

Web Scale Information Extraction TUTORIAL @ ECML/PKDD 2013

Anna Lisa Gentile Ziqi Zhang



Department of Computer Science, The University of Sheffield, UK

27th 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<http://lemurproject.org/clueweb09/>

^b<http://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<http://lemurproject.org/clueweb09/>

^b<http://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

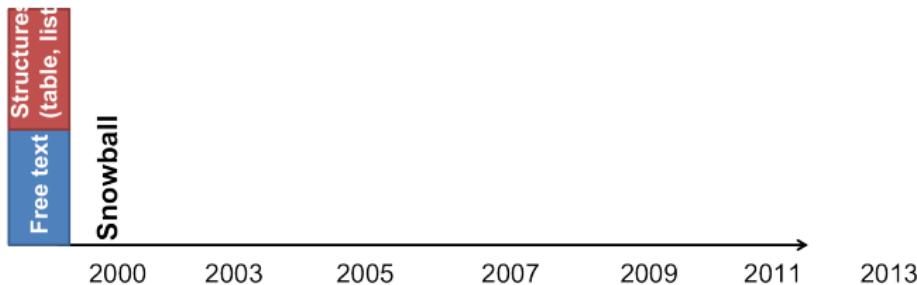
Web IE Methods and Systems

- 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]
- Probbase [Wu et al., 2012]
- LODIE [Ciravegna et al., 2012]

Web IE Methods and Systems

Snowball [Agichtein and Gravano, 2000]

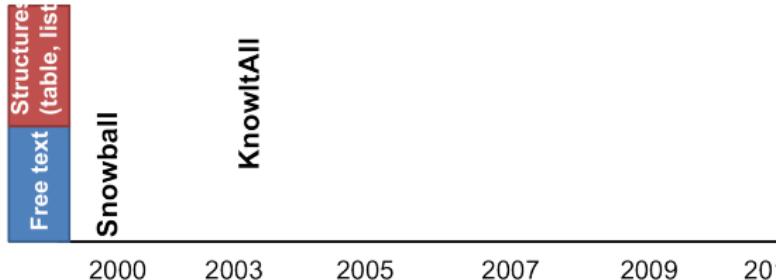
- Learn syntactic patterns to extract relation instances
 - 5-tuple <ordinal, url, left, middle, right>
 - selectiveness (Named Entity Recognition)
- Bootstrap with seed + domain specific heuristics
 - use case: <Organisation, Location> pairs



Web IE Methods and Systems

KnowItAll [Etzioni et al., 2004]

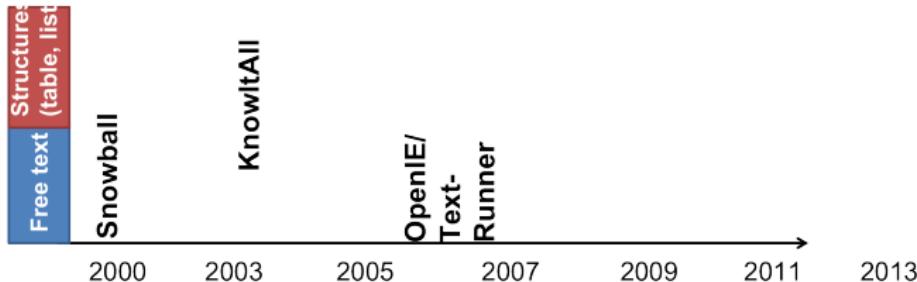
- Input: ontology and rule templates (Hearst patterns)
 - generic syntactic patterns (e.g., cities such as <?>) to extract relation/class instances
- statistically generated "discriminator phrases" for each class
- bootstrap with 2 names for each class



Web IE Methods and Systems

OpenIE/TextRunner [Banko et al., 2007]

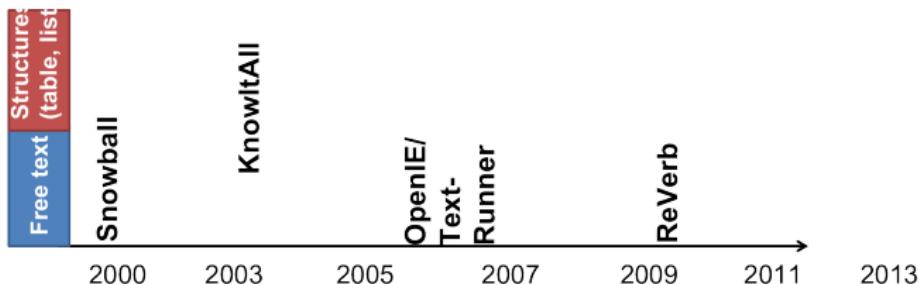
- Learn syntactic patterns to extract any relation instances from any domains (open IE)
- Completely unsupervised, no need for seeds
 - Input: corpus C, parser phase on a portion of C
 - pattern generation from parsed documents, t: $\langle e_1, r, e_2 \rangle$
 - feature vector generation and Naive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



Web IE Methods and Systems

ReVerb [Fader et al., 2011]

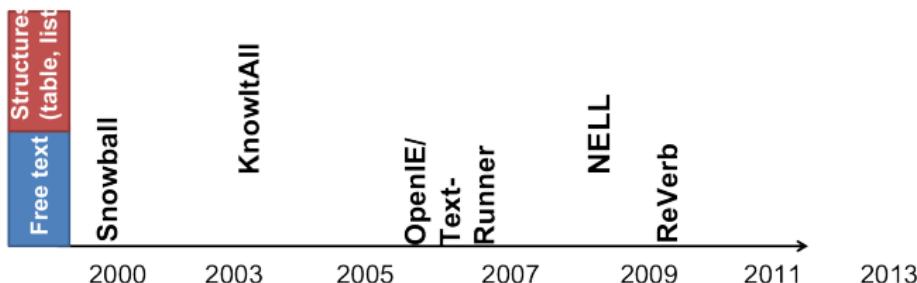
- Open IE improved:
 - syntactic and lexical constraints to select the target verbs
 - Input: POS-tagged and NP-chunked corpus



Web IE Methods and Systems

NELL [Carlson et al., 2010]

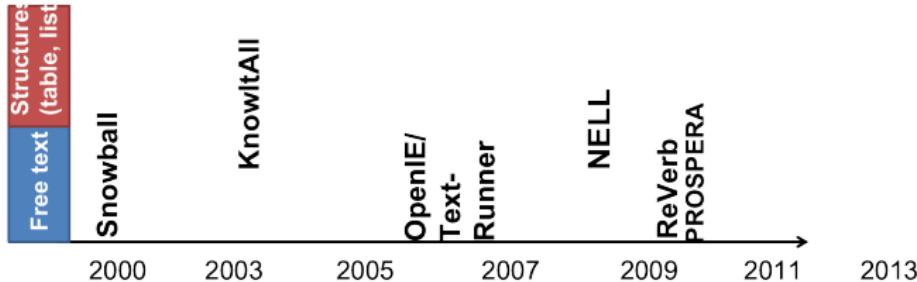
- Multi-strategy, “coupled” learning
 - enforcing constraints and/or strengthening evidences
- Bootstrap with seed ontology (class, relation, instance) + coupling rules



Web IE Methods and Systems

PROSPERA [Nakashole et al., 2011]

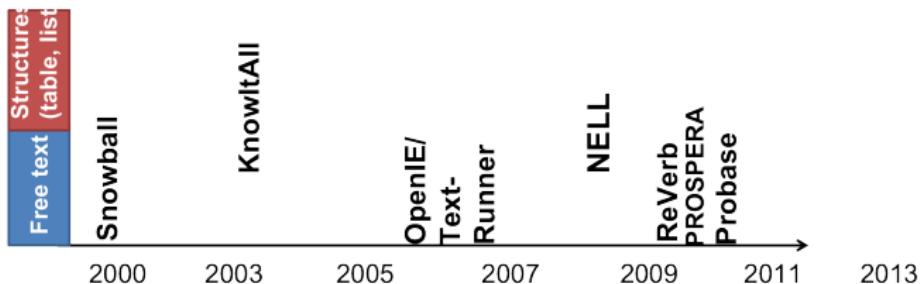
- N-gram item-set patterns to generalise narrow syntactic patterns to boost recall
- Reasoning with large KB (YAGO) to constrain extractions to boost precision
 - Input: target relations and type signature for involved entities
- Integration with KB (data reconciliation)



