从 Wikipedia 的一篇文章（也可以理解为一个网页，不限于 infobox）中抽取出来，比如抽取一个坐标或者标签。

d) 统计的抽取

一些自然语言处理相关的抽取器，将所有 Wikipedia 的页面中的数据汇总起来，使能够提供一些统计值，比如页面链接数量、词数。

3. 信息框“生”信息抽取

直接将信息框数据存入 RDF 数据。

如下图所示：

```
{{Infobox automobile
| name         = Ford GT40
| manufacturer = [[Ford Advanced Vehicles]]
| production   = 1964-1969
| engine       = 4181cc
(...)
}}
```

```
dbr:Ford_GT40 [
  dbp:name "Ford GT40"@en;
  dbp:manufacturer dbr:Ford_Advanced_Vehicles;
  dbp:engine 4181;
  dbp:production 1964;
  (...)
] .
```

可以发现生成的 RDF 有两个问题：一个是资源无法关联到相关类别；二是对于 engine，production 这样的属性，语义并不明确。这也是为什么需要一个基于映射的信息框抽取。

4. 基于映射的信息框抽取

由于信息框的多样性以及上文所讨论的，“生”数据质量有待提高，基于映射的信息框抽取被提出。

它完成了将一个 Infobox 对应到一个 DBpedia 的本体(ontology)中。其中 Infobox 的属性对应了

ontology 的属性。
比如一个映射如下：

```
{{TemplateMapping
|mapToClass = Automobile
|mappings =
 {{PropertyMapping
 | templateProperty = name
 | ontologyProperty = foaf:name }}
 {{PropertyMapping
 | templateProperty = manufacturer
 | ontologyProperty = manufacturer }}
 {{DateIntervalMapping
 | templateProperty = production
 | startDateOntologyProperty = productionStartDate
 | endDateOntologyProperty = productionEndDate }}
 {{IntermediateNodeMapping
 | nodeClass = AutomobileEngine
 | correspondingProperty = engine
 | mappings =
    {{PropertyMapping
    | templateProperty = engine
    | ontologyProperty = displacement
    | unit = Volume }}
    {{PropertyMapping
    | templateProperty = engine
    | ontologyProperty = powerOutput
    | unit = Power }}
 }}
 (...)
}}
```

上面的映射的例子中，将 infobox 的时间对应到了开始时间和结束时间，将 infobox 中的 engine 对应到了排放量和功率两个值，这样就可以更加准确、具有语义的将 Infobox 的信息映射出来。
得到了新的 RDF 节点如下：

```
dbr:Ford_GT40 [
  rdf:type    dbo:Automobile;
  rdfs:label "Ford GT40"@en;
  dbo:manufacturer
             dbr:Ford_Advanced_Vehicles;
  dbo:productionStartYear
             "1964"^^xsd:gYear;
  dbo:productionEndYear "1969"^^xsd:gYear;
  dbo:engine [
             rdf:type AutomobileEngine;
             dbo:displacement "0.004181";
  ]
  (...)
] .
```

DBpedia 的基于映射抽取器为了实现映射的准确性和实时性，允许用户新建和编辑，和 Wikipedia 的开发性相“映射”。
DBpedia 同时提供了三个工具，分别是映射检验器、抽取测试器、映射工具，供用户使用。
http://mappings.dbpedia.org

5. URI 模式
对于每一个 Wikipedia 中的文章，将会有一些 URI 与之对应。
在 DBpedia 中主要有三个命名空间：

http://dbpedia.org/resource/：与 Wikipedia 中的网页一一对应。
http://dbpedia.org/resource/：与 infobox 中的属性一一对应。
http://dbpedia.org/resource/：与 dbpedia 中的本体一一对应。
对于基于映射的抽取器，它可以通过编辑适应多国语言，所以 dbpedia 有两种数据集——本体数据集和标准数据集。这里的本地不是指存放在本体，而是指所在地，即用当地语言描述的东西。对于本地集，命名空间的 URI 前缀变化为：http://<lang>.dbpedia.org/resource/.

