Web Scale Information Extraction
TUTORIAL @ ECML/PKDD 2013

Anna Lisa Gentile   Ziqi Zhang

Department of Computer Science, The University of Sheffield, UK

27^{th} September 2013

Outline

1. Overview
2. Wrapper Induction
3. Table Interpretation
4. Conclusions

Web IE - Motivation

Data on the Web

- Very large scale
  - Unlimited domains
  - Unlimited documents
- Structured and unstructured

A promising route towards Tim Berners-Lee's vision of Semantic Web [Downey and Bhagavatula, 2013]

Web IE - Challenges

- Very large scale
- Coverage and quality
- Heterogeneity

Web IE - Challenges

Very large scale

- Web documents
  - ClueWeb09 - 1 billion web pages, 500 million English[a]
  - ClueWeb12 - 870 million English web pages[b]
- Knowledge Base (KB) examples
  - Linked Data
    - The Billion Triple Challenge (BTC) dataset - 1.4 billion facts
  - NELL [Carlson et al., 2010] KB - 3 million facts
  - Freebase KB - 40 million topics, 1.9 billion facts

[a] ttp://lemurproject.org/clueweb09/
[b] ttp://lemurproject.org/clueweb12/

Web IE - Challenges

Very large scale

- Web documents
  - ClueWeb09 - 1 billion web pages, 500 million English[a]
  - ClueWeb12 - 870 million English web pages[b]
- Knowledge Base (KB) examples
  - Linked Data
    - The Billion Triple Challenge (BTC) dataset - 1.4 billion facts
  - NELL [Carlson et al., 2010] KB - 3 million facts
  - Freebase KB - 40 million topics, 1.9 billion facts

Requiring efficient algorithms and evaluation methods

[a] ttp://lemurproject.org/clueweb09/
[b] ttp://lemurproject.org/clueweb12/

Web IE - Challenges

Coverage and quality

- Data redundancy - a crucial assumption behind typical Web IE methods
- Long tail can be equally interesting and important [Dalvi et al., 2012]
- Web pages contain substantial noise
  - E.g., only 1.1% of tables on the Web contain useful relational data [Cafarella et al., 2008]
- KBs are not perfect
  - E.g., DBpedia has erroneous facts [Gentile et al., 2013]

Web IE - Challenges

Coverage and quality

- Data redundancy - a crucial assumption behind typical Web IE methods
- Long tail can be equally interesting and important [Dalvi et al., 2012]
- Web pages contain substantial noise
  - E.g., only 1.1% of tables on the Web contain useful relational data [Cafarella et al., 2008]
- KBs are not perfect
  - E.g., DBpedia has erroneous facts [Gentile et al., 2013]

Requiring noise tolerant methods with reasonable coverage

Web IE - Challenges

Heterogeneity

- Natural language is highly expressive
  - Data redundancy may not be transparent
- Heterogeneity across KBs
  - E.g., [Wijaya et al., 2013]
- Heterogeneity inside KBs
  - E.g., [Zhang et al., 2013]

Web IE - Challenges

Heterogeneity

- Natural language is highly expressive
  - Data redundancy may not be transparent
- Heterogeneity across KBs
  - E.g., [Wijaya et al., 2013]
- Heterogeneity inside KBs
  - E.g., [Zhang et al., 2013]

YAGO
Dbpedia
Freebase
Entitycube
NELL
DeepDive
Probase
KnowItAll / ReVerb

PATTY
BabelNet
WikiNet
ConceptNet
WordNet

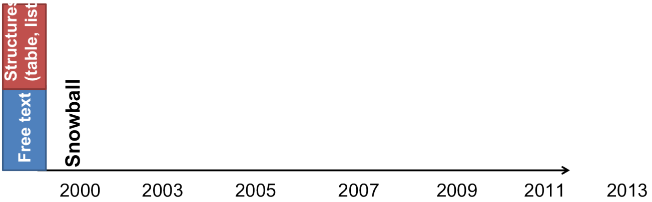
Web IE Methods and Systems

- 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]

Web IE Methods and Systems

Snowball [Agichtein and Gravano, 2000]

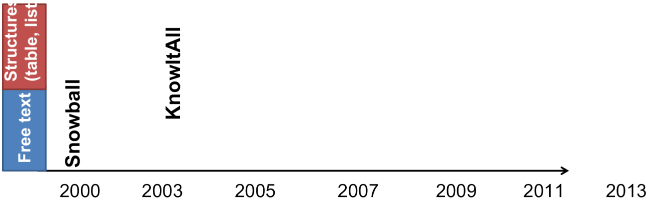
- Learn syntactic patterns to extract relation instances
  - 5-tuple <ordinal, url, left, middle, right>
  - selectiveness (Named Entity Recognition)
- Bootstrap with seed + domain specific heuristics
  - use case: <Organisation, Location> pairs



Web IE Methods and Systems

KnowItAll [Etzioni et al., 2004]

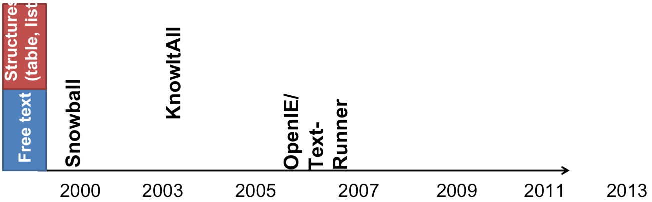
- Input: ontology and rule templates (Hearst patterns)
  - generic syntactic patterns (e.g., cities such as <?>) to extract relation/class instances
- statistically generated "discriminator phrases" for each class
- bootstrap with 2 names for each class



Web IE Methods and Systems

OpenIE/TextRunner [Banko et al., 2007]

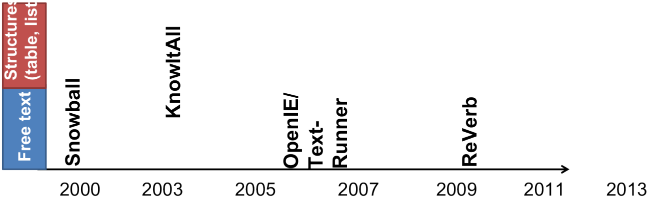
- Learn syntactic patterns to extract any relation instances from any domains (open IE)
- Completely unsupervised, no need for seeds
  - Input: corpus C, parser phase on a portion of C
  - pattern generation from parsed documents, t: <e_1, r, e_2>
  - feature vector generation and Naive Bayes learning



Web IE Methods and Systems

ReVerb [Fader et al., 2011]

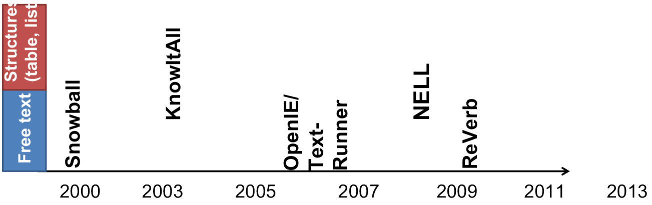
- Open IE improved:
  - syntactic and lexical constraints to select the target verbs
  - Input: POS-tagged and NP-chuncked corpus



Web IE Methods and Systems

NELL [Carlson et al., 2010]

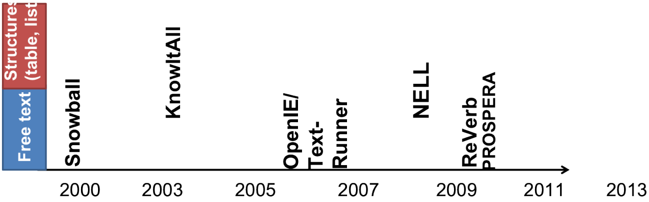
- Multi-strategy, “coupled” learning
  - enforcing constraints and/or strengthening evidences
- Bootstrap with seed ontology (class, relation, instance) + coupling rules



Web IE Methods and Systems

PROSPERA [Nakashole et al., 2011]

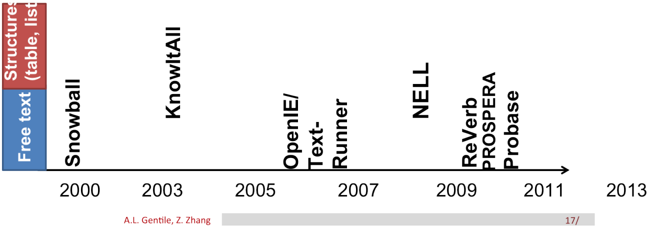
- N-gram item-set patterns to generalise narrow syntactic patterns to boost recall
- Reasoning with large KB (YAGO) to constrain extractions to boost precision
  - Input: target relations and type signature for involved entities
- Integration with KB (data reconciliation)



Web IE Methods and Systems

Probase [Wu et al., 2012]

- Building a probabilistic concept taxonomy
  - First iteration - bootstrap with syntactic patterns
  - Following iterations - previously learnt knowledge used to semantically constrain new extractions



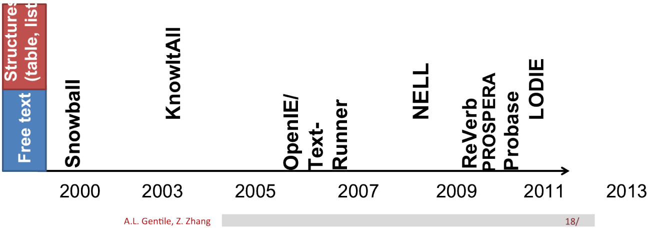
Organisations, Information and Knowledge

The University Of Sheffield.

Web IE Methods and Systems

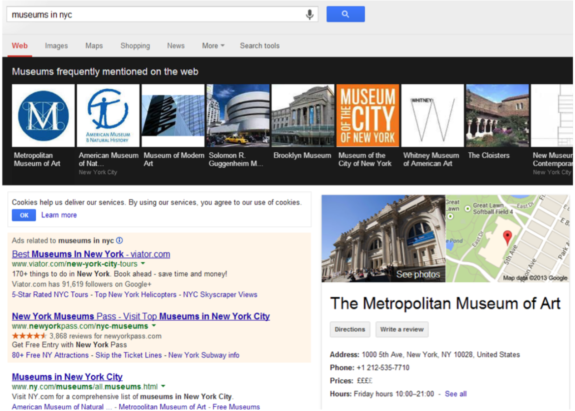
LODIE [Ciravegna et al., 2012]

- Multi-strategy learning
  - structured webpages, tables and lists, free text
- Focuses on using Linked Data to seed learning



Web IE Systems behind the Giants

Google Knowledge Graph



Web IE Systems behind the Giants

IBM Watson QA

Web IE - This tutorial

IS NOT about

- Any systems or their methodologies introduced in the previous slides
    - read corresponding publications
- Large scale KBs or KBs generated by previously introduced systems
    - see tutorial by [Suchanek and Weikum, 2013]

Instead

- Focus on “structured” data on the Web
    - Wrapping entity centric pages
    - Interpreting tables

Content of this tutorial

Entity centric structured web pages

- Regular, script generated Web pages containing entities of specific domains
  - high connectivity and redundancy of structured data on the Web [Dalvi et al., 2012]
- Great potential to extract high quality information

Content of this tutorial

Tables

- A widely used structure for relational information
  - Hundreds of millions of high quality, “useful” tables [Cafarella et al., 2008]
- Great potential to improve search quality
  - Tabular data complements free text
  - High demand for tabular data as seen by search engines

Content of this tutorial

Outline

Web IE methods

- Entity centric web pages
  - Wrapper Induction
- Tables
  - Table Interpretation
- Conclusions

Outline

1. Overview
2. Wrapper Induction
3. Table Interpretation
4. 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

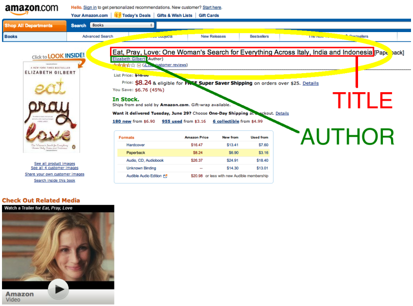
Web Scale Wrapper Induction

- 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]

Wrapper Induction: example

Extracting book attributes on e-commerce websites



Extraction lifecycle

1. Clustering pages within a Web site
2. Learning extraction rules
3. Detecting site structure changes
4. Re-learning broken rules

} Robust methods

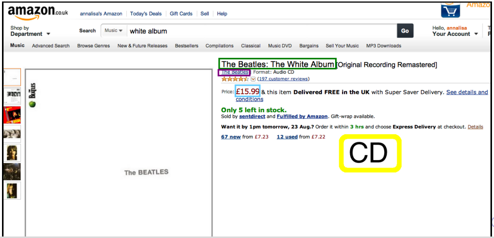
[Gulhane et al., 2011]

Organisations, Information and Knowledge

The University Of Sheffield.

Website Clustering Problem

Given a website, cluster the pages so that the pages generated by the same script are in the same cluster [Blanco et al., 2011]



Website Clustering Problem

Clustering approaches:

- URL, tag probability and tag periodicity features, using MiniMax algorithm [Crescenzi et al., 2002]
- XProj - XML clustering, linear complexity
  - $\sim$ 20 hours for a site with a million pages) [Aggarwal and Wang, 2007]
- ClustVX - XString representation of Web pages, which encapsulate tag paths and visual features [Grigalis, 2013]
- Shingle-signature [Gulhane et al., 2011]
- URLs of the webpages, simple content and structural features [Blanco et al., 2011]

Further reading survey [Gottron, 2008]

Scalable Clustering

Structurally clustering webpages for extraction

- URLs of the webpages
- simple content and structural features
- pair-wise similarity of URLs/documents is not meaningful [Aggarwal and Wang, 2007]
- look at URLs holistically and look at the patterns that emerge
- linear time complexity (700,000 pages in 26 seconds)

[Blanco et al., 2011]

Scalable Clustering

IDEA: use a simple encoding of each page which summarise URL and content features

- **URL** a sequence of tokens, delimited by /
- **URL pattern** is a sequence of tokens, with a special token *
  - The number of * is called the arity of the url pattern (e.g. www.2spaghi.it/ristoranti/*/*/*/*)
- $S = \{S_1, S_2, \ldots S_k\}$ be a set of **Scripts**
  - S_i a pair (p_i,D_i), with p_i a URL pattern, and D_i a database with same arity as p_i.