Web IE Methods and Systems

Probase [Wu et al., 2012]

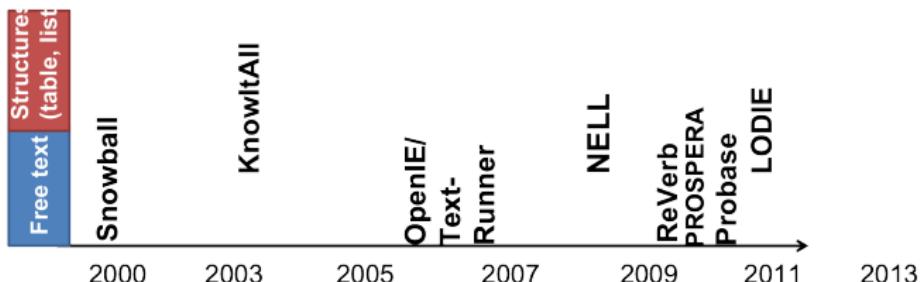
- Building a probabilistic concept taxonomy
 - First iteration - bootstrap with syntactic patterns
 - Following iterations - previously learnt knowledge used to semantically constrain new extractions



Web IE Methods and Systems

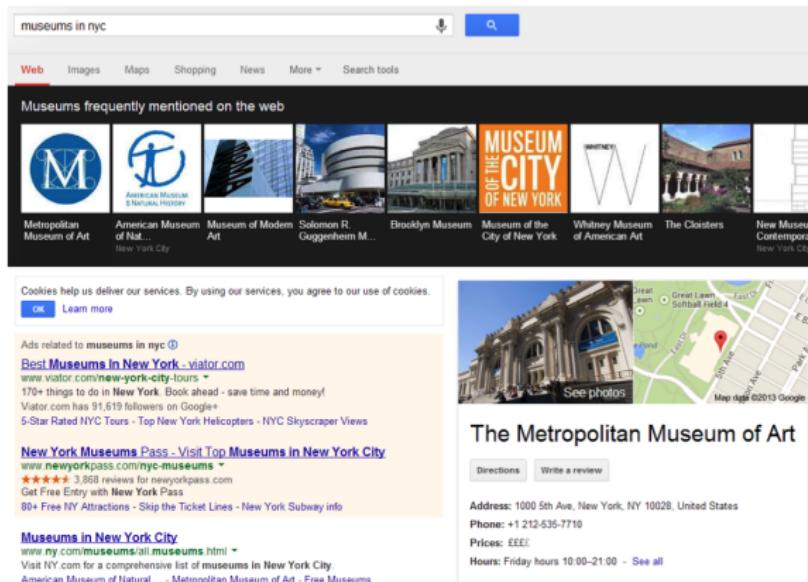
LODIE [Ciravegna et al., 2012]

- 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



museums in nyc

Web Images Maps Shopping News More Search tools

Museums frequently mentioned on the web

Metropolitan Museum of Art American Museum of Natural History Museum of Modern Art Solomon R. Guggenheim M... Brooklyn Museum Museum of the City of New York Whitney Museum of American Art The Cloisters New Museum Contemporary New York City

Cookies help us deliver our services. By using our services, you agree to our use of cookies. OK Learn more

Ads related to museums in nyc

[Best Museums In New York - viator.com](#)
[www.viator.com/new-york-city-tours](#) ▾
170+ things to do in New York. Book ahead - save time and money!
Viator.com has 91,619 followers on Google+
5-Star Rated NYC Tours - Top New York Helicopters - NYC Skyscraper Views

[New York Pass - Visit Top Museums in New York City](#)
[www.newyorkpass.com/nyc-museums](#) ▾
★★★★★ 3,668 reviews for newyorkpass.com
Get Free Entry with New York Pass
80+ Free NY Attractions - Skip the Ticket Lines - New York Subway info

[Museums in New York City](#)
[www.nycommune.com/all-museums.html](#) ▾
Visit NY.com for a comprehensive list of museums in New York City
American Museum of Natural History - Metropolitan Museum of Art - Free Museums

Great Lawn
Great Lawn Softball Field 4
East St.
West St.
Dad
Dad
5th Ave.
5th Ave.
Park Ave.
Map data ©2013 Google

The Metropolitan Museum of Art

Directions Write a review

Address: 1000 5th Ave, New York, NY 10028, United States
Phone: +1 212-535-7710
Prices: ~~EEEE~~
Hours: Friday hours 10:00-21:00 - See all

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

Wrapper Induction: example

Extracting book attributes on e-commerce websites

amazon.com Help... Books & more... personalised recommendations, New releases! Bestsellers
Your Account | Today's Deals | Gift & White List | Gift Cards

Search Books Advanced Search

Books

Click to LOOK INSIDE! Eat, Pray, Love: One Woman's Search for Everything Across Italy, India and Indonesia [Paperback]
Elizabeth Gilbert Author

Last Price: \$18.24 & eligible for FREE Super Saver Shipping on orders over \$25. Details
Now \$16.76 (45%)
In Stock.
Shipped by Amazon.com. Gift-wrap available.
Last delivered Tuesday, June 2017. Choose One-Day Shipping at checkout. Details
Buy Used from \$1.93 \$8.33 used from \$3.35 Kindle Edition from \$4.99

Format: Paperback
Author: Elizabeth Gilbert
ASIN: B000FQHJLW
ISBN-13: 9780345437758
ISBN-10: 034543775X
Audible CD: Audible
Unknown Binding
Audiobooks \$30.99 or less with new Audible membership

Check Out Related Media
Watch a video for Eat, Pray, Love

Amazon Video

TITLE
AUTHOR

BARNES & NOBLE www.bn.com GET FREE SHIPPING My Account | Sign In

Books textbooks eBooks music Kids Toys & Games DVD & Blu-ray Music Home & Gift Gift Cards More

SEARCH Books Books

Awakening (Mass Market Paperback - Release)
Kate Chopin Author Rating: #1 in Fiction > Classics > Feminist (166 ratings)
Read customer reviews Write a Review
PUBLISHED: February 1886
AGES: 12 and up
Series Rank: 1,491

TITLE
AUTHOR
DATE

NEW FROM BN.COM SPEND \$25, GET FREE SHIPPING DETAILS
\$4.49 Online Price (You Save 10%) Add to Bag
Usually ships within 24 hours

PICK ME UP Reserve & pick up in 60 minutes at your local store. Enter your zip FIND IN-STORE

OTHER FORMATS
Hardcover \$25.98
Paperback \$6.75
Mass Market Paperback - Release \$4.45
Other Format - Unabridged \$17.99
Compact Disc - Unabridged, 5 CDs, 5 hrs. 30 mins. \$17.05
MP3 on CD - Unabridged \$17.99
MP3 Basic - Abridged \$13.25

Customers who bought this also bought

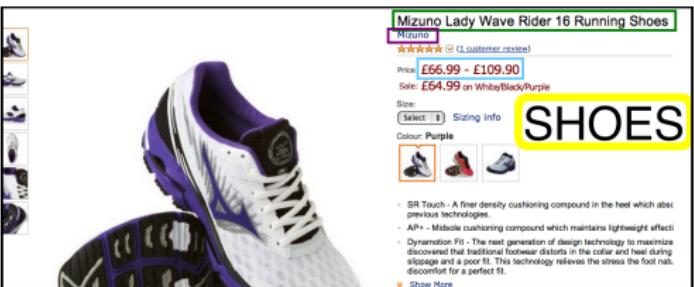
Related Subjects
- Politics & Social Issues - Politics
- Love & Relationships - Fiction
- American Fiction
More by This Author
- biographies -

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

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
 - ~ 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)

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/*/*/*/*)

- $S = \{S_1, S_2, \dots, S_k\}$ 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 .