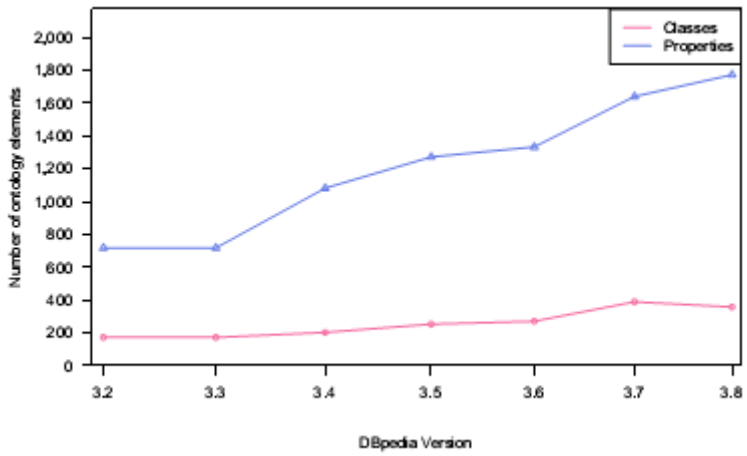
6. 自然语言处理抽取

DBpedia 提供了关于自然语言处理的数据集，目前有四个：话题标签(topic signatures)，文法类别(grammatical gender)， 词汇(lexicalization)，和主题概念(thematic concept)。

a) Lexicalization: 这是为了给出 DBpedia 中的别名统计信息而生成的数据集。它的作用是，给定一个词语，可以判断它可能表示的所有概念，包括以这个词为名字或别名的所有概念。同时会给出一个“分数”，这个分数表示了利用这个词表示这个概念的概率。
b) Topic signatures: 也就是给 DBpedia 中的每个 resource（就是与 Wikipedia 中的网页对应的资源）制作一个话题标签，以概括这个资源所围绕的话题。这个数据集的产生过程是：Wikipeida 中出现的每个词都是一个维度，每个 DBpedia 中的 resource 被表达成一个空间向量（VSM），对应这个多维空间中的一个点。对于每个与某 resource 相关的词，计算其的 tf-idf 的权重，然后选择出与这个 resource 关联最近的一些词，作为这个 resource 的话题标签。
c) Thematic: 这个抽取器旨在对 DBpedia 中的概念确定其主题，在 Wikipedia 中，许多类别下都有一篇文章来交待这个类别的主题，DBpedia 利用这个，标注了概念或实体的主题。
d) Grammatical gender: 这个部分可以针对 Person 这个本体，进行性别分析。在从 Wikipedia 到 DBpedia 的映射中，如果出现了 Person 这种实体，则统计这篇文章中出现的表征性别的形容词、代词等，然后以统计的方法确定这个人的性别。

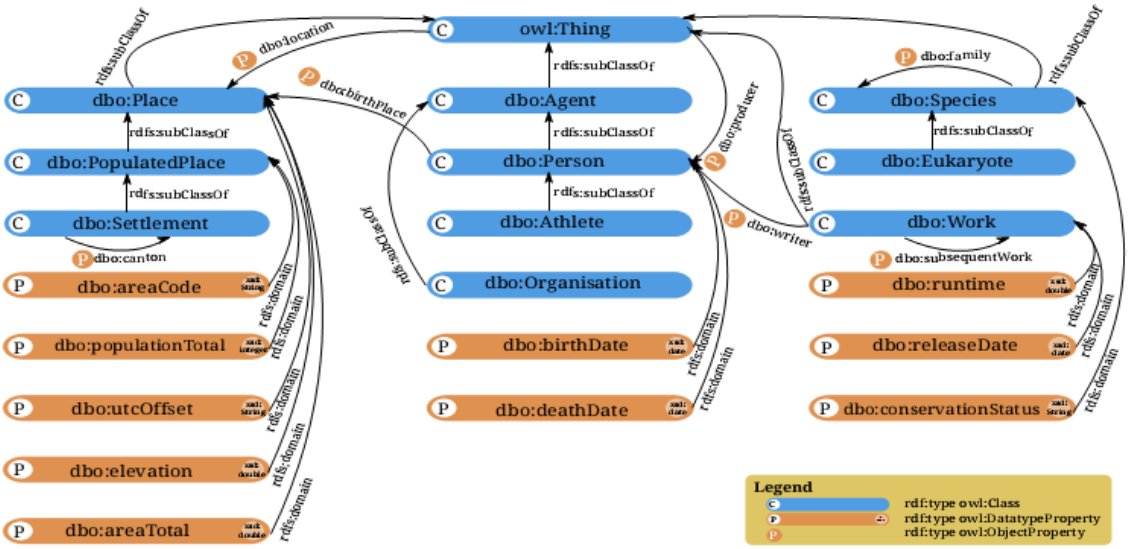
三. DBpedia 本体

DBpedia 本体目前包括了 320 个类别，类别之间包含层次关系，深度可以达到 5，深度控制在 5 以内是为了便于数据的使用，比如可视化或者导航。同时有 1650 个不同的属性来描述这些类别。下图展示了类别和属性的数量，随着 DBpedia 版本的变化：