[Blanco et al., 2011]

Scalable Clustering

- W set of Web pages
- $T(w)$ set of terms for each $w \in W$
  - e.g. URL sequence "site.com/a1/a2/..." represented as a set of terms $\{(pos_1 = site.com), (pos_2 = a_1), (pos_3 = a_2), \ldots\}$
- $W(t)$ set of Web pages containing t
- $C = \{W_1, \ldots, W_k\}$ a clustering of W

Principle of Minimum Description Length (MDL)

Given a set of urls U, find the set of scripts S that best explain U. Find the shortest hypothesis, i.e. S that minimize the description length of U

[Blanco 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

Supervised methods

Training data

- manually labelled training examples, with significant human effort
  - use reduced number of annotations (minimum 1 per website)
  - crowdsource the annotations

Web site specific

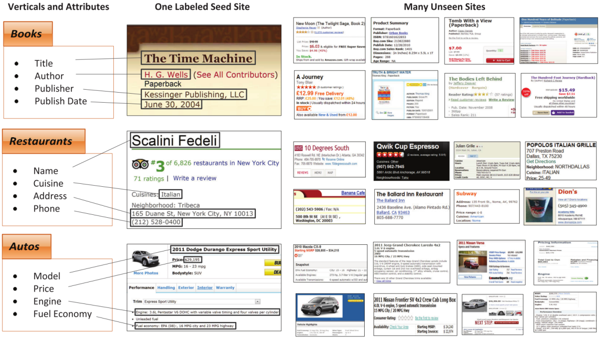
- learn a wrapper per each Web site
- assumption: structural consistency of the Web site
- porting wrappers across Web sites often require re-learnining [Wong and Lam, 2010]
- Web site change can cause wrappers to break
  - more training data required to enhance wrapper robustness [Carlson and Schafer, 2008, Dalvi et al., 2009, Dalvi et al., 2011, Hao et al., 2011]

Supervised methods in this tutorial

- Multi-view learner [Hao et al., 2011]
- Vertex! [Gulhane et al., 2011]

Supervised methods: Multi-view learners

- handle any vertical without re-implementation
- only requires one labeled example site from each vertical



[Hao et al., 2011]

Supervised methods: Multi-view learners

Based on the idea of having **strong** and **weak** features to train the wrappers

- Weak features
  - general across attributes, verticals and websites
  - identify a large amount of candidate attribute values
    - likely to contain noise
- Strong features
  - site-specifc
  - derived in an unsupervised manner

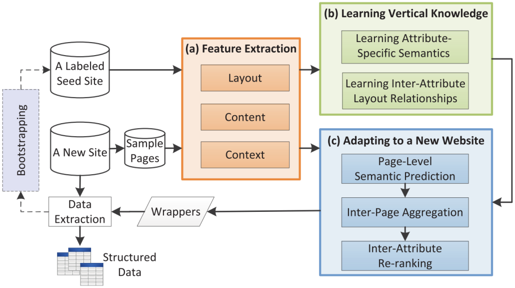
Characteristics

- improve robustness
- reduce the amount of manual annotations
- still require seed Web pages to be annotated (at least one website for each vertical)

[Hao et al., 2011]

Supervised methods: Multi-view learners

- (a) feature extraction
- (b) learning vertical knowledge
- (c) adapting to a new website

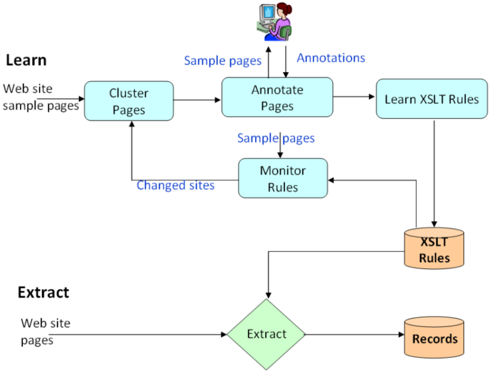


[Hao et al., 2011]

Supervised methods: Vertex!

Complete Wrapper lyfecycle

- clustering in 3 passes over the data
- greedy algorithm to pick pages to annotate
- Apriori style algorithm to learn rules
- site detection scheme
- optimisation of rule re-writing



[Gulhane et al., 2011]

Vertex! Learning rules

- $\boldsymbol{X_i}$, XPath
- $\boldsymbol{F(X_i)}$, frequency of X_i
- differentiate between informative and noisy XPaths
  - noisy sections share common structure and content
  - informative sections differ in their actual content
- $\boldsymbol{I(X_i)}$, informativeness of X_i

$$I(X_i) = 1 - \frac{\sum_{t \in T_i} F(X_i, t)}{M \cdot |T_i|} \tag{1}$$

  - T_i, set of content
  - $F(X_i, t)$ numb. pages containing content t, in nodes matched by X_i
  - M, total number of pages
- $\boldsymbol{w(X_i)} = F(X_i) \cdot I(X_i)$

[Gulhane et al., 2011]

Unsupervised methods

- Do not require training data

BUT

- do not recognise the semantics of the extracted data (i.e., attributes)
- rely on human effort as post-process to identify attribute values from the extracted content

Unsupervised methods in this tutorial

- RoadRunner [Crescenzi and Mecca, 2004]
- Yahoo! [Dalvi et al., 2011]
- SKES [He et al., 2013]
- LODIE Wrapper Induction [Gentile et al., 2013]

Unsupervised methods: RoadRunner

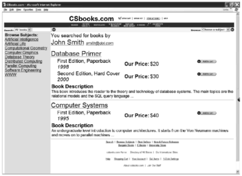
Schema finding problem

a. Source Dataset

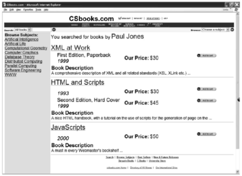
Name	Email	Books				
		Title	Descr.	Editions		
				Details	Year	Price
John Smith	smith@..	DB Primer	This book..	1st Ed., P.back	1998	20$
				2nd Ed., H. Cover	2000	30$
		Computer S.	An underg...	1st Ed., P.back	1995	40$
Paul Jones	*null*	XML at..	A compr..	1st Ed., P.back	1999	30$
		HTML..	A useful..	*null*	1993	30$
				2nd Ed., H. Cover	1999	45$
		JavaScript	A must in..	*null*	2000	50$
...	...	...	...	...	...	...

b. HTML Pages

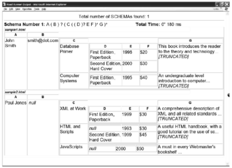
http://www.csbooks.com/author?Smith



http://www.csbooks.com/author?Jones



c. Data Extraction Output



[Crescenzi and Mecca, 2004]

Unsupervised methods: RoadRunner

- Definition of a class of regular languages: prefix mark-up languages
- Grammar inference with polynomial-time unsupervised learning algorithm
- no a priori knowledge about the target pages and their contents

[Crescenzi and Mecca, 2004]

Unsupervised methods: Yahoo!

- Objective: make wrapper induction noise-tolerant
- **Unsupervised** learning
  - automatically and cheaply obtained noisy training data (e.g. precompiled dictionaries, regular expressions...)
  - domain specific knowledge
- **Enumerate** all possible wrappers efficiently
  - Generate the *wrapper space* for a set of labels
  - bottom-up/top-down
- **Rank** wrappers in the space
  - probabilistic evaluation of
    - goodness of the annotators
    - good structure of the webpage

[Dalvi et al., 2011]

Unsupervised methods: SKES

- Cluster Web pages
- Represent each detail page as a collection of *tag paths*
- Wrapper extraction
    - Template induction
    - Structured data extraction
    - Data post-processing

[He et al., 2013]

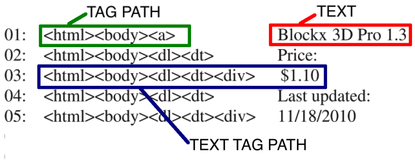
SKES - Page Representation

Definitions (from [Zhao et al., 2005]):

- *tag tree*: tree representation of a Web page, based on the tags in its source HTML
- *tag nodes*: root tag and internal nodes of the tree
- *tag path*: path to reach a specific tag node starting from the root

Page representation:

- content of all *text nodes* in the page
- their corresponding *tag paths*
- *text tag paths*: concatenation of a tag path and its carried text content



[He et al., 2013]

SKES - Page Representation example

TAG TREE

TAG NODES

```
<html><body>
<a>
   Archipelago 1.14
</a>
<dl>
   <dt>Price:</dt>
      <div>$2.99</div>
   <dt>Last updated:</dt>
      <div>09/26/2010</div>
</dl>
<a>
   Recommendations
</a>
<ul>
   <li>Par 72 Golf</li>
   <li>Mathpac 5.6</li>
   <li>FourNumGuess 1.0.6</li>
</ul>
</html></body>
```

SKES PAGE REPRESENTATION

TAG PATH

01:	<html><body><a>	Archipelago 1.14
02:	<html><body><dl><dt>	Price:
03:	<html><body><dl><dt><div>	$2.99
04:	<html><body><dl><dt>	Last updated:
05:	<html><body><dl><dt><div>	09/26/2010
06:	<html><body><a>	Recommendations
07:	<html><body>	Par 72 Golf
08:	<html><body>	Mathpac 5.6
09:	<html><body>	FourNumGuess 1.0.6

TEXT TAG PATH

[He et al., 2013]

SKES - wrapper extraction

INPUT: a set of HTML pages and a support threshold
OUTPUT: induced template
IDEA: counting the support of:

- tag paths
  - checking the presence of the tag path on the set of pages
- text tag paths
  - repetitive text tag paths are likely to be attributes indicators
  - unique text tag paths are likely to be data region

The University Of Sheffield.

Organisations, Information and Knowledge

SKES - pros and cons

- pros
  - completely unsupervised
  - no knowledge required (in the form of a schema)
- cons
  - no semantics for the attributes

[He et al., 2013]

Unsupervised methods: LODIE Wrapper Induction

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

[Gentile et al., 2013]

LODIE Wrapper Induction: 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

[Gentile et al., 2013]

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: 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]

