Unsupervised Wrapper Induction using Linked Data

Anna Lisa Gentile Z. Zhang I. Augenstein F. Ciravegna

Organisations, Information and Knowledge

The University Of Sheffield.

Department of Computer Science, The University of Sheffield, UK

KCAP 2015, 8^{th} October 2015

Organisations, Information and Knowledge

The University Of Sheffield.

Outline

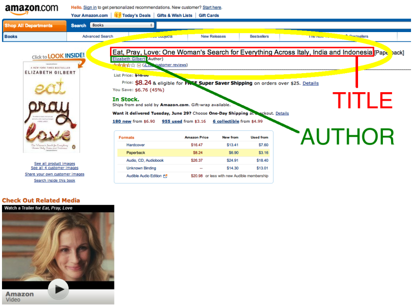
1. Wrapper Induction: Task definition
2. Proposed Methodology
3. Dataset
4. Experiment
5. Conclusions

Wrapper Induction: definition of the task

- Automatically learning wrappers using a collection of manually annotated Web pages as training data [Kushmerick, 1997, Muslea et al., 2003, Dalvi et al., 2009, Dalvi et al., 2011, Wong and Lam, 2010]
- Data is generally extracted from “detail” Web pages [Carlson and Schafer, 2008]
    - pages corresponding to a single data record (or entity) of a certain type or *concept* (also called *vertical* in the literature)
    - render various attributes of each record in a human-readable form

Wrapper Induction: example

Extracting book attributes on e-commerce websites



Web Scale Wrapper Induction

- Traditional wrapper induction task
  - schema
  - set of pages output from a single script
  - training data are given as input, and a wrapper is inferred that recovers data from the pages according to the schema.
- Web-scale wrapper induction task
  - large number of sites
  - each site comprising the output of an unknown number of scripts, along with a schema
  - per-site training examples can no longer be given

[Gulhane et al., 2011]

Learning Extraction rules: Characteristics

- Languages
  - Grammars
  - Xpath
  - OXpath
  - Xstring [Grigalis, 2013]
- Techniques
  - contextual rules (boundaries detection)
  - html-aware
  - visual features
  - hybrid approaches [Zhai and Liu, 2005, Zhao et al., 2005, Grigalis, 2013]
- Approaches
  - supervised
  - unsupervised
- Extraction dimensions
  - attribute-value pairs from tables
  - record level extractor (lists) [Álvarez et al., 2008, Zhai and Liu, 2005, Zhao et al., 2005]
  - detail page extractor

Outline

Proposed solution

- usage of *Linked Data* as background Knowledge
- flexible with respect to different domains
- no training data needed

Task definition

- C - set of *concepts* of interest $C = \{c_1, \ldots, c_i\}$
- their attributes $\{a_{i,1}, \ldots, a_{i,k}\}$
- a website containing Web pages that describe entities of each concept W_{c_i}
- **TASK**: retrieve attributes values for each entity on the Web pages

Methodology

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 $a_{i,k}$.

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

Dictionary Generation example

Find all concepts matching the keyword “university”

```
SELECT DISTINCT ?uni WHERE {
?uni rdf:type owl:Class ; rdfs:label ?lab .
FILTER regex(?lab,"university","i") }
```

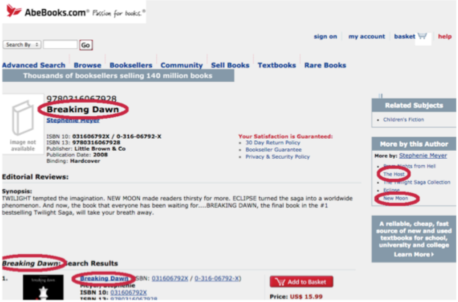
Identify all properties defined with this concept

```
SELECT DISTINCT ?prop WHERE {
?uni a <http://dbpedia.org/ontology/University> ; ?prop ?o . }
```

Extract all available values of this attribute

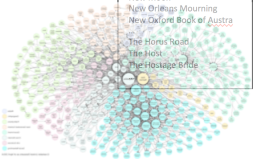
```
SELECT DISTINCT ?name WHERE{
?uni a <http://dbpedia.org/ontology/University> ;
<http://dbpedia.org/property/name> ?name .
FILTER (langMatches(lang(?name), 'EN')). }
```