可以发现，类别的数量并没有明显的变化，证明最初的类别体系是实用性很强的，而属性在不断递增，这得益于 DBpedia 与 Wikipeida 的映射关系越来越丰富，也得益于 Wikipedia 的内容越来越丰富。

下图是 DBpedia 本体模型的一个局部，其中 C 代表 owl 中定义的类别，白底的 P 代表特定数据类型的属性，黄底的 P 代表对象属性：



四. DBpedia 与 Linked Data

DBpedia 与正在不断第三方数据集相关联。

1. Outgoing Links：与外部数据相关联主要用到的工具有两种，一种是比较常用来进行关联数据集相关联的 Silk 和 LIMES，一种是针对某个数据集个性化定制的脚本。截至 2013 年 4 月份，已经与多个数据集相关联，如下图，其中第二列表示关联的关系，第三列表示建立的关联数量，最后一列表示使用的工具，空白部分是指还未与最新版本的数据集关联：

Data set	*Predicate*	*Count*	*Tool*
Amsterdam Museum	owl:sameAs	627	S
BBC Wildlife Finder	owl:sameAs	444	S
Book Mashup	rdf:type	9100	
	owl:sameAs		
Bricklink	dc:publisher	10100	
CORDIS	owl:sameAs	314	S
Dailymed	owl:sameAs	894	S
DBLP Bibliography	owl:sameAs	196	S
DBTune	owl:sameAs	838	S
Diseasome	owl:sameAs	2300	S
Drugbank	owl:sameAs	4800	S
EUNIS	owl:sameAs	3100	S
Eurostat (Linked Stats)	owl:sameAs	253	S
Eurostat (WBSG)	owl:sameAs	137	
CIA World Factbook	owl:sameAs	545	S
flickr wrappr	dbp:hasPhoto-	3800000	C
	Collection		
Freebase	owl:sameAs	3600000	C
GADM	owl:sameAs	1900	
GeoNames	owl:sameAs	86500	S
GeoSpecies	owl:sameAs	16000	S
GHO	owl:sameAs	196	L
Project Gutenberg	owl:sameAs	2500	S
Italian Public Schools	owl:sameAs	5800	S
LinkedGeoData	owl:sameAs	103600	S
LinkedMDB	owl:sameAs	13800	S
MusicBrainz	owl:sameAs	23000	
New York Times	owl:sameAs	9700	
OpenCyc	owl:sameAs	27100	C
OpenEI (Open Energy)	owl:sameAs	678	S
Revyu	owl:sameAs	6	
Sider	owl:sameAs	2000	S
TCMGeneDIT	owl:sameAs	904	
UMBEL	rdf:type	896400	
US Census	owl:sameAs	12600	
WikiCompany	owl:sameAs	8300	
WordNet	dbp:wordnet_type	467100	
YAGO2	rdf:type	18100000	
Sum		27211732	

Table 5

2. Incoming Links: 为了计算其他数据集对 DBpedia 的关联，DBpedia 使用了 sindice. Sindice 系统是一个基于 RDF 数据的搜索引擎，它利用爬虫将 RDF 数据收集起来，对每个实体，它的 URI 的二级域名为它所在的数据集。基于此，DBpedia 统计了关联至 DBpedia 的数据集。

domain	datasets	links
purl.org	498	6,717,520
dbpedia.org	248	3,960,212
creativecommons.org	2,483	3,030,910
identi.ca	1,021	2,359,276
l3s.de	34	1,261,487
rkbexplorer.com	24	1,212,416
nytimes.com	27	1,174,941
w3.org	405	658,535
geospecies.org	13	523,709
livejournal.com	14,881	366,025

Table 8

Top 10 datasets by incoming links in Sindice.

参考文献：

DBpedia - A Large-scale, Multilingual.