LODIE Wrapper Induction: 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')). }
```

[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]

LODIE Wrapper Induction: XPath identification

- find the distinct *xpath* in the set of annotation pairs $< xpath, value_{i,k} >$ for attribute $a_{i,n}$
- create a mapping between *xpath* and the set of distinct values matched by that *xpath* across the entire website collection
  - an entry in the map is a pair $< xpath, value_{i,k} | k = n >$ where n denotes the attribute of interest is $a_{i,n}$
- hypothesis
  - an attribute is likely to have various distinct values
  - the top ranked $< xpath, value_{i,k} | k = n >$ pairs by the size of $value_{i,k} | k = n$ are likely to be useful XPaths for extracting the attribute $a_{i,n}$ on the website collection

[Gentile et al., 2013]

LODIE - pros and cons

- pros
  - unsupervised
  - information need driven approach
    - search space for the wrappers is limited to possibly relevant portions of pages
- cons
  - user need definition still manual
    - BUT concept and attribute semantic is defined only once and valid to all websites of same domain

[Gentile et al., 2013]

Robust methods

- wrappers learnt without robustness considerations have short life (average of 2 months [Gulhane et al., 2011])
- probabilistic model to capture how Web pages evolve over time ([Dalvi et al., 2009])
    - trained using a collection of evolutions of Web pages
    - encoding the probability of each editing operation on a Web page over time
    - computing the probability of a Web page evolving from one state to another, by aggregating the probabilities of each edit operation
- “robustness” of wrappers is evaluated using the learnt probabilistic model

Outline

Table Interpretation

Table Interpretation

Table Interpretation – outline

- **Motivation**
- Problem definition
- Table Interpretation - Methods

Table Interpretation – Motivation

Specification criteria	Specification
Review details	
Test Date	June 2013
Reviewed using 2013 test programme	Yes
Design	
Smartphone	Yes
Phone style	Candy Bar
Height (mm)	120.0
Width (mm)	67.0
Depth (mm)	10.0
Weight (g)	139
Qwerty keyboard	Yes
Touchscreen	Yes
Display size (inches)	3.1

Blackberry Q10
Launch date: Apr 2013
Which? score: S

POS		LP	CLUB	P	W	D	L	GF	GA	GD	PTS
0	–	(0)	Arsenal	0	0	0	0	0	0	0	0
0	–	(0)	Aston Villa	0	0	0	0	0	0	0	0
0	–	(0)	Cardiff City	0	0	0	0	0	0	0	0
0	–	(0)	Chelsea	0	0	0	0	0	0	0	0
0	–	(0)	Crystal Palace	0	0	0	0	0	0	0	0
0	–	(0)	Everton	0	0	0	0	0	0	0	0
0	–	(0)	Fulham	0	0	0	0	0	0	0	0

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

Figure 1: Tables feature properties that favour the extraction of information.

TITLE	DESCRIPTION	ADRESS
Kankouji Temple	The temple has approximately 600 years history, an...	Uwano 267, Minamiuonuma-shi...
Untoan Temple	Untoan is a temple of the Soto school of Zen Buddh...	660 Unto, Minamiuonuma, Niigata...
Bishamon Temple	Fukoji is a "designated cultural asset" by the cit...	Bishamon-do, Urasa Fukoji Temple grounds ...
GOUZOKU PALACE RYUGON	RYUGON IS A TRADITIONAL JAPANESE STYLE HOTEL...	79, SAKADO, MINAMIUONUMA-SHI...
HOTEL KOJYOKAN	Welcome to the Kojyokan. Our hotel is located in t...	1873, ISHIUCHI, MINAMIUONUMA-SI...
LODGE MASHU	GET TO THE ISHIUCHI MARUYAMA SKI TRAIL JUST...	Go to Koide using route 17. Pass the bowlings...
Azumaken Restaurant	Hmmm... 	Get on route 291 towards Koide. Turn left at the S...
Café West Restaurant	Once a year, it can't hurt <img src="images\...	Drive on route 17 towards Koide. The place is on t...
Early's Restaurant	Oh gee, these burgers! <img src="images\pic5....	2035-2, ISHIUCHI, MINAMIUONUMA-SHI, NIIGATA

A widely used structure for (relational) information

Geometric tolerances							
Rotation (QD12-06-C)	mm	0.005	0.008	0.008	0.01	0.01	0.015
Flatness (QD12-07-C)	mm	0.008	0.008	0.008	0.012	0.012	0.015
Parallelism (QD12-08-C)	mm	0.013	0.015	0.015	0.02	0.02	0.03
Radial Runout (QD12-10-C)	mm	0.001	0.001	0.001	0.001	0.001	0.001
Axial Runout (QD12-11-C)	mm	0.001	0.001	0.001	0.001	0.001	0.001
Coning of Table Axis (QD12-35-B)	Arc Seconds	+/-0.5"	+/-0.5"	+/-0.5"	+/-0.5"	+/-0.5"	+/-0.5"
Concentricity of Centre Bore (QD12-36-B)	mm	0.005	0.005	0.005	0.005	0.005	0.005

Table 1. Classification results on 81 symbolic columns

manual \ computed	our method using the ontology				SMO			
	Food	Micro.	Resp.	Other	Food	Micro.	Resp.	Other
Food	34	0	0	12	45	0	0	1
Micro.	0	16	0	0	4	12	0	0
Response	0	0	1	0	0	0	0	1
Other	3	3	0	12	7	0	0	11

Technique	Major urban (< $10 million) Value often	Major urban (< $10 million) Value occasionally	Major urban (> $10 million) Value often	Major urban (> $10 million) Value occasionally
Adjusted yield approach	4.8% (1)	23.8% (10)	17.0% (9)	43.3% (16)
Cash flow (term and reversion approach)	38.1% (8)	19.1% (8)	22.6% (12)	18.9% (7)
Shortfall approach	52.4% (11)	47.6% (20)	56.6% (30)	27.0% (10)
Layer approach	4.7% (1)	9.5% (4)	3.8% (2)	10.8% (4)
Test of independence: chi-square	5.374 with DF 3 (dependence not proved)		11.46 with D.F.3 (highly dependent)	

Note: The figures in brackets are the actual number of responses in each category. All chi-square tests are based on numbers and not percentages.

Table Interpretation – Motivation

English only:
14.1 billion raw HTML tables
154 million tables containing relational data
[Cafarella et al., 2008a]

A **widely used** structure
for (relational)
information

Table Interpretation – Motivation

Structural regularity => semantic consistency
Simplifies interpretation of data and information extraction

A widely used **structure** for (relational) information

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

Table Interpretation – Motivation

Table structures embed important relational information – recall relational databases (RDBMs)

A widely used structure for (**relational**) information

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

Table Interpretation – Motivation

Enormous data source ....

... contains extractable, interpretable ...

A **widely used structure** for (**relational**) information

... semantically useful information that can be linked to/ complement KBs

Table Interpretation – Motivation

Tables often contain data that are unlikely to be found in a text [Quercini and Reynaud, 2013]

Great potential to improve search quality

Google: 30 millions queries/day lead to webpages containing tables with relational data
[Cafarrela et al., 2008a]

Table Interpretation

- Motivation
- **Problem definition**
- Table Interpretation - Methods

Table Interpretation – Definition

Input tables

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

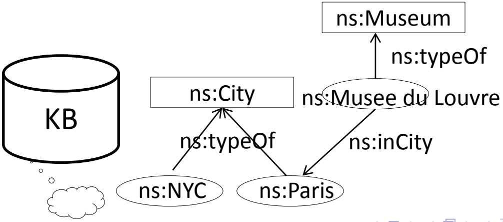


Table Interpretation – Definition

Input tables

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

Goal: link

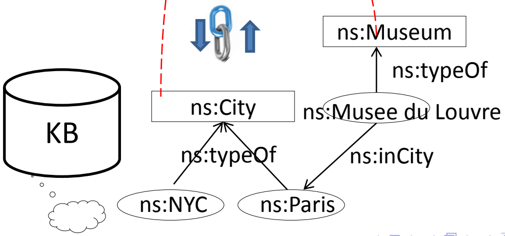


Table Interpretation – Definition

Input tables

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

Goal: link

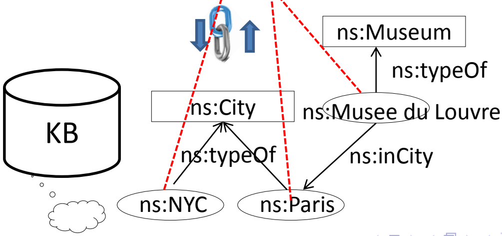


Table Interpretation – Definition

Input

tables

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

Goal: link

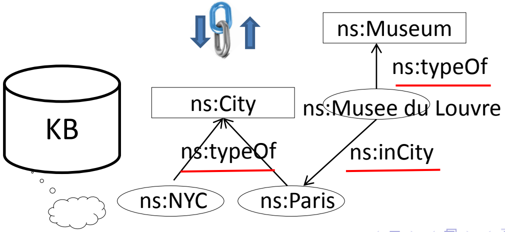


Table Interpretation – Definition

Input tables

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

Goal: enrich

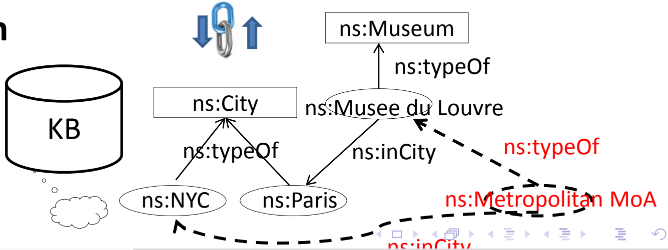


Table Interpretation – Definition

In words

Given an input table and a KB that defines semantic concepts, relations and instances, we require Table Interpretation to perform three types of annotations

- Semantic concept/class/type
- Entity instance
- Relation

Also enrich the KB with new concepts and entities

Table Interpretation – Challenges

Why is this a challenging task?

- Noise and diversity
- A multi-task problem
- Scalability

Table Interpretation – Challenges

Noise and diversity

Table Interpretation – Challenges

Noise and diversity

- Out of 14.1 billion HTML tables , only 1.1% of the raw HTML tables are true relations [Cafarrela et al., 2008a]
- Relational tables
  - tables containing relational data

Table Interpretation – Challenges

Noise and <u>diversity</u>

Table Interpretation – Challenges

Specification criteria	Specification
Review details	
Test Date	June 2013
Reviewed using 2013 test programme	Yes
Design	
Smartphone	Yes
Phone style	Candy Bar
Height (mm)	120.0
Width (mm)	67.0
Depth (mm)	10.0
Weight (g)	139
Qwerty keyboard	Yes
Touchscreen	Yes
Display size (inches)	3.1

Blackberry Q10
Launch date: Apr 2013
Which? score: S

POS		LP	CLUB	P	W	D	L	GF	GA	GD	PTS
0	–	(0)	Arsenal	0	0	0	0	0	0	0	0
0	–	(0)	Aston Villa	0	0	0	0	0	0	0	0
0	–	(0)	Cardiff City	0	0	0	0	0	0	0	0
0	–	(0)	Chelsea	0	0	0	0	0	0	0	0
0	–	(0)	Crystal Palace	0	0	0	0	0	0	0	0
0	–	(0)	Everton	0	0	0	0	0	0	0	0
0	–	(0)	Fulham	0	0	0	0	0	0	0	0

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

Figure 1: Tables feature properties that favour the extraction of information.

TITLE	DESCRIPTION	ADRESS
Kankouji Temple	The temple has approximately 600 years history, an...	Uwano 267, Minamiuonuma-shi...
Untoan Temple	Untoan is a temple of the Soto school of Zen Buddh...	660 Unto, Minamiuonuma, Niigata...
Bishamon Temple	Fukoji is a "designated cultural asset" by the cit...	Bishamon-do, Urasa Fukoji Temple grounds ...
GOUZOKU PALACE RYUGON	RYUGON IS A TRADITIONAL JAPANESE STYLE HOTEL...	79, SAKADO, MINAMIUONUMA-SHI...
HOTEL KOJYOKAN	Welcome to the Kojyokan. Our hotel is located in t...	1873, ISHIUCHI, MINAMIUONUMA-SI...
LODGE MASHU	GET TO THE ISHIUCHI MARUYAMA SKI TRAIL JUST...	Go to Koide using route 17. Pass the bowlings...
Azumaken Restaurant	Hmmm... 	Get on route 291 towards Koide. Turn left at the S...
Café West Restaurant	Once a year, it can't hurt <img src="images\...	Drive on route 17 towards Koide. The place is on t...
Early's Restaurant	Oh gee, these burgers! <img src="images\pic5....	2035-2, ISHIUCHI, MINAMIUONUMA-SHI, NIIGATA

A recap of example "relational" tables on the web...

Geometric tolerances							
Rotation (QD12-06-C)	mm	0.005	0.008	0.008	0.01	0.01	0.015
Flatness (QD12-07-C)	mm	0.008	0.008	0.008	0.012	0.012	0.015
Parallelism (QD12-08-C)	mm	0.013	0.015	0.015	0.02	0.02	0.03
Radial Runout (QD12-10-C)	mm	0.001	0.001	0.001	0.001	0.001	0.001
Axial Runout (QD12-11-C)	mm	0.001	0.001	0.001	0.001	0.001	0.001
Coning of Table Axis (QD12-35-B)	Arc Seconds	+/-0.5"	+/-0.5"	+/-0.5"	+/-0.5"	+/-0.5"	+/-0.5"
Concentricity of Centre Bore (QD12-36-B)	mm	0.005	0.005	0.005	0.005	0.005	0.005

Table 1. Classification results on 81 symbolic columns

manual \ computed	our method using the ontology				SMO			
	Food	Micro.	Resp.	Other	Food	Micro.	Resp.	Other