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 $\{(pos_1 = site.com), (pos_2 = a_1), (pos_3 = a_2), \dots\}$
- $W(t)$ set of Web pages containing t
- $C = \{W_1, \dots, W_k\}$ 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

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-learning
[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

Verticals and Attributes	One Labeled Seed Site	Many Unseen Sites
Books	The Time Machine H. G. Wells (See All Contributors) Paperback Kessinger Publishing, LLC June 30, 2004	New Moon (The Twilight Saga, Book 2) Stephenie Meyer (See All Contributors) Paperback \$12.99 Free Delivery Tomb With a View (The Twilight Saga, Book 3) Stephenie Meyer (See All Contributors) Paperback \$7.00 The Host (The Host, Book 1) Stephenie Meyer (See All Contributors) Paperback \$12.99 The Host (The Host, Book 1) Stephenie Meyer (See All Contributors) Paperback \$12.99 The Host (The Host, Book 1) Stephenie Meyer (See All Contributors) Paperback \$12.99
Restaurants	Scalini Fedeli 3 of 6,826 restaurants in New York City 71 ratings Write a review Cuisine: Italian Address: 165 Duane St, New York City, NY 10013 Phone: (212) 528-0400	10 Degrees South 3 of 1,000 restaurants in Dallas, TX 71 ratings Write a review Cuisine: Mexican Address: 2000 N Akard St, Dallas, TX 75201 Phone: (972) 852-7046 Qwik Cup Espresso 3 of 1,000 restaurants in Dallas, TX 71 ratings Write a review Cuisine: Italian Address: 1000 Main St, Dallas, TX 75202 Phone: (972) 234-9999 Italian Grille 3 of 1,000 restaurants in Dallas, TX 71 ratings Write a review Cuisine: Italian Address: 707 Preston Road, Dallas, TX 75238 Phone: (972) 243-1234 POPOLOS ITALIAN GRILLE 3 of 1,000 restaurants in Dallas, TX 71 ratings Write a review Cuisine: Italian Address: 707 Preston Road, Dallas, TX 75238 Phone: (972) 243-1234 Ben's Deli 3 of 1,000 restaurants in Dallas, TX 71 ratings Write a review Cuisine: Italian Address: 119 Main St, Dallas, TX 75202 Phone: (972) 243-0200 The Salted Iva Restaurant 3 of 1,000 restaurants in Dallas, TX 71 ratings Write a review Cuisine: Italian Address: 2436 Madison Ave, Dallas, TX 75201 Phone: (972) 243-7770 Redenzy 3 of 1,000 restaurants in Dallas, TX 71 ratings Write a review Cuisine: Italian Address: 119 Main St, Dallas, TX 75202 Phone: (972) 243-0200 Dion's 3 of 1,000 restaurants in Dallas, TX 71 ratings Write a review Cuisine: Italian Address: 119 Main St, Dallas, TX 75202 Phone: (972) 243-0200
Autos	2011 Dodge Durango Express Sport Utility Price: \$21,115 Mileage: 26 - 23 mpg Model: Durango Body Type: SUV 2011 Honda CR-V Price: \$21,115 Mileage: 26 - 23 mpg Model: CR-V Body Type: SUV 2011 Ford Escape Price: \$21,115 Mileage: 26 - 23 mpg Model: Escape Body Type: SUV 2011 Nissan Frontier SV 4x4 Crew Cab Long Box Price: \$21,115 Mileage: 26 - 23 mpg Model: Frontier Body Type: SUV 2011 Toyota Prius v Price: \$21,115 Mileage: 40 - 30 mpg Model: Prius v Body Type: Hatchback	2011 Chevrolet Cruze Price: \$21,115 Mileage: 30 - 23 mpg Model: Cruze Body Type: Sedan 2011 Mitsubishi Lancer Price: \$21,115 Mileage: 30 - 23 mpg Model: Lancer Body Type: Sedan 2011 Honda Accord Price: \$21,115 Mileage: 30 - 23 mpg Model: Accord Body Type: Sedan 2011 Ford Fusion Price: \$21,115 Mileage: 30 - 23 mpg Model: Fusion Body Type: Sedan 2011 Toyota Camry Price: \$21,115 Mileage: 30 - 23 mpg Model: Camry Body Type: Sedan 2011 Honda Accord Price: \$21,115 Mileage: 30 - 23 mpg Model: Accord Body Type: Sedan

[Hao et al., 2011]

Supervised methods: Multi-view learners

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

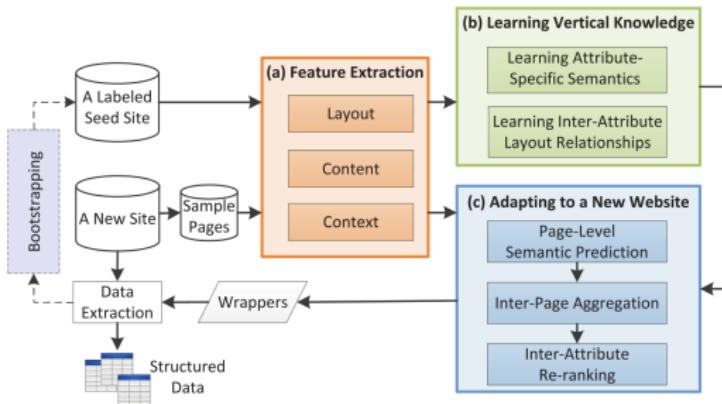
- Weak features
 - general across attributes, verticals and websites
 - identify a large amount of candidate attribute values
 - likely to contain noise
- Strong features
 - site-specific
 - derived in an unsupervised manner

Characteristics

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

Supervised methods: Multi-view learners

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

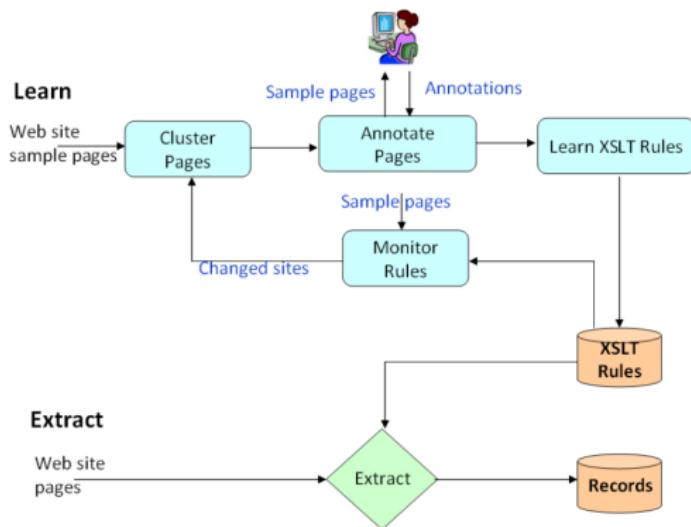


[Hao et al., 2011]

Supervised methods: Vertex!

Complete Wrapper lifecycle

- 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

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

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

- 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
- $w(X_i) = F(X_i) \cdot I(X_i)$

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\$
		Computer S.	An underg...	2nd Ed., H. Cover	2000	30\$
				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>

Books	CSBooks.com
You searched for books by John Smith (author)	
Database Primer	
First Edition, Paperback	Our Price: \$20
Second Edition, Hard Cover	Our Price: \$30
2005	
Book Description	
This book is designed for the theory and technology of database systems. The main topics are the relational model and the SQL query language.	
Computer Systems	
First Edition, Paperback	Our Price: \$40
1995	
Book Description	
This book is designed for students who want to learn computer architecture. It includes a brief history of computers and focuses on parallel machines.	

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

CSbooks.com

You searched for books by Paul Jones

XML at Work
First Edition, Paperback
1999

Stock Description
A book for the description of VMS, and related standards (VLS, CBL etc...)

HTML and Scripts

1990	Our Price: \$30	
Second Edition, Hard Cover	Our Price: \$45	
1999		

Book Description
A book for the description of VMS, and related standards (VLS, CBL etc...)

JavaScripts

2000	Our Price: \$50	
------	------------------------	--

Stock Description
A book for the description of VMS, and related standards (VLS, CBL etc...)

