Web scale Information Extraction with LODIE

Anna Lisa Gentile  Ziqi Zhang  Fabio Ciravegna

Department of Computer Science, The University of Sheffield, UK

Semantics for Big Data, 16th November 2013
Motivation

Web scale Information Extraction (IE)

Challenges

Conclusions

Anna Lisa Gentile, Ziqi Zhang, Fabio Ciravegna

Washington, District of Columbia
www.dc.gov/

The official web site of the Government of the District of Columbia. Includes news, web links, and information about the city and about local government services.

Washington, D.C. - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Washington,_D.C.

Washington, D.C., formally the District of Columbia and commonly referred to as Washington, "the District", or simply D.C., is the capital of the United States. History of Washington, DC - Federal district - Retrocession - Home rule

Washington.org - Official Tourism Site of Washington DC
washington.org/

Get travel information on visiting Washington, DC, including attractions, things to do, events, maps and hotels. Washington DC, Washington D.C.
Attractions - 100 Free (& Almost Free) - Visitor Request Form - Free Attractions

Washington DC Tourism and Vacations: 419 Things to Do in ...
www.lonelyplanet.com/usa/washington-dc

Nov 10, 2008
Washington, DC travel recommendations and tips from Lonely Planet. Compare and book from 422 places to ...

News for washington dc

ACC hoops tournament coming to Washington DC
USA TODAY - 5 hours ago
The Atlantic Coast Conference announced Wednesday that the Verizon Center in Washington, D.C., will host its 2016 men’s basketball ...

The Victims of Drones Take on Washington, D.C.
Huffington Post - 1 day ago
Washington, DC: Newsroom out to stay classy with ‘Anchorman’ exhibit
Los Angeles Times - 1 day ago

Washington, D.C., travel guide - Wikitravel
wikitravel.org » North America » United States of America » Mid-Atlantic » Washington, D.C.
Web IE Methods and Systems

- Snowball [Agichtein and Gravano, 2000]
- KnowItAll [Etzioni et al., 2004]
- OpenIE/TextRunner [Banko et al., 2007]
- ReVerb [Fader et al., 2011]
- NELL [Carlson et al., 2010]
- PROSPERA [Nakashole et al., 2011]
- Probase [Wu et al., 2012]
- LODIE [Ciravegna et al., 2012]
Linked Open Data for Web-scale Information Extraction

- **Web-scale IE**
  - number of documents, domains, facts
  - efficient and effective methods required

- **Linked Open Data to seed learning**
  - “[. . .] a recommended best practice for exposing, sharing, and connecting data [. . .] using URIs and RDF” (linkeddata.org).
  - a large KB of typed instances, relations, annotations (e.g., RDFa)

- **Adapting to specific user information needs**
  - users define specific IE tasks by specifying the types of instances and relations to be learnt
LODIE: challenges

- **Define user needs**: How to let users define Web-IE tasks tailored to their own needs
- **Gather training data**: How to automatically obtain training seeds (and filter noise) from the LOD
- **Multi-strategy learning**: How to combine multi-strategy learning (e.g., from both structured and unstructured contents)
Challenges: user Information needs

**SoA** defines a generic IE task - KnowItAll, StatSnowball, PROSPERA, NELL, ExtremeExtraction [Carlson et al., 2010, Etzioni et al., 2004, Freedman et al., 2011, Nakashole et al., 2011] extracts “people, organisation, location” etc and their generic relations

**RQ** how to let users define Web-IE tasks tailored to their own needs - “drugs that treat hayfever"
User needs formalisation

Goal: Support users in formalising their information needs in a machine understandable format

- Hypothesis
  - Users define information needs in terms of ontologies
  - Users use different vocabularies in ontology creation

- Methods
  - Baseline: manually identify relevant ontologies on the LOD and define a view on them using tools like neon-toolkit.org
  - Ontology Pattern: use **Ontology Patterns** to bridge the “vocabulary gap”
Vocabulary Gap

- Linked Data - diverse vocabulary
  - what if we don’t know the vocabulary
  - possible overlap in vocabularies

```sql
select distinct ?film ?title where {
  ?film a <http://dbpedia.org/ontology/Film> .
}
```
Vocabulary Gap

- Linked Data - diverse vocabulary
  - what if we don’t know the vocabulary
  - possible overlap in vocabularies
Statistical Knowledge Patterns (SKP)

[Zhang et al., 2013]
SKP Construction Overview

[Zhang et al., 2013]
Challenges: training data

SoA requires certain amount of training/learning resources to be manually specified

RQ how to automatically obtain these (and filter noise) from the LOD
Identification of Learning Seeds

Goal: Automatically identify training data in the forms of triples and annotations to seed learning

- Hypothesis:
  - LOD can already contain answers to user needs in the forms of triples and annotations
  - The Web contains additional linguistic realisations of triples

- Method
  - From LOD - SPARQL queries to fetch seed triples (and annotations) matching the user needs
  - From the Web - Search for linguistic realisations of triples (identified above):
    - co-occurrence of related instances in textual contexts e.g. sentences
    - structural elements e.g., tables
Learning seeds: Dictionary Generation