Website Annotation



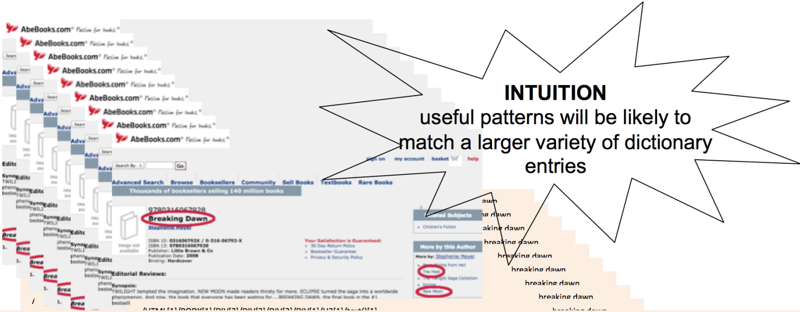
Book Titles

...
Breakfast of Champions
Breakfast on Pluto
Breakheart Pass
Breaking Dawn
...
Breaking Windows
Breaking news
Breaking point
...
New Mexico Sunrise
New Mexico Sunset
New Moon
New Orleans Mourning
New Oxford Book of Austra
...
The Horus Road
The Host
The Hostage Bride



/HTML[1]/BODY[1]/DIV[2]/DIV[2]/DIV[2]/DIV[1]/H2[1]/text()[1]	breaking dawn
/HTML[1]/BODY[1]/DIV[2]/DIV[2]/DIV[2]/DIV[4]/DIV[1]/H2[1]/EM[1]/text()[1]	breaking dawn
/HTML[1]/BODY[1]/DIV[2]/DIV[2]/DIV[2]/DIV[4]/TABLE[10]/TBODY[1]/TR[1]/TD[3]/B[1]/A[1]/text()[1]	breaking dawn
/HTML[1]/BODY[1]/DIV[2]/DIV[2]/DIV[2]/DIV[4]/TABLE[1]/TBODY[1]/TR[1]/TD[3]/B[1]/A[1]/text()[1]	breaking dawn
/HTML[1]/BODY[1]/DIV[2]/DIV[2]/DIV[2]/DIV[4]/TABLE[2]/TBODY[1]/TR[1]/TD[3]/B[1]/A[1]/text()[1]	breaking dawn
/HTML[1]/BODY[1]/DIV[2]/DIV[2]/DIV[2]/DIV[4]/TABLE[3]/TBODY[1]/TR[1]/TD[3]/B[1]/A[1]/text()[1]	breaking dawn
/HTML[1]/BODY[1]/DIV[2]/DIV[2]/DIV[2]/DIV[4]/TABLE[6]/TBODY[1]/TR[1]/TD[3]/B[1]/A[1]/text()[1]	breaking dawn
/HTML[1]/BODY[1]/DIV[2]/DIV[2]/DIV[2]/DIV[4]/TABLE[8]/TBODY[1]/TR[1]/TD[3]/B[1]/A[1]/text()[1]	breaking dawn
/HTML[1]/BODY[1]/DIV[2]/DIV[2]/DIV[3]/DIV[3]/UL[1]/LI[2]/A[1]/text()[1]	the host
/HTML[1]/BODY[1]/DIV[2]/DIV[2]/DIV[3]/DIV[3]/UL[1]/LI[5]/A[1]/text()[1]	new moon

XPath identification



/HTML[1]/BODY[1]/DIV[2]/DIV[2]/DIV[2]/DIV[1]/H2[1]/text()[1] 329
/HTML[1]/BODY[1]/DIV[2]/DIV[2]/DIV[2]/DIV[4]/DIV[1]/H2[1]/EM[1]/text()[1] 329
...
/HTML[1]/BODY[1]/DIV[2]/DIV[2]/DIV[2]/DIV[4]/TABLE[2]/TBODY[1]/TR[1]/TD[3]/B[1]/A[1]/text()[1] 197
/HTML[1]/BODY[1]/DIV[2]/DIV[2]/DIV[2]/DIV[4]/TABLE[1]/TBODY[1]/TR[1]/TD[3]/B[1]/A[1]/text()[1] 193
...
/HTML[1]/BODY[1]/DIV[2]/DIV[2]/DIV[2]/DIV[1]/DIV[6]/P[1]/B[1]/text()[1] 1

Outline

1. Wrapper Induction: Task definition
2. Proposed Methodology
3. Dataset
4. Experiment
5. Conclusions

Dataset

- 124K pages collected from 80 websites
- 8 verticals
  - textitAutos, Books, Cameras, Jobs, Movies, NBA Players, Restaurants, and Universities
  - 10 different websites (200 to 2,000 pages per website)
  - set of 3 to 5 common attributes to extract
- Ground truth
  - for each attribute-website pair, a file listing all possible attribute values found on the website is generated
  - using a few handcrafted regular expressions over each website
  - not all attributes are present on all websites (5 such cases in the dataset)