Search | **Logout** | **Customer Support** | **Help** | **Order Status** | **Checkouts** | **Checkouts** | **Checkouts**

c. Data Extraction Output

Title number of CD-ROMs made: 1			
Schema Number: A-1(A) C-1(C) D-1(F) F-1(G)		Total Time: 4' 182 sec	
John Smith	smith@mathworks.com	D	E
		Database Server	This book introduces the reader to the basic technology of databases.
		First Edition 1986 \$60 Second Edition 2000 \$30	
		Computer Systems	
		First Edition 1995 \$40 Paperback	An intermediate level introduction to computer... http://www.mathworks.com
<hr/>		<hr/>	
Paul Jensen	jensen@mathworks.com	XSL of Web	A comprehensive description of XSL, and all related standards... http://www.mathworks.com
		First Edition, 1998 \$30	
		HTML and SGML	
		First Edition 1997 \$30 Second Edition 1999 \$41 Hard Cover	A studio IT handbook, with a quick lesson to the... http://www.mathworks.com
		JavaScript	
		REF 2000 \$30	A must for Webmaster's...bookshelf

[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

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

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

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

	TAG PATH	TEXT
01:	<html><body><a>	Blockx 3D Pro 1.3
02:	<html><body><dl><dt>	Price:
03:	<html><body><dl><dt><div>	\$1.10
04:	<html><body><dl><dt>	Last updated:
05:	<html><body><dl><dt><div>	11/18/2010

SKES - Page Representation example

```
01: <html><body>
02: <a> TAG TREE
03:   Archipelago 1.14
04: </a>
05: <dl>
06:   <dt>Price:</dt> TAG PATH
07:     <div>$2.99</div>
08:   <dt>Last updated:</dt>
09:     <div>09/26/2010</div>
10: </dl>
11: <a> TAG NODES
12:   Recommendations
13: </a>
14: <ul> TEXT TAG PATH
15:   <li>Par 72 Golf</li>
16:   <li>Mathpac 5.6</li>
17:   <li>FourNumGuess 1.0.6</li>
18: </ul>
19: </html></body>
```

SKES PAGE REPRESENTATION

	<u>TAG PATH</u>	
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

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

SKES - pros and cons

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

Unsupervised methods: LODIE Wrapper Induction

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

LODIE Wrapper Induction: task definition

- C - set of *concepts* of interest $C = \{c_1, \dots, c_i\}$
- their attributes $\{a_{i,1}, \dots, a_{i,k}\}$
- 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

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 $\langle xpath, value_{i,k} \rangle$ 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 $a_{i,k}$.