Food	34	0	0	12	45	0	0	1
Micro.	0	16	0	0	4	12	0	0
Response	0	0	1	0	0	0	0	1
Other	3	3	0	12	7	0	0	11

Technique	Major urban (< $10 million) Value often	Major urban (< $10 million) Value occasionally	Major urban (> $10 million) Value often	Major urban (> $10 million) Value occasionally
Adjusted yield approach	4.8% (1)	23.8% (10)	17.0% (9)	43.3% (16)
Cash flow (term and reversion approach)	38.1% (8)	19.1% (8)	22.6% (12)	18.9% (7)
Shortfall approach	52.4% (11)	47.6% (20)	56.6% (30)	27.0% (10)
Layer approach	4.7% (1)	9.5% (4)	3.8% (2)	10.8% (4)
Test of independence: chi-square	5.374 with DF 3 (dependence not proved)		11.46 with D.F.3 (highly dependent)	

Note: The figures in brackets are the actual number of responses in each category. All chi-square tests are based on numbers and not percentages.

Table Interpretation – Challenges

Noise and diversity

Very difficult to generalise one universal solution

Specification criteria	Specification
Review details	
Test Date	June 2013
Reviewed using 2013 test programme	Yes
Design	
Smartphone	Yes
Phone style	Candy Bar
Height (mm)	120.0
Width (mm)	67.0
Depth (mm)	10.0
Weight (g)	139
Qwerty keyboard	Yes
Touchscreen	Yes
Display size (inches)	3.1

Blackberry Q10
Launch date: Apr 2013
Which? score: S

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

Figure 1: Tables feature properties that favour the extraction of information.

TITLE	DESCRIPTION	ADRESS
Kankouji Temple	The temple has approximately 600 years history, an...	Uwano 267, Minamiuonuma-shi...
Untoan Temple	Untoan is a temple of the Soto school of Zen Buddh...	660 Unto, Minamiuonuma, Niigata...
Bishamon Temple	Fukoji is a "designated cultural asset" by the cit...	Bishamon-do, Urasa Fukoji Temple grounds ...
GOUZOKU PALACE RYUGON	RYUGON IS A TRADITIONAL JAPANESE STYLE HOTEL...	79, SAKADO, MINAMIUONUMA-SHI...
HOTEL KOJYOKAN	Welcome to the Kojyokan. Our hotel is located in t...	1873, ISHIUCHI, MINAMIUONUMA-SI...
LODGE MASHU	GET TO THE ISHIUCHI MARUYAMA SKI TRAIL JUST...	Go to Koide using route 17. Pass the bowlings...
Azumaken Restaurant	Hmmm... 	Get on route 291 towards Koide. Turn left at the S...
Café West Restaurant	Once a year, it can't hurt <img src="images\...	Drive on route 17 towards Koide. The place is on t...
Early's Restaurant	Oh gee, these burgers! <img src="images\pic5....	2035-2, ISHIUCHI, MINAMIUONUMA-SHI, NIIGATA

	Geometric tolerances						
Rotation (QD12-06-C)	mm	0.005	0.008	0.008	0.01	0.01	0.015
Flatness (QD12-07-C)	mm	0.008	0.008	0.008	0.012	0.012	0.015
Parallelism (QD12-08-C)	mm	0.013	0.015	0.015	0.02	0.02	0.03
Radial Runout (QD12-10-C)	mm	0.001	0.001	0.001	0.001	0.001	0.001
Axial Runout (QD12-11-C)	mm	0.001	0.001	0.001	0.001	0.001	0.001
Coning of Table Axis (QD12-35-B)	Arc Seconds	+/-0.5"	+/-0.5"	+/-0.5"	+/-0.5"	+/-0.5"	+/-0.5"
Concentricity of Centre Bore (QD12-36-B)	mm	0.005	0.005	0.005	0.005	0.005	0.005

POS		LP	CLUB	P	W	D	L	GF	GA	GD	PTS
0	–	(0)	Arsenal	0	0	0	0	0	0	0	0
0	–	(0)	Aston Villa	0	0	0	0	0	0	0	0
0	–	(0)	Cardiff City	0	0	0	0	0	0	0	0
0	–	(0)	Chelsea	0	0	0	0	0	0	0	0
0	–	(0)	Crystal Palace	0	0	0	0	0	0	0	0
0	–	(0)	Everton	0	0	0	0	0	0	0	0
0	–	(0)	Fulham	0	0	0	0	0	0	0	0

Table 1. Classification results on 81 symbolic columns

computed / manual	our method using the ontology				SMO			
	Food	Micro.	Resp.	Other	Food	Micro.	Resp.	Other
Food	34	0	0	12	45	0	0	1
Micro.	0	16	0	0	4	12	0	0
Response	0	0	1	0	0	0	0	1
Other	3	3	0	12	7	0	0	11

	Major urban (< $10 million)		Major urban (> $10 million)	
Technique	Value often	Value occasionally	Value often	Value occasionally
Adjusted yield approach	4.8% (1)	23.8% (10)	17.0% (9)	43.3% (16)
Cash flow (term and reversion approach)	38.1% (8)	19.1% (8)	22.6% (12)	18.9% (7)
Shortfall approach	52.4% (11)	47.6% (20)	56.6% (30)	27.0% (10)
Layer approach	4.7% (1)	9.5% (4)	3.8% (2)	10.8% (4)
Test of independence: chi-square	5.374 with DF 3 (dependence not proved)		11.46 with D.F.3 (highly dependent)	

Note: The figures in brackets are the actual number of responses in each category. All chi-square tests are based on numbers and not percentages.

Table Interpretation – Challenges

Noise and **diversity**

Practically we narrow down the problem scope

Specification criteria	Specification
Review details	
Test Date	June 2013
Reviewed using 2013 test programme	Yes
Design	
Smartphone	Yes
Phone style	Candy Bar
Height (mm)	120.0
Width (mm)	67.0
Depth (mm)	10.0
Weight (g)	139
Qwerty keyboard	Yes
Touchscreen	Yes
Display size (inches)	3.1

Blackberry Q10
Launch date: Apr 2013
Which? score: S

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

Figure 1: Tables feature properties that favour the extraction of information.

TITLE	DESCRIPTION	ADRESS
Kankouji Temple	The temple has approximately 600 years history, an...	Uwano 267, Minamiuonuma-shi...
Untoan Temple	Untoan is a temple of the Soto school of Zen Buddh...	660 Unto, Minamiuonuma, Niigata...
Bishamon Temple	Fukoji is a "designated cultural asset" by the cit...	Bishamon-do, Urasa Fukoji Temple grounds ...
GOUZOKU PALACE RYUGON	RYUGON IS A TRADITIONAL JAPANESE STYLE HOTEL...	79, SAKADO, MINAMIUONUMA-SHI...
HOTEL KOJYOKAN	Welcome to the Kojyokan. Our hotel is located in t...	1873, ISHIUCHI, MINAMIUONUMA-SI...
LODGE MASHU	GET TO THE ISHIUCHI MARUYAMA SKI TRAIL JUST...	Go to Koide using route 17. Pass the bowlings...
Azumaken Restaurant	Hmmm... 	Get on route 291 towards Koide. Turn left at the S...
Café West Restaurant	Once a year, it can't hurt <img src="images\...	Drive on route 17 towards Koide. The place is on t...
Early's Restaurant	Oh gee, these burgers! <img src="images\pic5....	2035-2, ISHIUCHI, MINAMIUONUMA-SHI, NIIGATA

	Geometric tolerances						
Rotation (QD12-06-C)	mm	0.005	0.008	0.008	0.01	0.01	0.015
Flatness (QD12-07-C)	mm	0.008	0.008	0.008	0.012	0.012	0.015
Parallelism (QD12-08-C)	mm	0.013	0.015	0.015	0.02	0.02	0.03
Radial Runout (QD12-10-C)	mm	0.001	0.001	0.001	0.001	0.001	0.001
Axial Runout (QD12-11-C)	mm	0.001	0.001	0.001	0.001	0.001	0.001
Coning of Table Axis (QD12-35-B)	Arc Seconds	+/-0.5"	+/-0.5"	+/-0.5"	+/-0.5"	+/-0.5"	+/-0.5"
Concentricity of Centre Bore (QD12-36-B)	mm	0.005	0.005	0.005	0.005	0.005	0.005

POS		LP	CLUB	P	W	D	L	GF	GA	GD	PTS
0	▬	(0)	Arsenal	0	0	0	0	0	0	0	0
0	▬	(0)	Aston Villa	0	0	0	0	0	0	0	0
0	▬	(0)	Cardiff City	0	0	0	0	0	0	0	0
0	▬	(0)	Chelsea	0	0	0	0	0	0	0	0
0	▬	(0)	Crystal Palace	0	0	0	0	0	0	0	0
0	▬	(0)	Everton	0	0	0	0	0	0	0	0
0	▬	(0)	Fulham	0	0	0	0	0	0	0	0

Table 1. Classification results on 81 symbolic columns

manual \ computed	our method using the ontology				SMO			
	Food	Micro.	Resp.	Other	Food	Micro.	Resp.	Other
Food	34	0	0	12	45	0	0	1
Micro.	0	16	0	0	4	12	0	0
Response	0	0	1	0	0	0	0	1
Other	3	3	0	12	7	0	0	11

Technique	Major urban (< $10 million) Value often	Major urban (< $10 million) Value occasionally	Major urban (> $10 million) Value often	Major urban (> $10 million) Value occasionally
Adjusted yield approach	4.8% (1)	23.8% (10)	17.0% (9)	43.3% (16)
Cash flow (term and reversion approach)	38.1% (8)	19.1% (8)	22.6% (12)	18.9% (7)
Shortfall approach	52.4% (11)	47.6% (20)	56.6% (30)	27.0% (10)
Layer approach	4.7% (1)	9.5% (4)	3.8% (2)	10.8% (4)
Test of independence: chi-square	5.374 with DF 3 (dependence not proved)		11.46 with D.F.3 (highly dependent)	

Note: The figures in brackets are the actual number of responses in each category. All chi-square tests are based on numbers and not percentages.

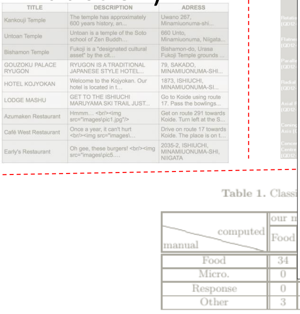
Table Interpretation – Challenges

Noise and diversity

Ignore “entity centric page” (consult wrapper induction)

Specification criteria	Specification
Review details	
Test Date	June 2013
Reviewed using 2013 test programme	Yes
Design	
Smartphone	Yes
Phone style	Candy Bar
Height (mm)	120.0
Width (mm)	67.0
Depth (mm)	10.0
Weight (g)	139
Qwerty keyboard	Yes
Touchscreen	Yes
Display size (inches)	3.1

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534



TITLE	DESCRIPTION	ADRESS
Kankouji Temple	The temple has approximately 600 years history, an...	Uwano 267, Minamiuonuma-shi...
Untoan Temple	Untoan is a temple of the Soto school of Zen Buddh...	660 Unto, Minamiuonuma, Niigata...
Bishamon Temple	Fukoji is a "designated cultural asset" by the cit...	Bishamon-do, Urasa Fukoji Temple grounds ...
GOUZOKU PALACE RYUGON	RYUGON IS A TRADITIONAL JAPANESE STYLE HOTEL...	79, SAKADO, MINAMIUONUMA-SHI...
HOTEL KOJYOKAN	Welcome to the Kojyokan. Our hotel is located in t...	1873, ISHIUCHI, MINAMIUONUMA-SI...
LODGE MASHU	GET TO THE ISHIUCHI MARUYAMA SKI TRAIL JUST...	Go to Koide using route 17. Pass the bowlings...
Azumaken Restaurant	Hmmm... 	Get on route 291 towards Koide. Turn left at the S...
Café West Restaurant	Once a year, it can't hurt <img src="images\...	Drive on route 17 towards Koide. The place is on t...
Early's Restaurant	Oh gee, these burgers! <img src="images\pic5....	2035-2, ISHIUCHI, MINAMIUONUMA-SHI, NIIGATA

Table Interpretation – Challenges

Noise and diversity

Ignore “complex” structures, e.g. col/row spans [Li et al., 2004; Adelfio & Samet, 2013]

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

TITLE	DESCRIPTION	ADRESS
Kankouji Temple	The temple has approximately 600 years history, an...	Uwano 267, Minamiuonuma-shi...
Untoan Temple	Untoan is a temple of the Soto school of Zen Buddh...	660 Unto, Minamiuonuma, Niigata...
Bishamon Temple	Fukoji is a "designated cultural asset" by the cit...	Bishamon-do, Urasa Fukoji Temple grounds ...
GOUZOKU PALACE RYUGON	RYUGON IS A TRADITIONAL JAPANESE STYLE HOTEL...	79, SAKADO, MINAMIUONUMA-SHI...
HOTEL KOJYOKAN	Welcome to the Kojyokan. Our hotel is located in t...	1873, ISHIUCHI, MINAMIUONUMA-SI...
LODGE MASHU	GET TO THE ISHIUCHI MARUYAMA SKI TRAIL JUST...	Go to Koide using route 17. Pass the bowlings...
Azumaken Restaurant	Hmmm... 	Get on route 291 towards Koide. Turn left at the S...
Café West Restaurant	Once a year, it can't hurt <img src="images\...	Drive on route 17 towards Koide. The place is on t...
Early's Restaurant	Oh gee, these burgers! <img src="images\pic5....	2035-2, ISHIUCHI, MINAMIUONUMA-SHI, NIIGATA

Table 1. Classification results on 81 symbolic columns

manual \ computed	our method using the ontology				SMO			
	Food	Micro.	Resp.	Other	Food	Micro.	Resp.	Other
Food	34	0	0	12	45	0	0	1
Micro.	0	16	0	0	4	12	0	0
Response	0	0	1	0	0	0	0	1
Other	3	3	0	12	7	0	0	11

Table 1. Classification results on 81 symbolic columns

manual \ computed	our method using the ontology				SMO			
	Food	Micro.	Resp.	Other	Food	Micro.	Resp.	Other
Food	34	0	0	12	45	0	0	1
Micro.	0	16	0	0	4	12	0	0
Response	0	0	1	0	0	0	0	1
Other	3	3	0	12	7	0	0	11