Hao et al. [Hao et al., 2011]

Dataset

Vertical	Web Sites	Web Pages	Attributes
Auto	10	17923	model (m), price (p), engine (e), fuel economy (f)
Book	10	20000	title (t), author (a), ISBN-13 (i), publisher (p), publish-date (pd)
Camera	10	5258	model (md), price (p), manufacturer (m)
Job	10	20000	title (t), company (c), location (l), date (d)
Movie	10	20000	title (t), director (d), genre (g), rating (r)
NBA player	10	4405	name (n), team (t), height (h), weight (w)
Restaurant	10	20000	name (n), address (a), phone (p), cuisine (c)
University	10	16705	name (n), phone (p), website (w), type (t)

Hao et al. [Hao et al., 2011]

Outline

1. Wrapper Induction: Task definition
2. Proposed Methodology
3. Dataset
4. Experiment
5. Conclusions

Experiments

- *Topline* experiment
    - artificially created dictionaries specifically tailored to the data
    - minimum level of noise
    - sets a higher limit of the performance of the method
- *Linked Data based WI* experiment
    - dictionaries generated from *Linked Data*
    - generated independently from the data
    - likely to contain noise

Experiments dictionaries

- *Topline* dictionaries
  - for each attribute of a vertical, collect all answers in the ground truth
    - each dictionary contains all (but not only) the true answers
- *Linked Data* dictionaries
  - manually explore *Linked Data* and create queries
  - query Sindice SPARQL endpoint[1]
  - not all verticals/attributes are covered by the *Linked Data*
  - results comparison only for covered attributes

[1] http://sparql.sindice.com/

Dictionaries statistics

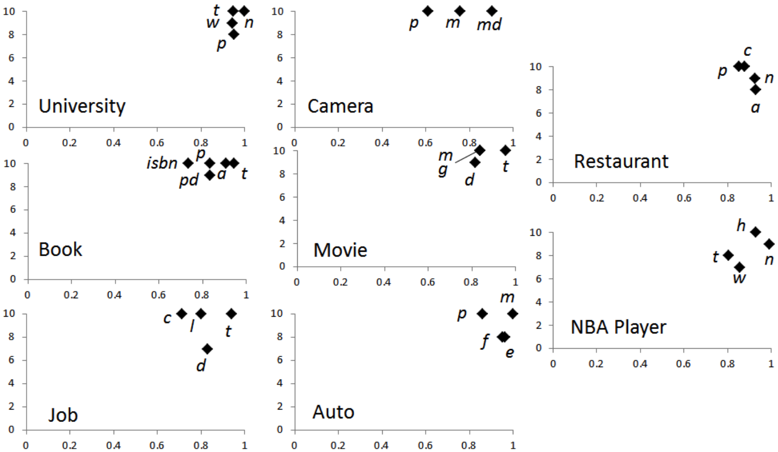
Vertical	Attribute	Topline	LD
University	phone	16973	283
	website	7968	12930
	name	9224	13144
	type	68	
Camera	model	5428	
	price	1524	
	manufacturer	253	
Book	isbn_13	19302	39112
	author	14228	13060
	title	17402	37485
	publication date	6645	3048
	publisher	6175	520
Movie	genre	1398	114
	title	17146	57292
	mpaa rating	3255	2
	director	7398	16079

Vertical	Attribute	Topline	LD
Job	title	17712	
	date posted	2381	
	location	5634	
	company	5655	
Auto	model	9916	
	price	10792	
	engine	2469	
	fuel economy	2051	
Restaurant	phone	19510	
	cuisine	2378	72
	address	29687	37
	name	16631	312
NBA player	weight	507	
	height	121	
	name	1457	9194
	team	60	677

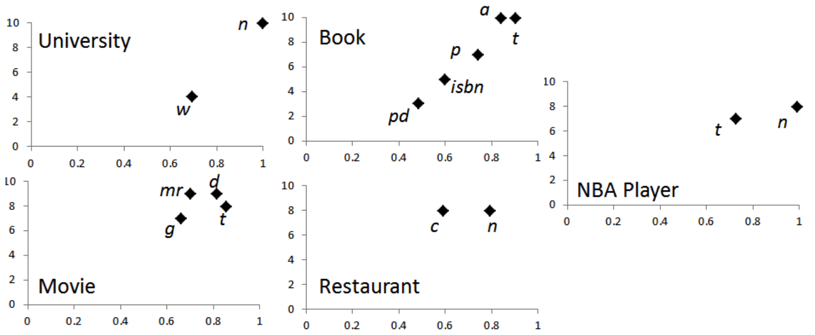
Results

- majority of cases the induced wrappers achieved very high accuracy
- number of cases where they failed
  - incorrect wrapper induced
  - failures often related to the nature of specific websites
  - proposed method is not suitable for all situations