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

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')). }
```

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

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

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

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

1 Overview

2 Wrapper Induction

3 Table Interpretation

4 Conclusions

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 programmes	Yes	
Device		
Smartphone	Yes	
Phone style	Candy bar	
Height (mm)	122.0	
Width (mm)	67.0	
Depth (mm)	10.0	
Weight (g)	139	
Screen resolution	Yes	
Touchscreen	Yes	
Display size (inches)	3.1	
BlackBerry Q10		
Launch date: Apr 2013		
Wikipedia score: 5		

POS	LP	CLUB	P	W	D	L	GF	GA	GD	PTS
0	#	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	#	Chelsea		0	0	0	0	0	0	0
0	#	Crystal Palace		0	0	0	0	0	0	0
0	#	Everton		0	0	0	0	0	0	0
0	#	Fulham		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, Minamionuma-shi...
Urban Temple	Ubano is the temple of the Soto school of Zen Buddhi...	660 Urito, Minamionuma, Niigata...
Bishamon Temple	Fukoji is a "designated cultural asset" by the cit...	Bishamon-do, Urusa Fukoji Temple grounds ...
GOUIZOKU PALACE RYUGON	RYUGON IS A TRADITIONAL JAPANESE STYLE HOTEL...	79, SAKAMOTO MINAMIONUMA-SHI...
HOTEL KOYOKAN	Welcome to the Koyokan. Our hotel is located in ...	1873, ISHIUCHI, MINAMIONUMA-SI...
LODGE MASHU	GET TO THE ISHIUCHI MARUYAMA SKI TRAIL JUST...	Go to Kode using route 17. Pass the bowings...
Azumaken Restaurant	Hmmm... 	Get on route 291 towards Kode. Turn left at the S...
Café West Restaurant	Once a year, it can't hurt MINAMIONUMA-SHI, NIIGATA	

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

Generalization / Subdivision							
Rotation	(E010-08-C3)	mm	0.005	0.000	0.000	0.01	0.01
Distance	(E010-07-C2)	mm	0.000	0.000	0.000	0.012	0.012
Position	(E010-06-C3)	mm	0.013	0.015	0.015	0.02	0.02
Radial Runout	(E010-05-C3)	mm	0.001	0.001	0.001	0.001	0.001
Axial Runout	(E010-01-C3)	mm	0.001	0.001	0.001	0.001	0.001
Corner of Table Axis	(E010-05-G5)	Arc Seconds	+/-0.5°	+/-0.5°	+/-0.5°	+/-0.5°	+/-0.5°
Concentricity of Center Holes	(E010-06-G5)	mm	0.005	0.005	0.005	0.005	0.005

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

Technique	Major urban (< \$10 million)		Major urban (> \$10 million)	
	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		11.46 with DF 3 (dependence not proved)	(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
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

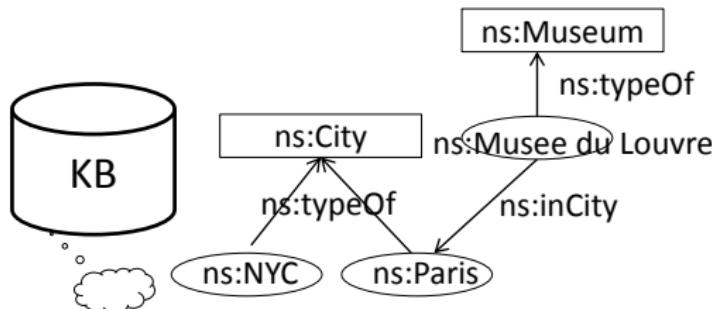


Table Interpretation – Definition

Input tables

Museum	City	Country	Visitor count
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

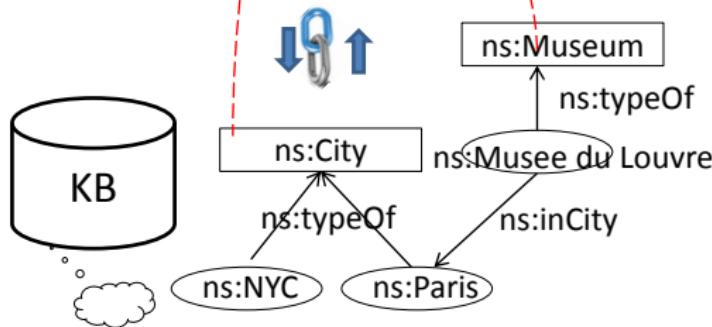


Table Interpretation – Definition

Input tables

Museum	City	Country	Visitor count
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

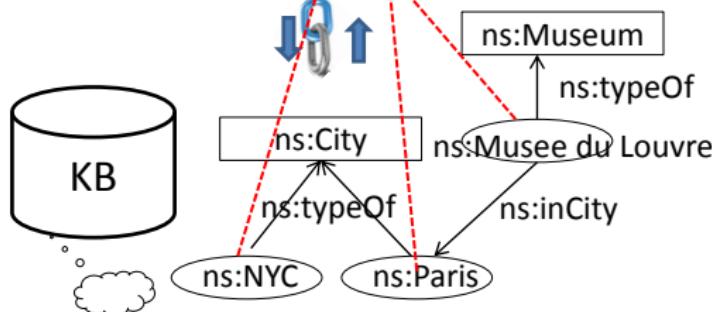


Table Interpretation – Definition

Input

tables

Museum	City	Country	Visitor count
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

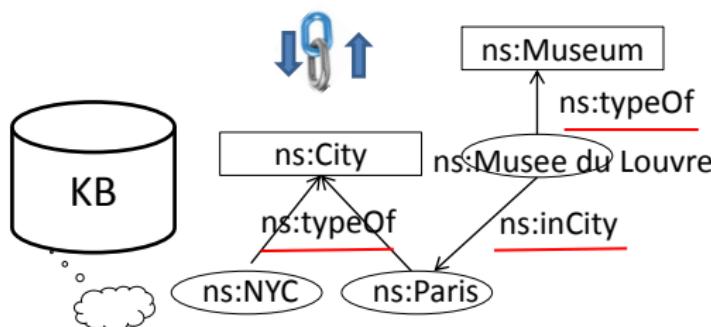


Table Interpretation – Definition

Input tables

Museum	City	Country	Visitor count
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

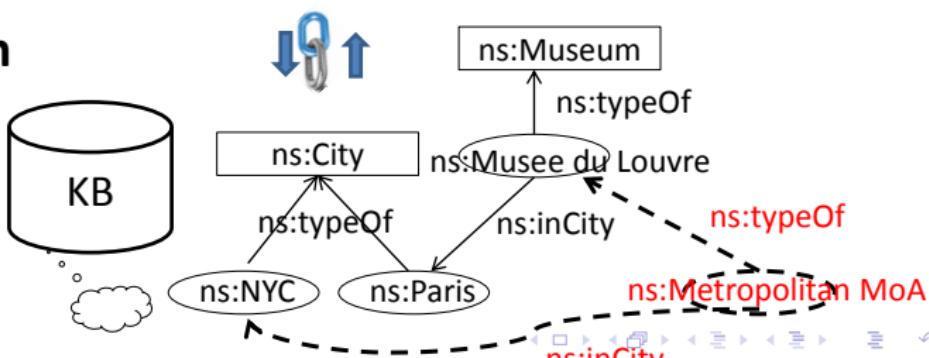


Table Interpretation – Definition

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 diversity

Table Interpretation – Challenges

Specification criteria		Specification
Review details		
Test Date	June 2013	
Reviewed using 2013 test programmes	Yes	
Device		
Smartphone	Yes	
Phone style	Candy bar	
Height (mm)	122.0	
Width (mm)	67.0	
Depth (mm)	10.0	
Weight (g)	139	
Screen resolution	Yes	
Touchscreen	Yes	
Display size (inches)	3.1	
BlackBerry Q10		
Launch date: Apr 2013		
Wikipedia score: 5		

POS	LP	CLUB	P	W	D	L	GF	GA	GD	PTS
0	#	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	#	Chester		0	0	0	0	0	0	0
0	#	Crystal Palace		0	0	0	0	0	0	0
0	#	Everton		0	0	0	0	0	0	0
0	#	Fulham		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, Minamionuma-shi...
Urban Temple	Ubano is the temple of the Soto school of Zen Buddhi...	660 Uto, Minamionuma, Niigata...
Bishamon Temple	Fukoji is a "designated cultural asset" by the cit...	Bishamon-do, Urasa Fukoji Temple grounds ...
GOUIZOKU PALACE RYUGON	RYUGON IS A TRADITIONAL JAPANESE-STYLE HOTEL...	79, SAKAMOTO, MINAMIONUMA-SHI...
HOTEL KOYOKAN	Welcome to the Koyokan. Our hotel is located in t...	1873, ISHIUCHI, MINAMIONUMA-SI...
LODGE MASHU	GET TO THE ISHIUCHI MARUYAMA SKI TRAIL JUST...	Go to Kode using route 17. Pass the bowings...
Azumaken Restaurant	Hmmm... 	Get on route 291 towards Kode. Turn left at the S...
Café West Restaurant	Once a year, it can't hurt MINAMIONUMA-SHI, NIIGATA	

A recap of example “relational” tables on the web...

Generalization factors						
Iteration (E010 - 06-C3)	min	0.005	0.000	0.000	0.01	0.01
Iteration (E010 - 07-C3)	min	0.000	0.000	0.000	0.012	0.012