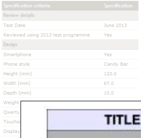
Technique	Major urban (< $10 million) Value often	Major urban (< $10 million) Value occasionally	Major urban (> $10 million) Value often	Major urban (> $10 million) Value occasionally
Adjusted yield approach	4.8% (1)	23.8% (10)	17.0% (9)	43.3% (16)
Cash flow (term and reversion approach)	38.1% (8)	19.1% (8)	22.6% (12)	18.9% (7)
Shortfall approach	52.4% (11)	47.6% (20)	56.6% (30)	27.0% (10)
Layer approach	4.7% (1)	9.5% (4)	3.8% (2)	10.8% (4)
Test of independence: chi-square	5.374 with DF 3 (dependence not proved)		11.46 with D.F.3 (highly dependent)	

Note: The figures in brackets are the actual number of responses in each category. All chi-square tests are based on numbers and not percentages.

Table Interpretation – Challenges

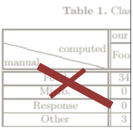
Noise and <u>diversity</u>

Ignore long texts (see free text IE)



Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

TITLE	DESCRIPTION	ADRESS
Kankouji Temple	The temple has approximately 600 years history, an...	Uwano 267, Minamiuonuma-shi...
Untoan Temple	Untoan is a temple of the Soto school of Zen Buddh...	660 Unto, Minamiuonuma, Niigata...
Bishamon Temple	Fukoji is a "designated cultural asset" by the cit...	Bishamon-do, Urasa Fukoji Temple grounds ...
GOUZOKU PALACE RYUGON	RYUGON IS A TRADITIONAL JAPANESE STYLE HOTEL...	79, SAKADO, MINAMIUONUMA-SHI...
HOTEL KOJYOKAN	Welcome to the Kojyokan. Our hotel is located in t...	1873, ISHIUCHI, MINAMIUONUMA-SI...
LODGE MASHU	GET TO THE ISHIUCHI MARUYAMA SKI TRAIL JUST...	Go to Koide using route 17. Pass the bowlings...
Azumaken Restaurant	Hmmm... 	Get on route 291 towards Koide. Turn left at the S...
Café West Restaurant	Once a year, it can't hurt <img src="images\...	Drive on route 17 towards Koide. The place is on t...
Early's Restaurant	Oh gee, these burgers! <img src="images\pic5....	2035-2, ISHIUCHI, MINAMIUONUMA-SHI, NIIGATA



TITLE	DESCRIPTION	ADRESS
Kankouji Temple	The temple has approximately 600 years history, an...	Uwano 267, Minamiuonuma-shi...
Untoan Temple	Untoan is a temple of the Soto school of Zen Buddh...	660 Unto, Minamiuonuma, Niigata...
Bishamon Temple	Fukoji is a "designated cultural asset" by the cit...	Bishamon-do, Urasa Fukoji Temple grounds ...
GOUZOKU PALACE RYUGON	RYUGON IS A TRADITIONAL JAPANESE STYLE HOTEL...	79, SAKADO, MINAMIUONUMA-SHI...
HOTEL KOJYOKAN	Welcome to the Kojyokan. Our hotel is located in t...	1873, ISHIUCHI, MINAMIUONUMA-SI...
LODGE MASHU	GET TO THE ISHIUCHI MARUYAMA SKI TRAIL JUST...	Go to Koide using route 17. Pass the bowlings...
Azumaken Restaurant	Hmmm... 	Get on route 291 towards Koide. Turn left at the S...
Café West Restaurant	Once a year, it can't hurt <img src="images\...	Drive on route 17 towards Koide. The place is on t...
Early's Restaurant	Oh gee, these burgers! <img src="images\pic5....	2035-2, ISHIUCHI, MINAMIUONUMA-SHI, NIIGATA

Table Interpretation – Challenges

Noise and diversity

Ignore numeric tables – those with many numeric values

Specification criteria	Specification
Review details	
Test Date	June 2013
Reviewed using 2013 test programme	Yes
Design	
Smartphone	Yes
[illegible] style	Candy Bar
Height (mm)	120.0
Width (mm)	67.0
Depth (mm)	10.0
We[illegible]	139
[illegible]	Yes
Touchscreen	Yes
Display size (inches)	3.1

Blackberry Q10
Launch date: Apr 2013
Which? score: S

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

Figure 1: Tables feature properties that favour the extraction of information.

TITLE	DESCRIPTION	ADRESS
Kankouji Temple	The temple has approximately 600 years history, an...	Uwano 267, Minamiuonuma-shi...
Untoan Temple	Untoan is a temple of the Soto school of Zen Buddh...	660 Unto, Minamiuonuma, Niigata...
Bishamon Temple	Fukoji is a "designated cultural asset" [illegible] the cit...	Bishamon-do, Urasa Fukoji Temple grounds ...
GOUZOKU PALACE RYUGON	RYUGON [illegible] TRADITIONAL JAPANESE ST[illegible]OTEL...	79, SAKADO, MINAMIUONUMA-SHI...
HOTEL KOJYOKAN	Welcome to [illegible] Kojy[illegible] Our hotel is loca[illegible]d in t...	1873, ISHIUCHI, MINAMIUONUMA-SI...
LODGE MASHU	GET TO THE ISHIUCHI MARUYAMA SKI TRAIL JUST...	Go to Koide using route 17. Pass the bowlings...
Azumaken Restaurant	Hmmm... 	Get on route 291 towards Koide. Turn left at the S...
Café West Restaurant	Once a year, it can't hurt <img src="images\...	Drive on route 17 towards Koide. The place is on t...
Early's Restaurant	Oh gee, these burgers! <img src="images\pic5....	2035-2, ISHIUCHI, MINAMIUONUMA-SHI, NIIGATA

	Geometric tolerances						
Rotation (QD12-06-C)	mm	0.005	0.008	0.008	0.01	0.01	0.015
Flatness (QD12-07-C)	mm	0.008	0.008	0.008	0.012	0.012	0.015
Parallelism (QD12-08-C)	mm	0.013	0.015	0.015	0.02	0.02	0.03

POS		LP	CLUB	P	W	D	L	GF	GA	GD	PTS
0	–	(0)	Arsenal	0	0	0	0	0	0	0	0
0	–	(0)	Aston Villa	0	0	0	0	0	0	0	0
0	–	(0)		0	0	0	0	0	0	0	0

CLUB	P	W	D	L	GF	GA	GD	PTS
Arsenal	0	0	0	0	0	0	0	0
Aston Villa	0	0	0	0	0	0	0	0
Cardiff City	0	0	0	0	0	0	0	0

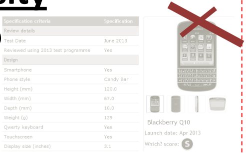
Table 1. Clas[illegible]

computed / manual	our
	Foo[illegible]
Fo[illegible]	34
Mi[illegible]o.	0
Response	0
Other	3

Table Interpretation – Challenges

Noise and <u>diversity</u>

Many studies also ignore vertical tables



Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

Figure 1: Tables feature properties that favour the extraction of information.

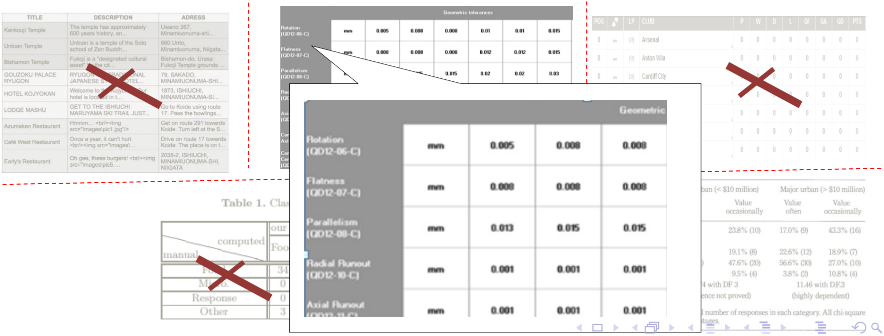
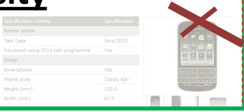


Table Interpretation – Challenges

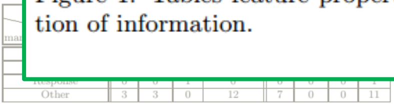
Noise and <u>diversity</u>

Typical tables addressed in the literature



Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

Figure 1: Tables feature properties that favour the extraction of information.



Note: The figures in brackets are the actual number of responses in each category. All chi-square tests are based on numbers and not percentages.

Table Interpretation – Challenges

Noise and <u>diversity</u>

Typical tables

- Each column describes data of the same type
- Each row describes relational data
- Often have a subject column [75% in Venetis et al., 2011; 95% in Wang et al., 2012]
- Represents the majority of “useful” (e.g., containing information useful for search) tables on the web [Venetis et al., 2011]

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

Figure 1: Tables feature properties that favour the extraction of information.

Table Interpretation – Challenges

A multi-task problem

What type?

What relations?

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

Figure 1: Tables feature properties that favour the extraction of information.

What entities/attributes?

Table Interpretation – Challenges

A multi-task problem

Classification/Typing

Relation Extraction

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

Figure 1: Tables feature properties that favour the extraction of information.

Disambiguation

Table Interpretation – Challenges

A multi-task problem

- Task dependency:
  - The output of one task becomes the input of others
- Complexity v.s. Effectiveness
  - What is the time complexity and effectiveness of
    - Sequentially performing each task
    - Other "holistic" models that address them simultaneously

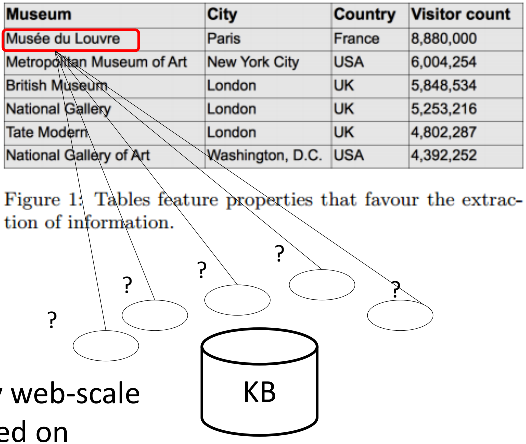
Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

Figure 1: Tables feature properties that favour the extraction of information.

Table Interpretation – Challenges

Scalability

- A search problem
- What is the search space given
  - the KB with millions/billions of nodes
  - the millions/ billions of tables
- Consider that many web-scale IE systems are developed on clusters [e.g., Carlson et al., 2010]



Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

Figure 1: Tables feature properties that favour the extraction of information.

Table Interpretation

- Motivation
- Problem definition
- **Table Interpretation - Methods**

Table Interpretation - Methods

General workflow

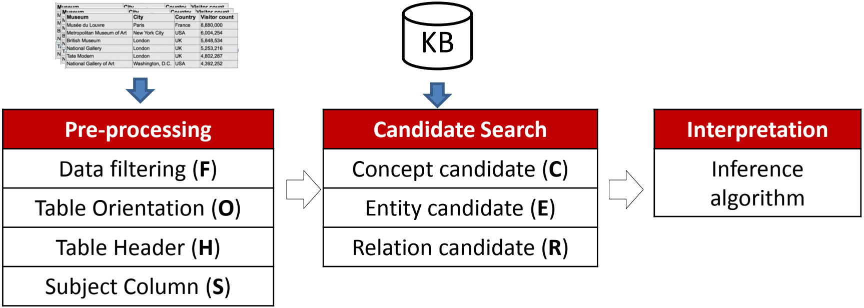


Table Interpretation - Methods

General workflow

(F) Use typical features/ methods [Cafarella et al. 2008a]
(O,H,S) Make typical assumptions

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

KB

Pre-processing
Data filtering (**F**)
Table Orientation (**O**)
Table Header (**H**)
Subject Column (**S**)

Candidate Search
Concept candidate (**C**)
Entity candidate (**E**)
Relation candidate (**R**)

Interpretation
Inference algorithm

Table Interpretation - Methods

General workflow

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

Pre-processing
Data filtering (**F**)
Table Orientation (**O**)
Table Header (**H**)
Subject Column (**S**)

KB

Candidate Search
Concept candidate (**C**)
Entity candidate (**E**)
Relation candidate (**R**)

- KB APIs or customised index of C, E, R
- Differs largely in KBs, indexing/ranking methods

Interpretation
Inference algorithm

Table Interpretation - Methods

General workflow

For each study pointers will be given...

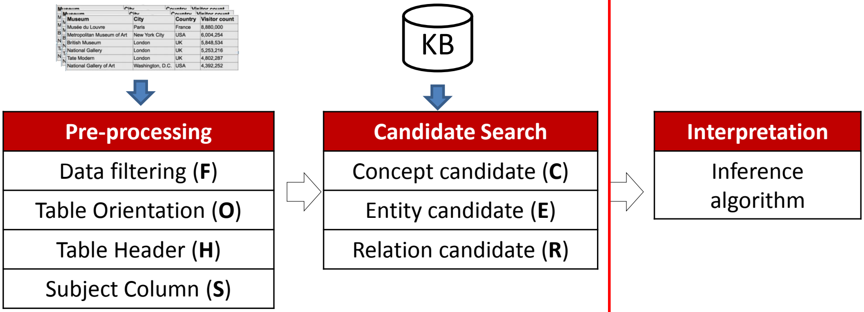


Table Interpretation - Methods

General workflow

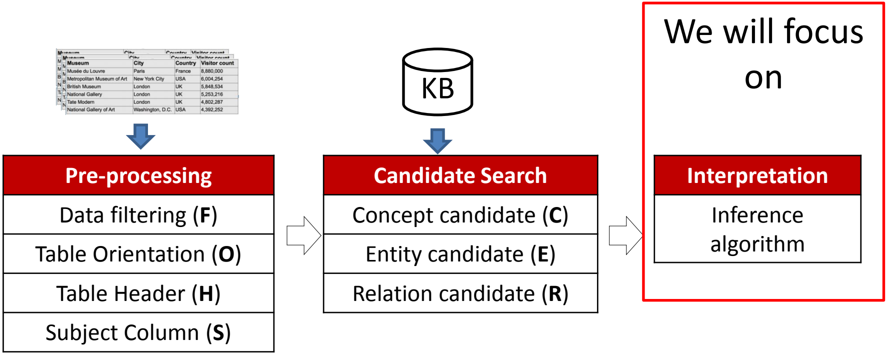


Table Interpretation - Methods

Interpretation – a closer look

Concept_001_cityInTheUK
Concept_023_cityInTheUS
Concept_125_city

Relation_a01_capitalOf
Relation_a87_cityOf
Relation_a91_locatedIn

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

? (not found in KB)

Entity_1_London_UK
Entity_2_London_USA

Table Interpretation - Methods

Interpretation – a closer look

~~Concept_001_cityInTheUK~~
~~Concept_023_cityInTheUS~~
Concept_125_city

~~Relation_a01_capitalOf~~
Relation_a87_cityOf
~~Relation_a91_locatedIn~~

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

Entity_101_MMoA

Entity_1_London_UK
~~Entity_2_London_USA~~

Table Interpretation - Methods

A couple of highly influential work in table information extraction in general

Table Interpretation - Methods

WebTables – Cafarella et al. [2008a, 2008b, 2011]

First to study relational tables on the Web

- 14.1 billion raw HTML tables
- Filter non-relational tables, produced 154 million or 1.1% high quality relational tables
- Detect headers – 71% of relational tables have a header row
- Schema co-occurrence statistics
  - used for schema auto-completion

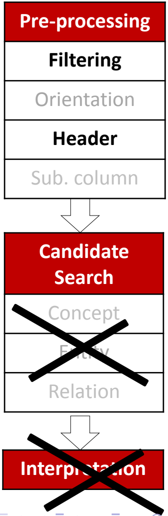


Table Interpretation - Methods

Closely related – Google Fusion Tables

- A collaborative table creation, integration and publication service
- Search: relational table corpus on the Web
- Edit: merge different tables
- Data type annotations (e.g., location, date, number)

Web Tables

Fusion Tables

Send Feedback