Topline results



Linked Data results



Overall results

Concept	**Hao**	***Topline***	***LD***
auto	0.71	0.94	
book	0.87	0.85	0.78
camera	0.91	0.76	
job	0.85	0.82	
movie	0.79	0.86	0.76
nbaplayer	0.82	0.9	0.87
restaurant	0.96	0.89	0.69
university	0.83	0.96	0.91

Organisations, Information and Knowledge

The University Of Sheffield.

Outline

1. Wrapper Induction: Task definition
2. Proposed Methodology
3. Dataset
4. Experiment
5. Conclusions

Recap

- is Linked Data suitable Knowledge source for Web Scale Information Extraction?
  - investigation on Wrapper Induction task
- Contributions
  - study on the suitability of *Linked Data* to build dictionaries
  - good results overall for Wrapper Induction
  - some failure cases

Recap

- Simple idea
  - generate knowledge resources from Linked Data, in the form of dictionaries
  - use the dictionaries to annotate websites
  - look for recurrent patterns
- Advantages
  - no training material required
  - dictionaries are reusable across all websites of a pertinent domain
  - adaption across domains and websites with little human effort
- Limitations
  - not all concepts are covered by *Linked Data*
  - not all concepts are easy to locate in the *Linked Data*
  - lack of robustness in the learnt wrappers
    - irregular structure of the website
    - quality of the dictionary

Organisations, Information and Knowledge

The University Of Sheffield.

Interesting future directions

- investigation on the quality of dictionaries
  - is the dictionary sufficiently large for a task?
  - distributional features of the dictionary
  - compatibility between dictionary and the set of answers

Organisations, Information and Knowledge

The University Of Sheffield.

Further reading I

Álvarez, M., Pan, A., Raposo, J., Bellas, F., and Cacheda, F. (2008).
Finding and Extracting Data Records from Web Pages.
Journal of Signal Processing Systems, 59(1):123–137.

Carlson, A. and Schafer, C. (2008).
Bootstrapping information extraction from semi-structured web pages.
e European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases.

Dalvi, N., Bohannon, P., and Sha, F. (2009).
Robust web extraction: an approach based on a probabilistic tree-edit model.
Proceedings of the 35th SIGMOD international conference on Management of data.

Dalvi, N., Kumar, R., and Soliman, M. (2011).
Automatic wrappers for large scale web extraction.
Proceedings of the VLDB Endowment, 4(4):219–230.

Grigalis, T. (2013).
Towards web-scale structured web data extraction.
Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13, page 753.

Gulhane, P., Madaan, A., Mehta, R., Ramamirtham, J., Rastogi, R., Satpal, S., Sengamedu, S. H., Tengli, A., and Tiwari, C. (2011).
Web-scale information extraction with vertex.
2011 IEEE 27th International Conference on Data Engineering, pages 1209–1220.

Organisations, Information and Knowledge

The University Of Sheffield.

Further reading II

Hao, Q., Cai, R., Pang, Y., and Zhang, L. (2011).
From One Tree to a Forest : a Unified Solution for Structured Web Data Extraction.
In *SIGIR 2011*, pages 775–784.

Kushmerick, N. (1997).
Wrapper Induction for information Extraction.
In *IJCAI97*, pages 729–735.

Muslea, I., Minton, S., and Knoblock, C. (2003).
Active Learning with Strong and Weak Views : A Case Study on Wrapper Induction.
IJCAI'03 8th international joint conference on Artificial intelligence, pages 415–420.

Wong, T. and Lam, W. (2010).
Learning to adapt web information extraction knowledge and discovering new attributes via a Bayesian approach.
Knowledge and Data Engineering, IEEE, 22(4):523–536.

Zhai, Y. and Liu, B. (2005).
Web data extraction based on partial tree alignment.
. . . the 14th international conference on World Wide Web, pages 76–85.

Zhao, H., Meng, W., Wu, Z., Raghavan, V., and Yu, C. (2005).
Fully automatic wrapper generation for search engines.
Proceedings of the 14th international conference on World Wide Web - WWW '05, page 66.