Parallelism (E010 - 06-C3)	max	0.013	0.015	0.015	0.02	0.02
Radial Parallel (E010 - 07-C3)	min	0.001	0.001	0.001	0.001	0.001
Arcal Parallel (E010 - 01-C3)	min	0.001	0.001	0.001	0.001	0.001
Coming of Table Axis (E010 - 05-C3)	Arc Seconds	+0.5°	+0.5°	+0.5°	+0.5°	+0.5°
Concentricity of Center Axes (E010 - 06-C3)	min	0.005	0.005	0.005	0.005	0.005

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

Technique	Major urban (< \$10 million)		Major urban (> \$10 million)	
	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 DF 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
Device		
Smartphone	Yes	
Phone	Candy Bar	
Height (mm)	128.0	
Width (mm)	67.0	
Depth (mm)	10.0	
Weight (g)	138	
Qwerty keyboard	Yes	
Touchscreen	Yes	
Display size (inches)	3.1	
		
		BlackBerry Q10
		Launch date: Apr 2013
		Whicht! score: 5

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	ADDRESS	
Kankou Temple	The temple has approximately 600 years history, etc...	Ueno 267, Minamisuna-shi... 660 Units.	
Unzen Temple	Unzen is a temple of the Soto school of Zen Buddhi...	Minamisuna, Nagata...	
Bishamon Temple	Fukko is a "regenerated cultural asset" by the city of...	Sakurahon-dō, Unzen Fukko town, Nagasaki...	
GOUZOKU PALACE RYUGON	RYUGON IS A TRADITIONAL JAPANESE STYLE HOTEL...	79, SAKADO MINAMISUNA-SHI...	
HOTEL KOYOKAN	Welcome to the Koyokan Onsen Hotel, located in the city...	1873, ISHUCHI, MINAMISUNA-SHI...	
LODGE MASHI	GET TO THE ISHUCHI MARUYAMA SKI TRAIL JUST...	Go to Kotsu using route 17, Pass the bowings...	
Azumakan Restaurant	Hmmm... 	Get on route 291 towards Kotsu. Turn left at the 5...	
Cafe West Restaurant	Once a year, it can't hurt 	Drive on route 17 towards Kotsu. The place is on the ...	
		2008-2 (SHUUCHI, MINAMISUNA-SHI, NAGATA)	
Early's Restaurant	Oh gee, these burgers! \$10 million)		
------------------------------	--------------------	------------------------------	--------------------
Value often	Value occasionally	Value often	Value occasionally
4.8% (1)	23.8% (10)	17.0% (6)	43.3% (16)
9.5% (4)	38.5% (2)	10.8% (4)	
38.1% (8)	19.1% (8)	22.6% (12)	18.9% (7)
52.4% (11)	47.6% (20)	56.6% (30)	27.0% (10)
4.7% (1)	9.5% (4)	3.8% (2)	
5.374 with DF 3		11.46 with DF 3	
(dependence not proved)		(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

TITLE	DESCRIPTION	ADRESS
Kankouji Temple	The famous temple of the 1000 years history, etc.	Uezo-cho, Minamimurase-shi, Nara-ken.
Unzen Temple	Unsan is a temple of the Soto school of Zen-Buddhism.	990 Uno, Minazumicho, Nagasaki-ken.
Benten Temple	Famous for a magnificated cultural treasure.	Benten-cho, Urusa Mura, Nagasaki-ken.
GOUZOKU PALACE RYUGON	RYUGON IS A TRADITIONAL JAPANESE STYLE HOTEL.	79, SAKADO, MINAMIRANJIMA-SHII, 1873, ISHIKAWA-CHO, NAGASAKI-KEN.
HOTEL KOYOKAN	Stay here to the Koyokan, Old Japanese style hotel.	Go to the Koyokan Hotel.
LODGE MASHU	GET TO THE ISHUCHI MARYAMA SHI TRAIL JUST.	Go to the Onsen route using 17. Pass the bowlings.
Azumakan Restaurant	Hmm... "sushi+ring sushi+ring" is the best!	Get on route 291 towards the Onsen.
Cafe West Restaurant	Once a year, it comes here! "ring+ring" is great!	Drive on route 17 toward Koda. The place is on 1.
Early's Restaurant	Oh ge, here, these burger! "ring+ring" "image+image" is great!	2035-2, (ISHUCHI), MINAMIRANJIMA-SHII, NAGASAKI.

Specification Criteria	Specification
Review details	
Test Date	June 2013
Reviewed using	2013 test programme
Design	
Smartphone	Yes
Phone style	Candy Bar
Height (mm)	128.0
Width (mm)	67.0
Depth (mm)	15.0
Weight (g)	139
QWERTY keyboard	Yes
Touchscreen	Yes
Display size (inches)	3.1




Blackberry Q10
Launch date: Apr 2013
Which? score: **5**

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.

	Geometric Inferences						
Iteration (1020-06-C)	avg	0.005	0.008	0.000	0.01	0.01	0.015
Distance (1020-07-C)	avg	0.000	0.008	0.000	0.012	0.012	0.015
Parallelogram (1020-08-C)	avg	0.013	0.015	0.015	0.02	0.02	0.03
Radial Prism (1020-09-C)	avg	0.001	0.001	0.001	0.001	0.001	0.001
Acute Prism (1020-11-C)	avg	0.001	0.001	0.001	0.001	0.001	0.001
Coring of A Face (1020-25-05)	Avg Seconds	+0.5*	+0.5*	+0.5*	+0.5*	+0.5*	+0.5*
Concavity of A Face (1020-36-C)	avg	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

	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 revision 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)
Layover 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 DF 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)

TITLE	DESCRIPTION	ADDRESS
Kankou Temple	The temple has approximately 600 years history; en...	Ueno 267, Minamisomoma-ishi...
Unzen Temple	Unzen is a temple of the Soto school of Zen Buddhi...	660 Utsuts...
Bishamon Temple	FUKO is a "nostragated" cultural center of...	Bishamon-do, Urata
GOUZOKU PALACE RYUGOKU	RYUGOKU IS A TRADITIONAL JAPANESE STYLE HOTEL...	79, SAKADO, MINAMIAKUMUMA-ISHI...
HOTEL KOYOKAN	Welcome to the Koyokan. Our hotel is located at...	1873, IRIUCHI,