List of **countries** and dependencies and their capitals in native ...
http://en.wikipedia.org/.../List_of_**countries**_and_dependencies_and_their_**capitals**_i...

country (exonym)	saint barthélemy	saint helena ...	saint kitts and ...

Show less (35 rows / 5 columns total) - Import data

Country (exonym)	Capital (exonym)	Country (endonym)	Capital (endonym)	Official or native
Saint Barthélemy	Gustavia	Saint-Barthélemy	Gustavia	French
Saint Helena, Ascension	Jamestown			English
Saint Kitts and	Basseterre			English
Saint Martin	Marigot	Saint-Martin	Marigot	French
Saint Lucia	Castries			English

List of **countries** and dependencies and their capitals in native ...

Table Interpretation - Methods

Table Interpretation

Table Interpretation - Methods

Syed et al. [2010], Mulwad et al. [2010, 2011]

KBs:

- DBpedia
- Wikitology [Syed et al. 2008]
  - A specialised IR index of Wikipedia entities
  - Searchable fields: article content, title, redirects, first sentence, categories, types/concepts (Freebase, DBpedia, Yago types), DBpedia info box properties and values, etc.

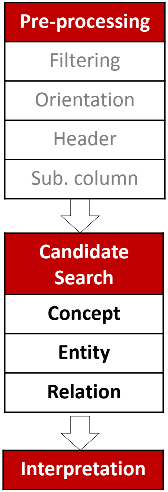


Table Interpretation - Methods

Syed et al. [2010], Mulwad et al. [2010, 2011]

A sequential model

Step 1) Label columns

- Query Wikitology to find **top *N*** candidate entities for each cell value (except header) based on its context in the table
- Get each candidate's type
- Candidates **vote** to determine a uniform type for the column

City
Paris
New York City
London
London
London
Washington, D.C.

Table Interpretation - Methods

Syed et al. [2010], Mulwad et al. [2010,

Custom query considers "context" of each value

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254

Query: title[1]="Paris" & redirects[1]="Paris" & firstSent[1]="City" & linkedConcepts[1]={"Musee du Louvre", "France", "8,880,000"} & infoboxPropertyValues ={"Musee du Louvre", "France", "8,880,000"}

Result: ***N*** candidates matching the query "Paris……"

[1] Searchable fields in Wikitology

Table Interpretation - Methods

Syed et al. [2010], Mulwad et al. [2010, 2011]

A sequential model

Step 1) Label columns

- Query Wikitology to find top *N* candidate entities for each cell value (except header) based on its context in the table
- Get each candidate's type
- Candidates **vote** to determine a uniform type for the column

City ?
Paris
New York City
London
London
London
Washington, D.C.

Table Interpretation - Methods

Syed et al. [2010], Mulwad et al. [2010, 2011]

A sequential model

Step 2) Disambiguate entities

- For each mention, modify the same query in Step 1
  - Adding a constraint on the “type” field to be the concept just learnt
- Re-send the query asking for “exact match” to obtain one entity

City concept_city
Paris ?
New York City ?
London ?
London ?
London ?
Washington, D.C. ?

Table Interpretation - Methods

Syed et al. [2010], Mulwad et al. [2010,

Custom query considers “context” of each value

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254

Query: title[1]=“Paris” & redirects[1]=“Paris” & firstSent[1]=“City” & linkedConcepts[1]={”Musee du Louvre”, “France”, “8,880,000”} & infoboxPropertyValues ={”Musee du Louvre”, “France”, “8,880,000”} **& type=“concept_city”**

[1] Searchable fields in Wikitology

Table Interpretation - Methods

Syed et al. [2010], Mulwad et al. [2010, 2011]

A sequential model

Step 3) Relation Enumeration

- Query each pair of entities of two columns in DBpedia for the predicate connecting them, e.g., <ent:Paris, ?, Ent:France>
- The **majority** wins

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254

Table Interpretation - Methods

Venetis et al. [2011]

KBs:

- A class label database built by mining the Web with Hearst patterns [Hearst 1992]
- A relation database built by running TextRunner [Yates et al., 2007] over the Web

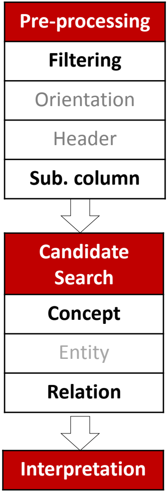


Table Interpretation - Methods

Venetis et al. [2011]

A sequential model:

Step 1) Label columns

- A maximum likelihood model – the best class label $l(A)$ is the one maximising the probability of the values given the class label for the column:
- l_i – candidate label; v_n – values in the cells of a column (except header)

$$l(A) = \arg\max_{l_i} \{\Pr[v_1, \ldots, v_n \mid l_i]\}$$
$$= \prod_j \frac{\Pr[l_i \mid v_j] \times \Pr[v_j]}{\Pr[l_i]} \propto \prod_j \frac{\Pr[l_i \mid v_j]}{\Pr[l_i]}$$

Table Interpretation - Methods

Venetis et al. [2011]

A sequential model:

Step 1) Label columns

- A maximum likelihood model – the best class label *l(A)* is the one maximising the probability of the values given the class label for the column:
- l_i – candidate label; v_n – values in the cells of a column (except header)

$$l(A) = \arg\max_{l_i} \{\Pr[v_1, \ldots, v_n \mid l_i]\}$$
$$= \prod_j \frac{\Pr[l_i \mid v_j] \times \Pr[v_j]}{\Pr[l_i]} \propto \prod_j \frac{\Pr[l_i \mid v_j]}{\Pr[l_i]}$$

Based on frequencies of *v* and *l* in the class label DB

Table Interpretation - Methods

Venetis et al. [2011]

A sequential model:

Step 2) Label relations

- depending on the pairs of cell values from columns A and B
- If a substantial number of values from A and B occur in extractions of the form (a, R, b) in the relations DB
- "substantial number": assessed using the same maximum likelihood model

Table Interpretation - Methods

Shen et al. [2012]

KB:

- YAGO [Suchanek et al., 2007]
- Wikipedia
- Entity mention dictionary (<mention, entity>) for fast candidate entity lookup
  - Built on Wikipedia (page titles, disambiguation, links etc.)

- Entity linking in lists
- Generalisable to tables
- Evaluated against entity linking in tables

Table Interpretation - Methods

Shen et al. [2012]

Component 1: a candidate mapping entity is "good" if the **prior probability** of the entity being mentioned is high

- "A Tale of Two Cities" => the *musical* or *novel*?
- Each candidate entity $r_{i,j} \in Ri$ having the same mention form l_i has different popularity
- Some are obscure and rare for the given mention

$$P_{pr}(r_{i,j}) = \frac{count(r_{i,j})}{\sum_{u=1}^{|R_i|} count(r_{i,u})}$$

Table Interpretation - Methods

Shen et al. [2012]

Component 1: a candidate mapping entity is "good" if the **prior probability** of the entity being mentioned is high

- "A Tale of Two Cities" => the *musical* or *novel*?
- Each candidate entity $r_{i,j} \in Ri$ having the same mention form l_i has different popularity
- Some are obscure and rare for the given mention

$$P_{pr}(r_{i,j}) = \frac{count(r_{i,j})}{\sum_{u=1}^{|R_i|} count(r_{i,u})}$$

Frequency of the entity being mentioned by label l_j in Wikipedia

Table Interpretation - Methods

Shen et al. [2012]

Component 2: a candidate mapping entity is “good” if its type is **coherent** with types of the other mapping entities in the list

- *Sim* calculates semantic similarity
  - Based on YAGO hierarchy using Lin [1998]
  - Based on Wikipedia article corpus using distributional similarity [Harris 1954]

$$Coh(r_{i,j}) = \frac{1}{|L|-1} \sum_{u=1,u\neq i}^{|L|} Sim(r_{i,j}, m_u)$$

L – the entire list

m_u – the mapping entity for the list item u

Table Interpretation - Methods

Shen et al. [2012]

Final form (LQ = linking quality)

$$LQ(r_{i,j}) = \alpha * P_{pr}(r_{i,j}) + (1-\alpha) * Coh(r_{i,j})$$

$$LQ(M) = \alpha * \sum_{s=1}^{|L|} P_{pr}(m_s) + (1-\alpha) * \sum_{s=1}^{|L|} Coh(m_s)$$

- Weight parameter must be learnt
- An iterative substitution algorithm that reduces computation
  - Initialise mappings based on maximum prior only
  - Keep trying new mappings until LQ maximised (local maximum)

Table Interpretation - Methods

Limaye et al. [2010]

A holistic approach based on collective inference

- Markov Network (MN)

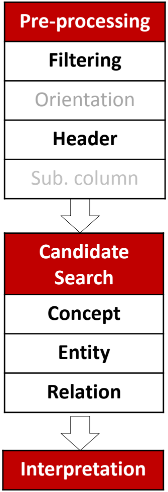
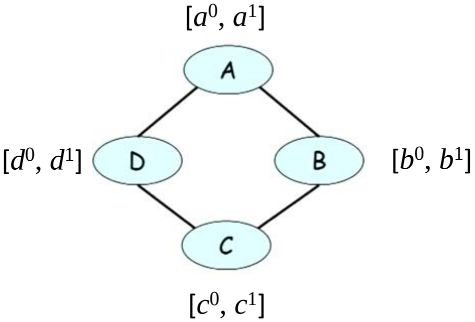


Table Interpretation - Methods

MN – a primer

A graphical representation of dependency between variables



(based on https://class.coursera.org/pgm/lecture/preview)

Table Interpretation - Methods

MN – a primer

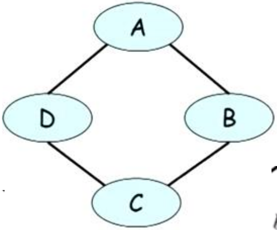
Factors (ϕ) to encode "compatibility" between variables

$\phi_4[D, A]$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1[A, B]$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10



$\phi_3[C, D]$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_2[B, C]$

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

(based on https://class.coursera.org/pgm/lecture/preview)

Table Interpretation - Methods

MN – a primer

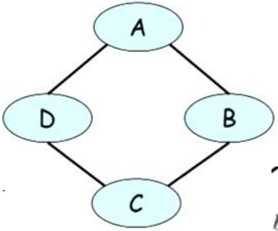
Goal: what is the optimal setting of the variables such that collective "compatibility" is maximised?

$\phi_4[D, A]$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1[A, B]$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10



$\phi_3[C, D]$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_2[B, C]$

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

(based on https://class.coursera.org/pgm/lecture/preview)

Table Interpretation - Methods

Table as an MN

Variables: column type, cell entity, pairwise column relation

Values: candidates from KBs

Compatibility: dependency between candidates

Concept_001_cityInTheUK
Concept_023_cityInTheUS
Concept_125_city

Relation_a01_capitalOf
Relation_a87_cityOf
Relation_a91_locatedIn

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,004,254
British Museum	London	UK	5,848,534
National Gallery	London	UK	5,253,216
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,392,252

Entity_1_London_UK
Entity_2_London_USA

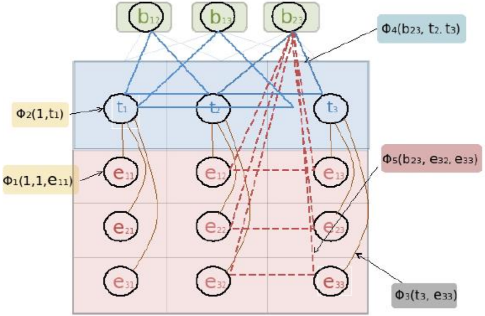
Table Interpretation - Methods

Table as an MN

Variables: column type, cell entity, pairwise column relation

Values: candidates from KBs

Compatibility: dependency between candidates



Limaye's model

Table Interpretation - Methods

Limaye et al. [2010]

Graph construction

- Each column type, cell entity, pairwise column-column relation becomes a variable
- Retrieve candidates (variable values) from YAGO
- Modelling compatibility
  - Cell text and entity label
  - Column header text and type label
  - Column type and cell entity
  - Relation and pair of column types
  - Relation and entity pairs

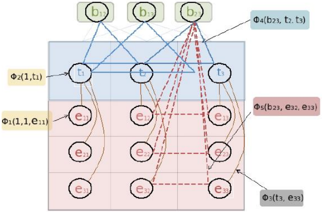
Table Interpretation - Methods

Limaye et al. [2010]

Inference

$$\max_{\mathbf{e},\mathbf{t},\mathbf{b}} \underbrace{\prod_{c,c'} \phi_4(b_{cc'}, t_c, t_{c'}) \prod_{r} \phi_5(b_{cc'}, e_{rc}, e_{rc'})}_{\text{relation}}$$

$$\underbrace{\prod_{c} \phi_2(c, t_c)}_{\text{columns}} \underbrace{\prod_{r} \phi_1(r, c, e_{rc})\, \phi_3(t_c, e_{rc})}_{\text{cells}} .$$