- User Information Need formalisation
  - translate the concept and attributes of interest to the vocabularies used within the Linked Data
- given a SPARQL endpoint, query the exposed Linked Data to identify the relevant concepts
- select the most appropriate class and properties that describe the attributes of interest
- using the SPARQL endpoint, query the Linked Data to retrieve instances of the properties of interest

[Gentile et al., 2013]
Dictionary Generation example

Find all concepts matching the keyword “university”

SELECT DISTINCT ?uni WHERE {
FILTER regex(?lab,"university","i") }

Identify all properties defined with this concept

SELECT DISTINCT ?prop WHERE {
}

Extract all available values of this attribute

SELECT DISTINCT ?name WHERE{
?uni a <http://dbpedia.org/ontology/University> ;
<http://dbpedia.org/property/name> ?name .
FILTER (langMatches(lang(?name), 'EN')). 
}

[Gentile et al., 2013]
Preliminary experiment on Wrapper Induction task
LODIE Wrapper Induction: method

1. **Dictionary Generation**
   - for each attribute $a_{i,k}$ of each concept $c_i$, generate a dictionary $d_{i,k}$ for $a_{i,k}$ by exploiting Linked Data

2. **Page annotation**
   - $W_{j,i}$, Web pages from a website $j$ containing entities of $c_i$
   - annotate pages in $W_{j,i}$ by matching every entry in $d_{i,k}$ against the text content in the leaf nodes
   - for each match, create the pair $<xpath, value_{i,k}>$ for $W_{j,i}$

3. **Xpath identification**
   - for each attribute, gather all xpaths of matching annotations and their matched values
   - rate each path based on the number of different values it extracts
   - apply $wp_{j,i,k}$ best scoring xpath to re-annotate the website $j$ for attribute $a_{i,k}$.

[Gentile et al., 2013]
LODIE Wrapper Induction: Website Annotation

Get all annotations for attribute $a_{i,n}$, as ($<xpath, value_{i,k}>)$ pairs

- incompleteness of the auto-generated dictionaries
- the number of false negatives can be large (i.e., low recall)
- possible ambiguity in the dictionaries (e.g., 'Home' is a book title that matches part of navigation paths on many Web pages)
  - annotation does not involve disambiguation

---

[Gentile et al., 2013]
Filtering Learning Seeds

**Goal:** Filter noisy training data and select the most informative examples for learning

- **Hypothesis:**
  - Identified learning seeds can contain noise (causing “drifting-away”)
  - ... and can be redundant (causing unnecessary overheads)
  - good learning examples are **consistent** w.r.t. the learning task and **diverse**.

- **Method**
  - initial idea: use clustering of the seeds to define **consistency** and **variability**
  - **ongoing work:** exploiting the statistical distribution of seeds
Challenges: learning strategies

**SoA** Typically semi-supervised bootstrapping based learning from unstructured texts, prone to propagation of errors

**RQ** how to combine multi-strategy learning (e.g., from both structured and unstructured contents) to avoid drifting away from the learning task
Multi-Strategy Learning

**Goal:** Learning from different types (e.g., structured, unstructured) with different strategies to improve both recall and precision

- **Hypothesis:**
  - The same pieces of knowledge can be repeated in different forms, e.g., in tables v.s. sentences (re-enforcing evidence, Precision)
  - Some knowledge may be found only in one form or another (Recall)

- **Method:** multi-strategy learning
  - Learning from structures such as tables and lists [Milne and Witten, 2007, Limaye et al., 2010]
  - Inducing wrappers for regular pages [Kushmerick, 1997]
  - Lexical-syntactic pattern learning from free texts
  - Combine outputs from different processes
Challenges: publication of triples

**SoA** No integration with existing KB

**RQ** how to integrate with LOD
Integration with the LOD

**Goal:** Assign unique identifier to the extracted knowledge

- **Hypothesis:**
  - Knowledge that already exists in the LOD can be re-extracted and must be integrated

- **Method:**
  - simple, scalable disambiguation methods, e.g., by feature overlapping [Banerjee and Pedersen, 2002] and string distance metrics
Impact

LODIE timeliness

- LOD: first very large-scale information resource available for IE
- covering for a growing number of domains

LODIE output

- Web-scale IE task corpora, linked resources, etc.
- developed code will be available as open source
- all the data generated will be made available using a licence such as Open Data Commons (opendatacommons.org)
Further reading I

Snowball: extracting relations from large plain-text collections.

An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet.

Open information extraction from the web.
In *IJCAI*, pages 2670–2676.

Toward an architecture for never-ending language learning.

Lodie: Linked open data for web-scale information extraction.
Further reading II

Web-scale information extraction in knowitall: (preliminary results).
In Proceedings of the 13th international conference on World Wide Web, WWW ’04, pages 100–110, New York, NY, USA. ACM.

Identifying relations for open information extraction.

Extreme Extraction – Machine Reading in a Week.

Unsupervised wrapper induction using linked data.
In Proceedings of the seventh international conference on Knowledge capture, K-CAP ’13, pages 41–48, New York, NY, USA. ACM.

Wrapper Induction for information Extraction.
In IJCAI97, pages 729–735.

Annotating and Searching Web Tables Using Entities, Types and Relationships.
Further reading III