LODGE MASHI	GET TO THE IRUUCHI MARYAMA SKI TRAIL JUST...	MINAMIAKUMUMA-ISHI,
Azumakan Restaurant	Hennin... 	NIGATA
Cafe West Restaurant	Once in year, it can't hurt 	Get on route 291 towards
Early's Restaurant	Oh gee, there's burgen! 	Route 17. Turn left at the S...

Table 1. Class

	computed	our n
manual		Food
Food		34
Micro.		0
Response	0	0 1 0 0 0 0 1
Other	3	3 0 12 7 0 0 11

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

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

Specification criteria

Specification

Review details

June 2013

Reviewed using 2013 test programme Yes

Design

Yes

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

(dependence not proved)

(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 “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

Table 1. Classification results on 81 symbolic columns

TITLE	DESCRIPTION	ADDRESS							
Kankou Temple	The temple has approximately 600 years history, etc..	Ueno 257, Minamisomoma-ishi...							
Unzen Temple	Unzen is a temple of the Soto school of Zen Buddhism.	660 Unzen, Minamisomoma, Nagata...							
Bishamon Temple	Fukko is a "reconstructed cultural temple" of the 19th century.	Bishamon-do, Urabe							
GOUZOKU PALACE RYUGOKU	RYUGOKU IS A TRADITIONAL JAPANESE STYLE HOTEL..	79, SAKADO, MINAMINAGANO-ISHI...							
HOTEL KOYOKAN	Welcome to the Koyokan. Our hotel is located at the center of...	1873, IJIKUCHI, MINAMINAGANO-ISHI...							
LODGE MASHI	GET TO THE MASHI HACHIMARUYAMA SKI TRAIL JUST...	On the route 291 towards Kode. Pass the bottom of...							
Azumaya Restaurant	Hennin... 	Get on route 291 towards Kode. Turn left at the S...							
Cafe West Restaurant	Once a year, it can't hurt 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	

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

Technique	Major urban (< \$10 million)		Major urban (> \$10 million)	
	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		11.46 with DF 3	
				(dependence not proved)
				(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 long texts
(see free text IE)

TITLE	DESCRIPTION	ADDRESS
Kankūji Temple	The temple has approximately 600 years history, etc.	Ueno 287, Minamimurono-shi, Gifu, JAPAN
Unzen Temple	Unzen is one of the Seven Schools of Zen Buddha.	600, Unzen-cho, Minamimurono, Nigata, Japan
Bishamon Temple	Fujii is a "designated cultural asset" by the city.	Bishamon-dō, Urusa Fujii Temple grounds
GOZOKUJI PALACE RYUGON	RYUJION IS A THERMAL SPRINGS HOTEL. WELCOME TO THE KOKYOIN. Our hotel is located in it.	79, SAKADO MINAMIMURONO-SHI, Gifu, JAPAN
HOTEL KOKYOJIN		1873, ISHUKI MINAMIMURONO-SI, Gifu, JAPAN
LODGE MASHU	GET TO THE ISHUCHI MARYAMA SO TRAIL JUST...	Go to Kurobe route using 17. Pass the 7.
Azumazumi Restaurant	https://www.google.com/search?q=azumazumi+jpg	Turn left at the 7. Kurobe. Turn left at the S.
Café West Restaurant	Once a year, it can't hurt •-img-img	Drive on route 17 towards Kurobe. The sign is on I...
Early's Restaurant	Oh gee, these burgers! -img-img	203-2, (ISHUCHI MINAMIMURONO-SHI, NIGATA, JAPAN)

Table 1. Clas-

Manual	Computed	Food
For	34	
Micro.	0	
Response	0	
Other	3	



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 – Challenges

Noise and diversity

Ignore numeric tables – those with many numeric values

TITLE	DESCRIPTION	ADRESS
Kankou Temple	The temple was approximately 600 years old.	Usami, Minamimurama-ishi.
Unsan Temple	Unsan is a temple of the Soto school of Zen Buddhis.	690 Uruo, Minamimurama, Nagata.
Bahamont Temple	Fukuo is a "designated cultural property".	Bahamont-dori, Uruse Minamimurama, Nagata.
GOUZOKU PALACE RYUGYON	RYUJOKEI, THE ORIGINAL JAPANESE PALACE. Welcome to the original Japanese Palace.	79, SAKADO, MINAMURUNUMA-SHI, 1873, ICHIUCHI, MINAMURUNUMA-SHI.
HOTEL KOJYUKAN		
LODGE MASHU	GET TO THE ISHICHIU MARUYAMA SKI TRAIL JUST.	Go to Kuro to ride route 17. Pass the bowings.
Amanoi Restaurant	Irrem...-eeri-ring Irrem...-eeri-ring	Get on route 231 towards Rikudo. Turn left at the Drive bridge. The place is on it.
Cafe West Restaurant	Once a year, it's hot! Irrem...-eeri-ing...irrem...-eeri-	2032-2, ISHICHIU, MINAMURUNUMA-SHI, NAGATA.
Early's Restaurant	Oh gee, these burger!~irrem...-eeri-ing irrem...-eeri-ing...irrem...-eeri-	

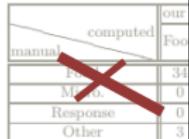


Table 1. Clas-

Anna Lisa Gentile, Ziqi Zhang

Table Interpretation – Challenges

Noise and diversity

Many studies also ignore vertical tables

TITLE	DESCRIPTION	ADDRESS
Kankou Temple	The temple has approximately 600 years history, etc..	Ueno 267, Minamisomoma-ishi.. 660 Units, Minamisomoma, Nigata..
Unison Temple	Unison is a temple of the Soto school of Zen Buddhist...	
Bishamon Temple	FUKUOKA's integrated cultural complex.	Bishamon-do, Urabe
GOUZOKU PALACE RYUGON	JAPANESE PALACE PAL	79, SAKADO, MINAMISOMOMA-SHI..
HOTEL KOYUKAN	Welcome to Koyukan Hotel	1873, IJIKUCHI, MINAMISOMOMA-SHI..
LODGE MASHI	GET TO THE IJIKUCHI MARYAMA SKI TRAIL JUST...	17. Pass the border