- Implementation: belief propagation

Table Interpretation - Methods

One potential limitation of these methods: candidate search

Name	Birthdate	Political Party	Assumed Office	Height
Barack Obama	4 Aug 1961	Democratic	2009	6'1
Arnold Schwarzenegger	30 Jul 1947	Republican	2003	6'2
Hillary Clinton	26 Oct 1947	Democratic	2009	5'8

Table Interpretation - Methods

Wang et al. [2010]

Tables describing a single entity type (concept) and its attributes

- An entity column + attribute columns
- Goal:
  - Find the entity column and...
  - ... the best matching concept schema

Name	Birthdate	Political Party	Assumed Office	Height
Barack Obama	4 Aug 1961	Democratic	2009	6'1
Arnold Schwarzenegger	30 Jul 1947	Republican	2003	6'2
Hillary Clinton	26 Oct 1947	Democratic	2009	5'8

(US presidents, {Birthdate, Political Party, Assumed Office}, 0.90)

(politicians, {Birthdate, Political Party, Assumed Office}, 0.88)

(NBA players, {Birthdate, Height}, 0.65)

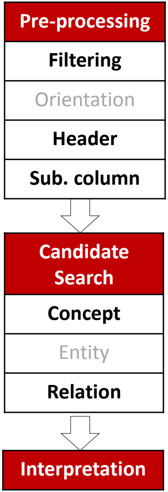


Table Interpretation - Methods

Wang et al. [2010]

KB:

- Probase – a probabilistic KB of concepts, entities and attributes
  - Search API supports:
    - *f(c)* - given a concept c, return its attributes and entities
    - *f(A)* - Given an attribute set A return triples (c, a, $prob$) $a \in A$
    - *g(E)* - Given an entity set E return triples (c, e, $prob$), $e \in E$

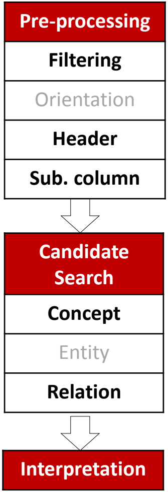


Table Interpretation - Methods

Wang et al. [2010]

The *row* of headers describes a specific concept

Name	Birthdate	Political Party	Assumed Office
Barack Obama	4 Aug 1961	Democratic	2009
Arnold Schwarzenegger	30 Jul 1947	Republican	2003
Hillary Clinton	26 Oct 1947	Democratic	2009

Table Interpretation - Methods

Wang et al. [2010]

The *row* of headers describes a specific concept

The entity *column* should contain entities of a certain concept

Name	Birthdate	Political Party	Assumed Office
Barack Obama	4 Aug 1961	Democratic	2009
Arnold Schwarzenegger	30 Jul 1947	Republican	2003
Hillary Clinton	26 Oct 1947	Democratic	2009

Table Interpretation - Methods

Wang et al. [2010]

The *row* of headers describes a specific concept

The entity *column* should contain entities of a certain concept

$\boxed{\text{The conclusion should be consistent}}$

Name	Birthdate	Political Party	Assumed Office
Barack Obama	4 Aug 1961	Democratic	2009
Arnold Schwarzenegger	30 Jul 1947	Republican	2003
Hillary Clinton	26 Oct 1947	Democratic	2009

Table Interpretation - Methods

Wang et al. [2010]

$SC_A = f(A^s)^1$ $(c, a, prob), a \in A$

Name	Birthdate	Political Party	Assumed Office
Barack Obama	4 Aug 1961	Democratic	2009
Arnold Schwarzenegger	30 Jul 1947	Republican	2003
Hillary Clinton	26 Oct 1947	Democratic	2009

[1] Simplified for explanation. Consult Wang et al. for full details

Table Interpretation - Methods

Wang et al. [2010]

$SC_A = f(A^s)^1$ $(c, a, prob), a \in A$

$SC_E = g(E^{col})^1$ $(c, e, prob), e \in E$

Name	Birthdate	Political Party	Assumed Office
Barack Obama	4 Aug 1961	Democratic	2009
Arnold Schwarzenegger	30 Jul 1947	Republican	2003
Hillary Clinton	26 Oct 1947	Democratic	2009

[1] Simplified for explanation. Consult Wang et al. for full details

Table Interpretation - Methods

Wang et al. [2010]

$SC_A = f(A^s)^1$ $(c, a, prob), a \in A$

$SC_E = g(E^{col})^1$ $(c, e, prob), e \in E$

Name	Birthdate	Political Party	Assumed Office
Barack Obama	4 Aug 1961	Democratic	2009
Arnold Schwarzenegger	30 Jul 1947	Republican	2003
Hillary Clinton	26 Oct 1947	Democratic	2009

$$h(s, col) = \max\{sa_i \cdot se_j \mid (c_i, A_i^S, sa_i) \in SC_A, (c_j, E_j^{col}, se_j) \in SC_E, c_i = c_j\}$$

$$(final\ schema, entity\ column) = \operatorname*{argmax}_{s,col} h(s, col)$$

[1] Simplified for explanation. Consult Wang et al. for full details

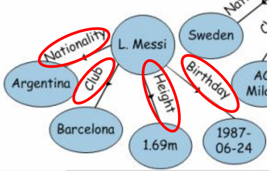
Table Interpretation - Methods

Guo et al. [2011]

Each tuple in a table describes a single entity and each value describes one of its properties

- Goal:
  - Map tuples to entities in a KB
  - Create schema based on mapping

Lionel Messi	Argentina	Barcelona	1.69m	24 June 1987	99
Zlatan Ibrahimovic	Sweden	AC Milan	1.95m	3 October 1981	80
Cristinano Ronaldo	Portugal	Real Madrid	1.86m	5 February 1985	90



Pre-processing
- Filtering
- Orientation
- Header
- **Sub. column**

Candidate Search
- **Concept**
- **Entity**
- **Relation**

Interpretation

Table Interpretation - Methods

Guo et al. [2011]

KB:

- YAGO RDF triples <subject, predicate, object>, e.g., <ent_L.Messi, club, “Barcelona”>
- An inverted free text index of YAGO entities
  - Each entity is an article
  - All objects concatenated as text
  - Enables candidate search by tuples

Lionel Messi	Argentina	Barcelona	1.69m	24
Zlatan Ibrahimovic	Sweden	AC Milan	1.95m	3
Cristinano Ronaldo	Portugal	Real Madrid	1.86m	5

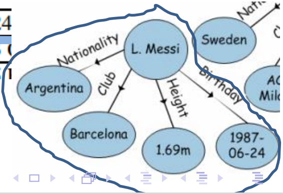


Table Interpretation - Methods

Guo et al. [2011]

Mapping between a tuple (t) and an entity (e)

Based on string similarity

$$score(M) = \sum_{\forall (t(A_i), e(P_j)) \in M} sim(t(A_i), e(P_j))$$

Column index

Predicate index

Lionel Messi	Argentina	Barcelona	1.69m	24
Zlatan Ibrahimovic	Sweden	AC Milan	1.95m	3
Cristinano Ronaldo	Portugal	Real Madrid	1.86m	5

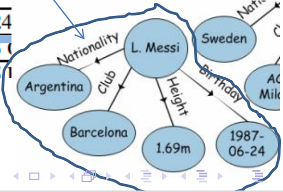


Table Interpretation - Methods

Guo et al. [2011]

Optimal mapping between a tuple (t) and an entity (e)

$$score(M_1) = 0.8 + 0.2 + 0.9 = 1.9$$

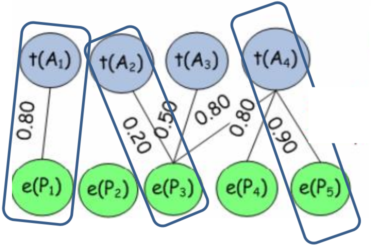


Table Interpretation - Methods

Guo et al. [2011]

Optimal mapping between a tuple (t) and an entity (e)

$$score(M_2) = 0.8 + 0.5 + 0.8 = 2.1$$

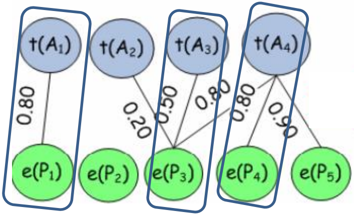


Table Interpretation - Methods

Guo et al. [2011]

Optimal mapping between a tuple (t) and an entity (e)

$$score(M_3) = 0.8 + 0.5 + 0.9 = 2.2$$

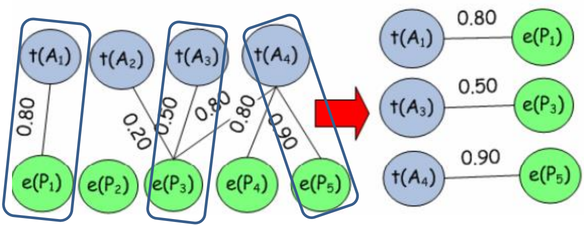


Table Interpretation - Methods

Guo et al. [2011]

Optimal schema for the table

- For each tuple, generate optimal tuple-entity mapping
- Some $t(A_i)$may not be mapped to anything

Lionel Messi	Argentina	Barcelona	1.69m	24 June 1987	99
Zlatan Ibrahimovic	Sweden	AC Milan	1.95m	3 October 1981	80
Cristinano Ronaldo	Portugal	Real Madrid	1.86m	5 February 1985	90
…					
Hao Junmin	China	Schalke 04	1.78m	24 March 1987	60
Shinji Kagawa	Japan	Dormund	1.73m	17 March 1989	88

Table Interpretation - Methods

Guo et al. [2011]

Optimal schema for the table

- For each tuple, generate optimal tuple-entity mapping
- Some $t(A_i)$may not be mapped to anything
- $t(A_i)$ and $t`(A_i)$ (i.e., same column in different tuples) may map to different predicates of an entity type or even different entity types

Lionel Messi	Argentina	Barcelona	1.69m	24 June 1987	99
Zlatan Ibrahimovic	Sweden	AC Milan	1.95m	3 October 1981	80
Cristinano Ronaldo	Portugal	Real Madrid	1.86m	5 February 1985	90
...					
Hao Junmin	China	Schalke 04	1.78m	24 March 1987	60
Shinji Kagawa	Japan	Dormund	1.73m	17 March 1989	88

Table Interpretation - Methods

Guo et al. [2011]

Optimal schema for the table – Maximum weight independent set problem

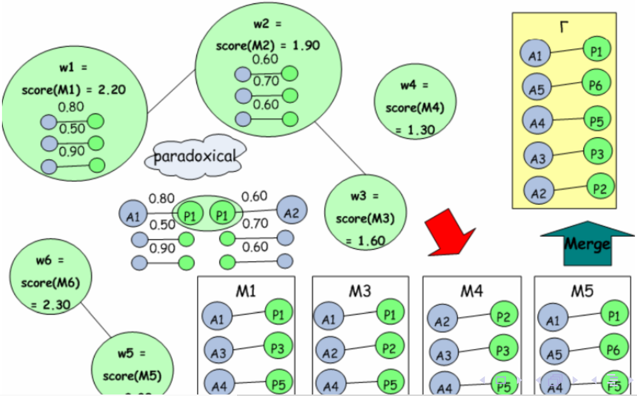


Table Interpretation - Methods

Guo et al. [2011]

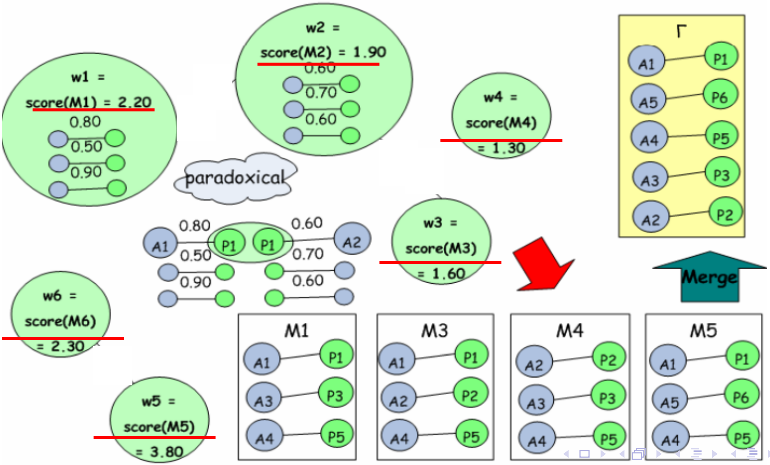


Table Interpretation - Methods

Guo et al. [2011]

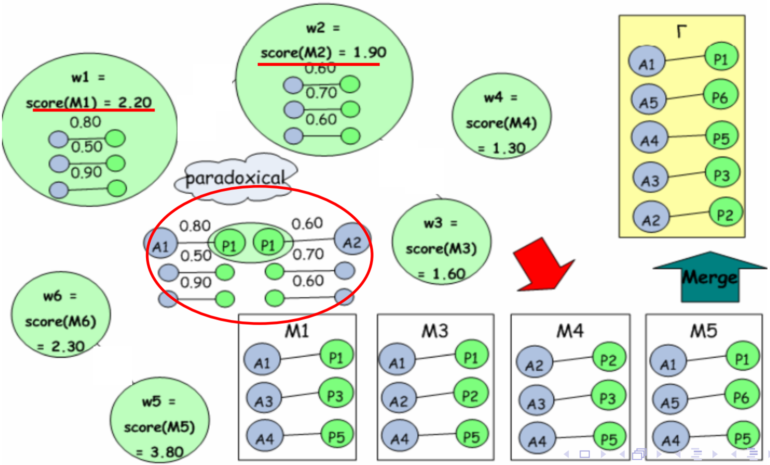


Table Interpretation - Methods

Guo et al. [2011]

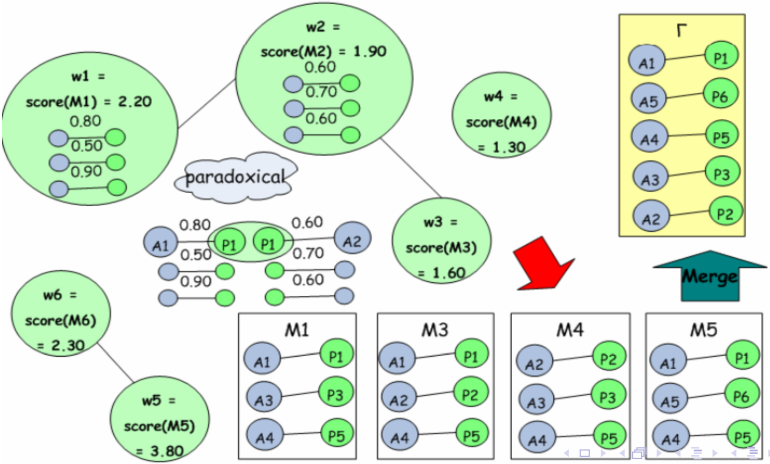


Table Interpretation - Methods

Guo et al. [2011]

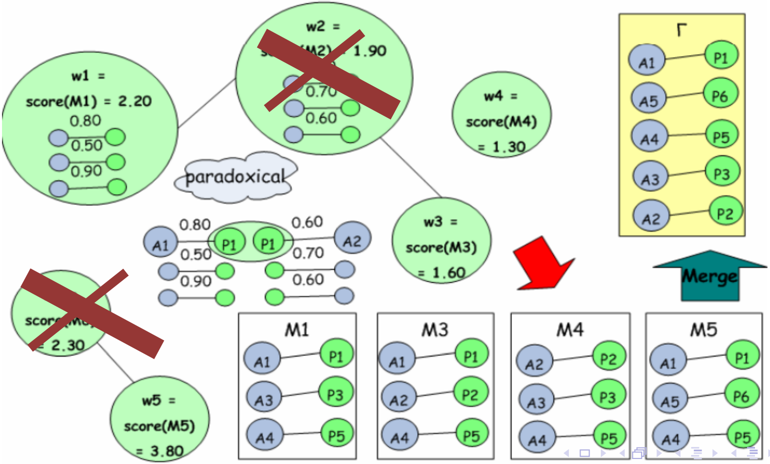


Table Interpretation - Methods

Evaluation and comparison

	Dataset	Evaluation methods	Remark
Limaye	Wiki manual, Web manual, Web relation, Wiki link (≈6500 tables)	Annotation accuracy (% of correct annotation)	
Venetis	Wiki manual, Web manual, additional crawled corpus	Annotation accuracy, table search	Column labelling outperforms Limaye
Shen	Wiki manual, Web manual, Web list (≈ 660 lists)	Annotation accuracy	Entity linking outperforms Limaye (even using only a basic model)

Table Interpretation - Methods

Evaluation and comparison

	Dataset	Evaluation methods	Remark
Wang	69 million tables filtered from the Web	Table search, taxonomy expansion	
Guo	100 Google Fusion Tables	Annotation accuracy	
Syed	5 tables	Annotation accuracy	

Table Interpretation - Methods

A couple of other table interpretation work

Table Interpretation - Methods

DEMO: https://d5gate.ag5.mpi-sb.mpg.de/webaida/

Yosef et al. [2011]

- KB: YAGO
- Goal: linking entities in tables to YAGO
- Intuition: maximising semantic relatedness between entities in a table

Hignette et al. [2007, 2009]; Buche et al. [2013]

- KB: A domain specific ontology densely populated with concepts, relations, and instances
- Goal: column typing, entity linking, relation recognition

Table Interpretation - Methods

Some related research areas

Table Interpretation - Methods

Table (schema) matching and integration

[Assaf et al., 2012; Yakout et al., 2012; Zhang et al., 2013]

Airport Code		Organization	Cost
LHR	England	Microsoft	123.2
LGA	United States	Apple	232.12
HUU	Peru	Orange	321.7
DBO	Australia	IBM	354.64
BGY	Italy	Accenture	243.8

Table 1. Source Table

Airport	Pays	OR_lbl	Cost
LaGuardia	Estados Unidos	MS	201.41
Heathrow	Angleterre	Yahoo	90.5
Queen Alia	الأردن	Samsung	198
Prestwick	Scozia	GOOG	211.27
Beauvais	Frankreich	HP	55.99

Table 2. Target Table

Table Interpretation - Methods

Table (schema) matching and integration

Demo http://downey-n1.cs.northwestern.edu/public/

WikiTable [Bhagavatula et al. 2013]

- Release a normalised Wikipedia table corpus
- Table search
- Table join and integration

Table Interpretation - Methods

Table (schema) matching and integration

WikiTable [Bhagavatula et al. 2013]

Demo http://downey-n1.cs.northwestern.edu/public/

Wikitables List of United States cities by population

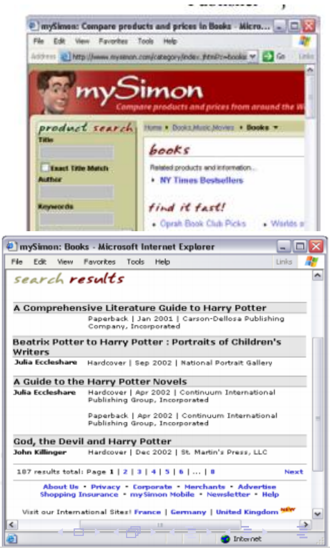
Wikipedia Table: **Cities Formerly over 100,000 People** from **List of United States cities by population** (see Wikipedia)
<-see other tables

Hidden Columns (click to display): Cities Formerly over 100,000 People:Notes | Final standings:Points | Final standings:Rank | Cities Formerly over 100,000 People:Percent decline from peak populatior

City	State	2010 population	Numeric decline from peak population	African American Population	Rank
compare with..	compare with..	compare with..	compare with..	compare with..	compare with..
Albany	New York	97,856	-37,139	505,200 87,897 100,774 ...	43 44 37 50
Allegheny	Pennsylvania	N/A	–	661,839 25,957	30 17
Brooklyn	New York	N/A	–	505,200 87,897 100,774 ...	43 44 37 50
Camden	New Jersey	77,344	-47,211	46,314 145,085	18 54
Canton	Ohio	73,007	-43,905	211,672 133,039 60,705	32 16 28
Dearborn	Michigan	98,153	-13,854	590,226 57,939 23,127	10 1 27

Table Interpretation - Methods

Table from the "hidden" web
[Wang 2003]



Outline

Web Scale Information Extraction

Summary & Conclusion

Summary & Conclusion

Web as a text corpus

- Unlimited domains, unlimited documents
- Structured and unstructured

Web IE

- Promising route towards Semantic Web
- Many challenges
    - Scalability, coverage and quality, heterogeneity

Web IE systems & methods

- > 10 years history
- Free form text v.s. structures

Summary & Conclusion

Wrapper induction

- Many Web pages are
  - automatically generated using scripts
  - present regular structures
  - good opportunity for IE
- Schema and semantic are not known in advance
  - training material required
    - Schema, annotations...
  - minimize user input

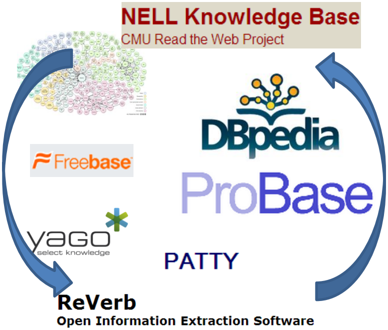
Summary & Conclusion

Table Interpretation

- Tables contain complementary information to free text
- Great opportunities to search engines
- Very large amount but very noisy
- A complex, multi-task problem
- Computation-demanding

The Take-away Message

Knowledge Base



IE

Web Texts

Organisations, Information and Knowledge

The University Of Sheffield.

Further reading I

Adelfio, M. D. and Samet, H.
Schema extraction for tabular data on the web.
Proceedings of VLDB Endowment.

Aggarwal, C. C. and Wang, J. (2007).
XProj : A Framework for Projected Structural Clustering of XML Documents.
pages 46–55.

Agichtein, E. and Gravano, L. (2000).
Snowball: extracting relations from large plain-text collections.
In *Proceedings of the fifth ACM conference on Digital libraries*, DL '00, pages 85–94, New York, NY, USA. ACM.

Ahmad, A., Eldad, L., Aline, S., Corentin, F., Raphaël, T., and David, T. (2012).
Improving schema matching with linked data.
In *First International Workshop On Open Data*.

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.

Banko, M., Cafarella, M., Soderland, S., Broadhead, M., and Etzioni, O. (2007).
Open information extraction from the web.
In *IJCAI*, pages 2670–2676.

Further reading II

Blanco, L., Dalvi, N., and Machanavajjhala, A. (2011).
Highly efficient algorithms for structural clustering of large websites.
Proceedings of the 20th international conference on World wide web - WWW '11, page 437.

Buche, P., Dibie-Barthélemy, J., Ibanescu, L., and Soler, L. (2013).
Fuzzy web data tables integration guided by an ontological and terminological resource.
IEEE Transactions on Knowledge and Data Engineering, 25(4):805–819.

Cafarella, M. J., Halevy, A., and Madhavan, J. (2011).
Structured data on the web.
Communications of the ACM, 54(2):72–79.

Cafarella, M. J., Halevy, A., Wang, D. Z., Wu, E., and Zhang, Y. (2008).
Webtables: exploring the power of tables on the web.
Proceedings of VLDB Endowment, 1(1):538–549.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Jr., E. H., and Mitchell, T. (2010).
Toward an architecture for never-ending language learning.
In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, pages 1306–1313. AAAI Press.

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.

Further reading III

Ciravegna, F., Gentile, A. L., and Zhang, Z. (2012).
Lodie: Linked open data for web-scale information extraction.
In Maynard, D., van Erp, M., and Davis, B., editors, *SWAIE*, volume 925 of *CEUR Workshop Proceedings*, pages 11–22. CEUR-WS.org.

Crescenzi, V. and Mecca, G. (2004).
Automatic information extraction from large websites.
Journal of the ACM, 51(5):731–779.

Crescenzi, V., Mecca, G., and Merialdo, P. (2002).
Wrapping-oriented classification of web pages.
. . . of the 2002 ACM symposium on . . ., (ii):1108–1112.

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.

Dalvi, N., Machanavajjhala, A., and Pang, B. (2012).
An analysis of structured data on the web.
Proc. VLDB Endow., 5(7):680–691.

Further reading IV

Downey, D. and Bhagavatula, C. S. (2013).
Using natural language to integrate, evaluate, and optimize extracted knowledge bases.
In *The 3rd Workshop on Knowledge Extraction at CIKM 2013*, AKBC '13.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2004).
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.

Fader, A., Soderland, S., and Etzioni, O. (2011).
Identifying relations for open information extraction.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1535–1545, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gentile, A. L., Zhang, Z., Augenstein, I., and Ciravegna, F. (2013).
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.

Gottron, T. (2008).
Clustering template based web documents.
Advances in Information Retrieval, pages 40–51.

Further reading V

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.

Guo, X., Chen, Y., Chen, J., and Du, X. (2011).
Item: extract and integrate entities from tabular data to rdf knowledge base.
In *Proceedings of the 13th Asia-Pacific web conference on Web technologies and applications*, APWeb'11, pages 400–411, Berlin, Heidelberg. Springer-Verlag.

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.

Harris, Z. (1954).
Distributional structure.
Word, 10(23):146–162.

He, J., Gu, Y., Liu, H., Yan, J., and Chen, H. (2013).
Scalable and noise tolerant web knowledge extraction for search task simplification.
Decision Support Systems.

Further reading VI

Hignette, G., Buche, P., Dibie-Barthélemy, J., and Haemmerlé, O. (2007).
An ontology-driven annotation of data tables.
In *Proceedings of the 2007 international conference on Web information systems engineering*, WISE'07, pages 29–40, Berlin, Heidelberg. Springer-Verlag.

Hignette, G., Buche, P., Dibie-Barthélemy, J., and Haemmerlé, O. (2009).
Fuzzy annotation of web data tables driven by a domain ontology.
In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications*, ESWC 2009 Heraklion, pages 638–653, Berlin, Heidelberg. Springer-Verlag.

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

Laender, A. H. F., Ribeiro-Neto, B. a., da Silva, A. S., and Teixeira, J. S. (2002).
A brief survey of web data extraction tools.
ACM SIGMOD Record, 31(2):84.

Li, S., Peng, Z., and Liu, M. (2004).
Extraction and integration information in html tables.
In *Proceedings of the The Fourth International Conference on Computer and Information Technology*, CIT '04, pages 315–320, Washington, DC, USA. IEEE Computer Society.

Further reading VII

Limaye, G., Sarawagi, S., and Chakrabarti, S. (2010).
Annotating and searching web tables using entities, types and relationships.
Procedings. VLDB Endowment, 3(1):1338–1347.

Lin, D. (1998).
An information-theoretic definition of similarity.
In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Mulwad, V., Finin, T., and Joshi, A. (2011).
Automatically generating government linked data from tables.
In *Working notes of AAAI Fall Symposium on Open Government Knowledge: AI Opportunities and Challenges*.

Mulwad, V., Finin, T., Syed, Z., and Joshi, A.
T2ld: Interpreting and representing tables as linked data.
In Polleres, A. and Chen, H., editors, *ISWC Posters&Demos*, CEUR Workshop Proceedings. CEUR-WS.org.

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.

Nakashole, N., Theobald, M., and Weikum, G. (2011).
Scalable knowledge harvesting with high precision and high recall.
In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 227–236, New York, NY, USA. ACM.

Organisations, Information and Knowledge

The University Of Sheffield.

Further reading VIII

Quercini, G. and Reynaud, C. (2013).
Entity discovery and annotation in tables.
In *Proceedings of the 16th International Conference on Extending Database Technology*, EDBT '13, pages 693–704, New York, NY, USA. ACM.

Shen, W., Wang, J., Luo, P., and Wang, M.
In Yang, Q., Agarwal, D., and Pei, J., editors, *KDD*, pages 1424–1432. ACM.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007).
Yago: a core of semantic knowledge.
In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA. ACM.

Suchanek, F. M. and Weikum, G. (2013).
Knowledge harvesting in the big-data era.
In Ross, K. A., Srivastava, D., and Papadias, D., editors, *SIGMOD Conference*, pages 933–938. ACM.

Syed, Z., Finin, T., and Joshi, A. (2008).
Wikipedia as an ontology for describing documents.
In *Proceedings of the Second International Conference on Weblogs and Social Media*. AAAI Press.

Syed, Z., Finin, T., Mulwad, V., and Joshi, A. (2010).
Exploiting a web of semantic data for interpreting tables.
In *Proceedings of the Second Web Science Conference*.

Further reading IX

Venetis, P., Halevy, A., Madhavan, J., Paşca, M., Shen, W., Wu, F., Miao, G., and Wu, C. (2011).
Recovering semantics of tables on the web.
Proceedings of VLDB Endowment, 4(9):528–538.

Wang, J. and Lochovsky, F. H. (2003).
Data extraction and label assignment for web databases.
In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 187–196, New York, NY, USA. ACM.

Wang, J., Wang, H., Wang, Z., and Zhu, K. Q. (2012).
Understanding tables on the web.
In *Proceedings of the 31st international conference on Conceptual Modeling*, ER'12, pages 141–155, Berlin, Heidelberg. Springer-Verlag.

Wijaya, D., Talukdar, P. P., and Mitchell, T. (2013).
Pidgin: Ontology alignment using web text as interlingua.
In *Proceedings of the Conference on Information and Knowledge Management (CIKM 2013)*, San Francisco, USA. Association for Computing Machinery.

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.

Further reading X

Wu, W., Li, H., Wang, H., and Zhu, K. Q. (2012).
Probase: a probabilistic taxonomy for text understanding.
In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 481–492, New York, NY, USA. ACM.

Yakout, M., Ganjam, K., Chakrabarti, K., and Chaudhuri, S. (2012).
Infogather: entity augmentation and attribute discovery by holistic matching with web tables.
In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 97–108, New York, NY, USA. ACM.

Yosef, M. A., Hoffart, J., Bordino, I., Spaniol, M., and Weikum, G. (2011).
AIDA: an online tool for accurate disambiguation of named entities in text and tables.
In *Proceedings of the 37th International Conference on Very Large Databases*, VLDB'11, pages 1450–1453.

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.

Zhang, Z., Gentile, A. L., Augenstein, I., Blomqvist, E., and Ciravegna, F. (2013).
Mining equivalent relations from linked data.
In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 289–293, Sofia, Bulgaria. Association for Computational Linguistics.

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.

Further reading XI