Azumakan Restaurant	Hennin... img/eng...	Get on route 291 towards
Cafe West Restaurant	Once a year, it can't hurt Henninimg/...	Kode. Turn left at the S...
Early's Restaurant	Oh yes, there's burgen! Henninimg/...	Drive on route 17 towards

Table 1. Clas

our	computed
Food	
Manual	
Response	34
Other	0

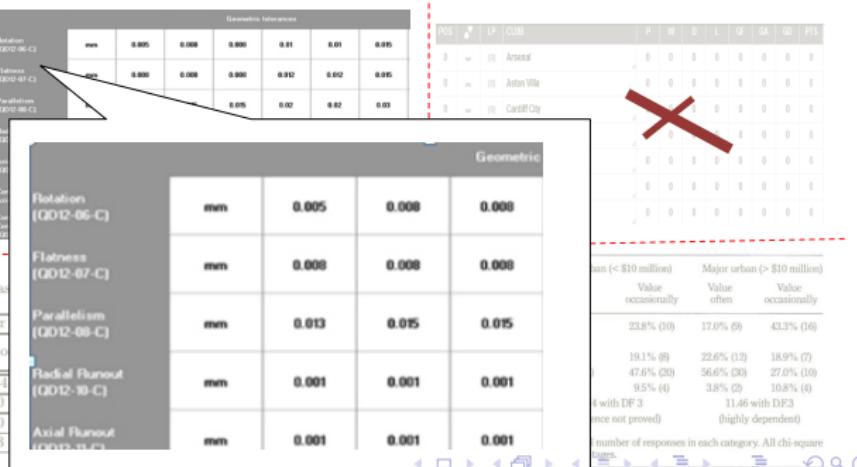


Table Interpretation – Challenges

Noise and diversity

Typical tables
addressed in the
literature

TITLE	DESCRIPTION
Kankou Temple	The temple has approximately 600 years history; en...
Unseen Temple	Unseen is a temple of the Soto school of Zen Buddhism.
Bishamon Temple	FUKUOKA's "most integrated cultural complex" of...
GOUZOKU PALACE RYUGOON	RYUGOON is a GOURMET HOTEL JAPANESE RESTAURANT...
HOTEL KOYOKAN	Welcome to Koyokan Hotel! Located in the heart of...
LODGE MASHI	GET TO THE HIRUCHI MARYAMA SKI TRAIL JUST...
Azumakan Restaurant	Hmm... <rb>ring ring image</p></p></p>
Cafe West Restaurant	Get on Kode 1 Once in year, it can't hurt <rb>ring ring image<...
Early's Restaurant	Drive on Kode 2 2004-2 MINAMI NIGATI Oh gee, there's burgen! <rb>ring ring image</p>

Classification criteria	Specification
Test Date	June 2013
Reviewed using 2013 test programme	Yes
Order	
Smartphone	Yes
Phone style	Candy Bar
Height [mm]	122.0
Width [mm]	67.0

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 Galer...	USA	USA	4,392,252

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.

Interpretation	0	0	1	0	0	0	1
Other	3	3	0	12	7	0	11

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

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

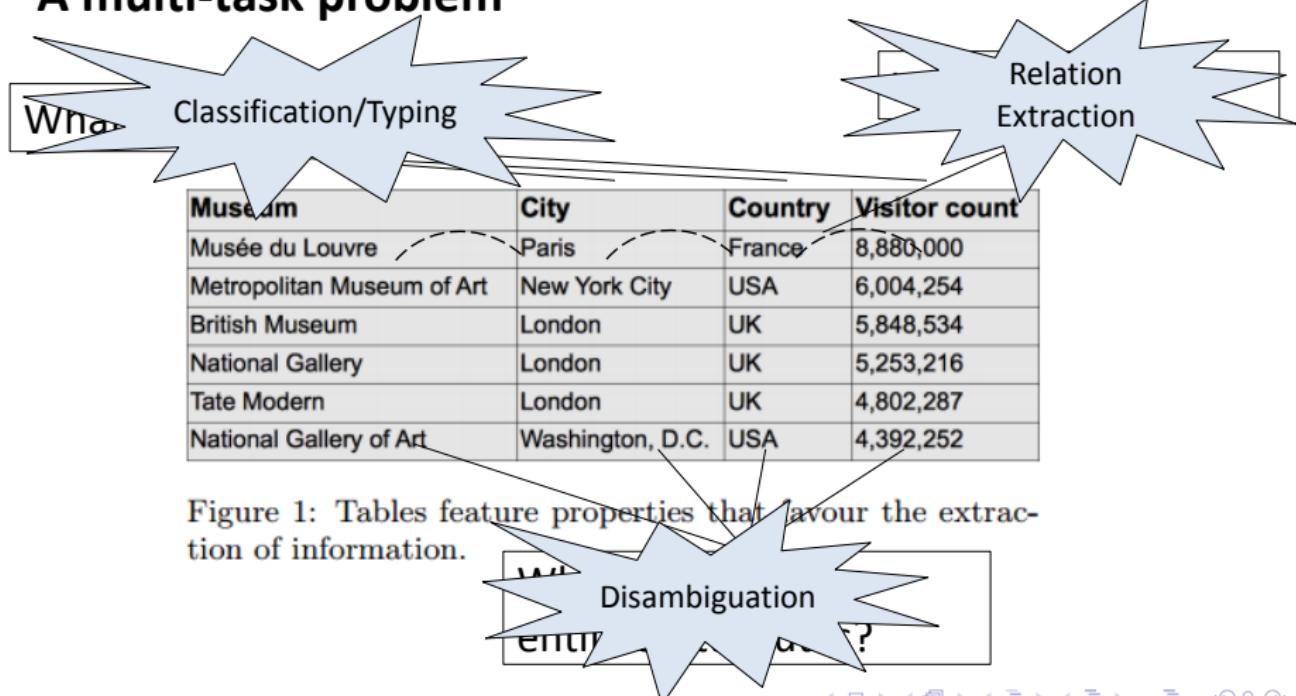


Figure 1: Tables feature properties that favour the extraction of information.

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.

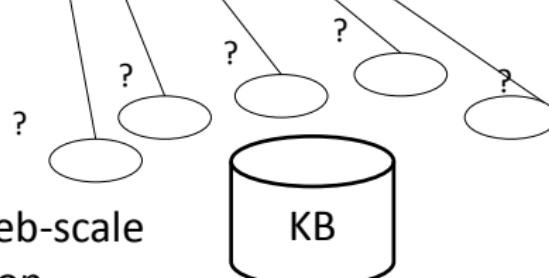


Table Interpretation

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

Table Interpretation - Methods

General workflow

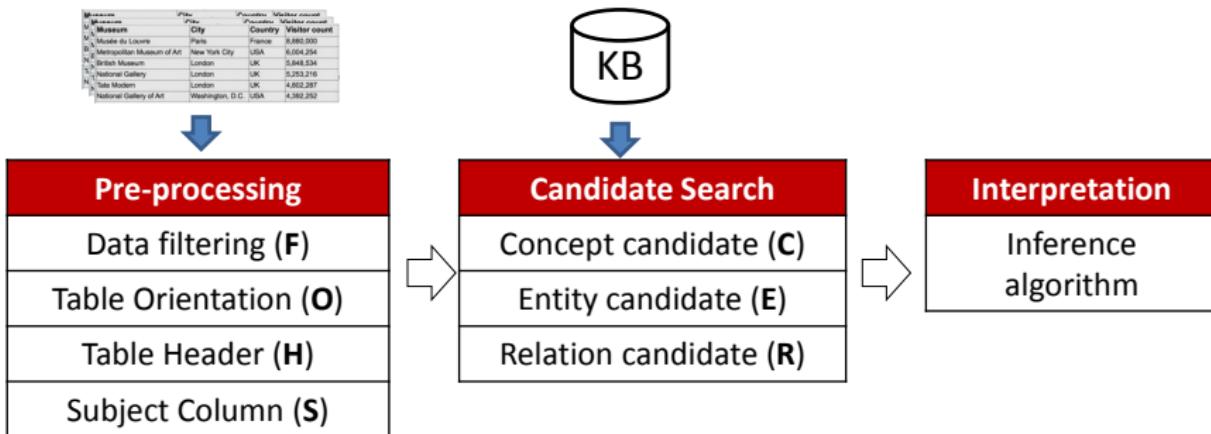


Table Interpretation - Methods

General workflow

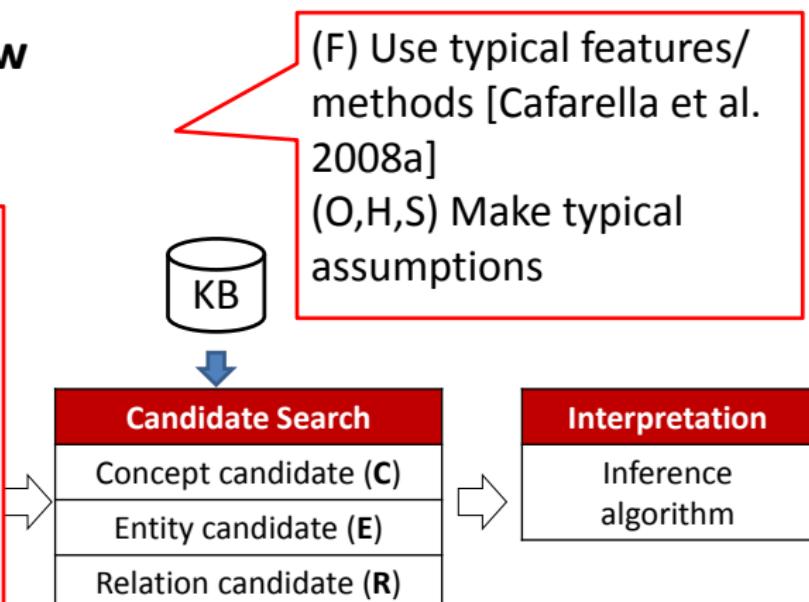
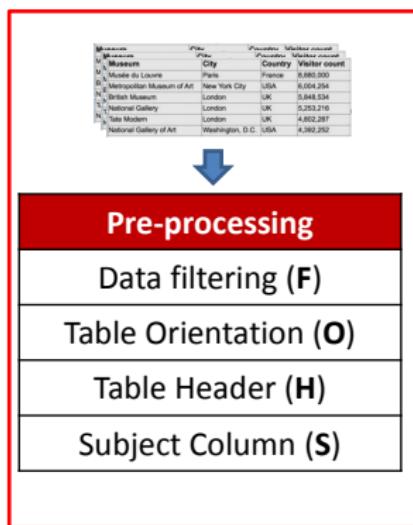


Table Interpretation - Methods

General workflow

Museum	City	Country	Visitor count
Musée du Louvre	Paris	France	8,880,000
Metropolitan Museum of Art	New York City	USA	6,044,254
British Museum	London	UK	5,944,534
National Gallery	London	UK	3,271,118
Tate Modern	London	UK	4,862,287
National Gallery of Art	Washington, D.C.	USA	4,382,252



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

For each study pointers will be given...

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	5,648,534
National Gallery	London	UK	3,278,118
Tate Modern	London	UK	4,802,287
National Gallery of Art	Washington, D.C.	USA	4,382,252

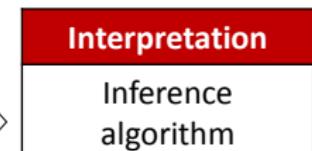
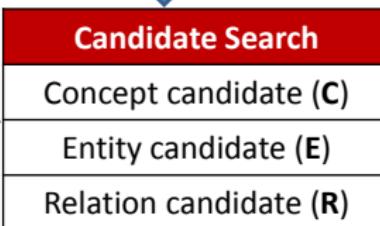
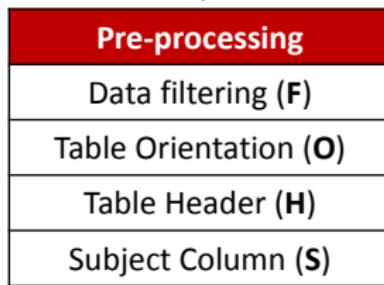


Table Interpretation - Methods

General workflow

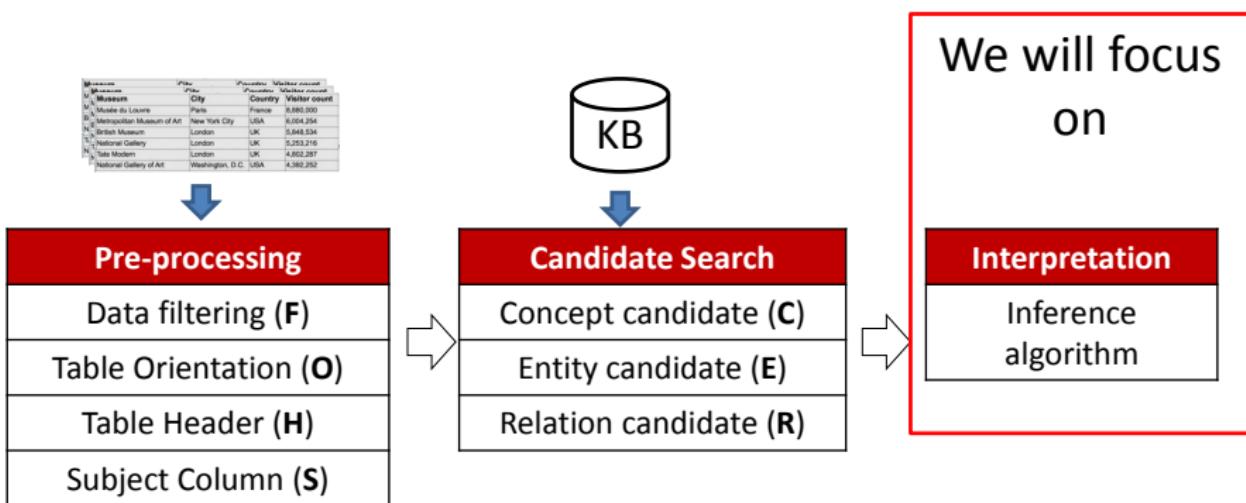


Table Interpretation - Methods

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

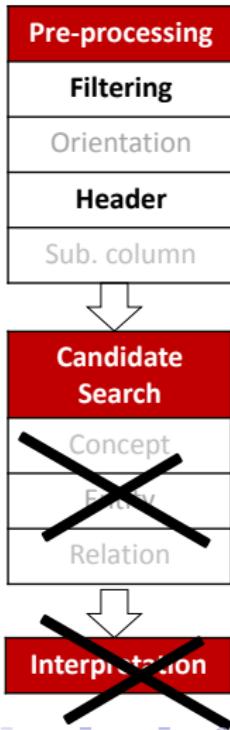


Table Interpretation - Methods

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 [List of countries and dependencies and their capitals in native ...](#)
http://en.wikipedia.org.../List_of_countries_and_dependencies_and_their_capitals_i...

Fusion Tables

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

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

Send Feedback

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.

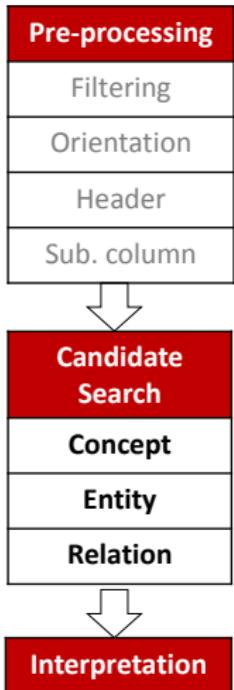


Table Interpretation - Methods

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: title¹=“Paris” & redirects¹=“Paris” &
firstSent¹=“City” & linkedConcepts¹={"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: title¹=“Paris” & redirects¹=“Paris” &
firstSent¹=“City” & linkedConcepts¹={"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

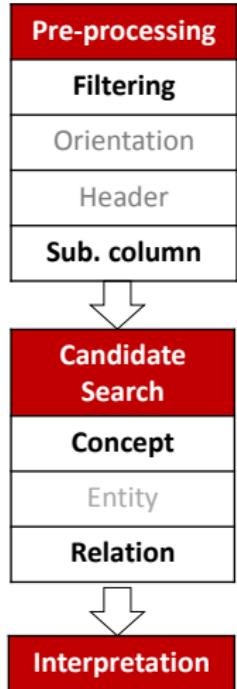


Table Interpretation - Methods

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)

$$l(A) = \arg \max_{l_i} \{ \Pr [v_1, \dots, v_n \mid l_i] \}$$

$$= \prod_j \frac{\Pr [l_i \mid v_j] \times \Pr [v_j]}{\Pr [l_i]} \propto \prod_j \frac{\Pr [l_i \mid v_j]}{\Pr [l_i]}$$

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)

$$l(A) = \arg \max_{l_i} \{ \Pr [v_1, \dots, v_n \mid l_i] \}$$

$$= \prod_j \frac{\Pr [l_i \mid v_j] \times \Pr [v_j]}{\Pr [l_i]} \propto \prod_j \frac{\Pr [l_i \mid v_j]}{\Pr [l_i]}$$

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 $r_{i,j} \in R_i$ 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 $r_{i,j} \in R_i$ having the same mention form l_i has different popularity
- Some are obscure and rare for the given mention

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

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)

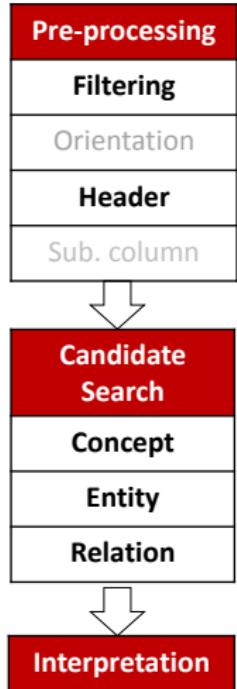
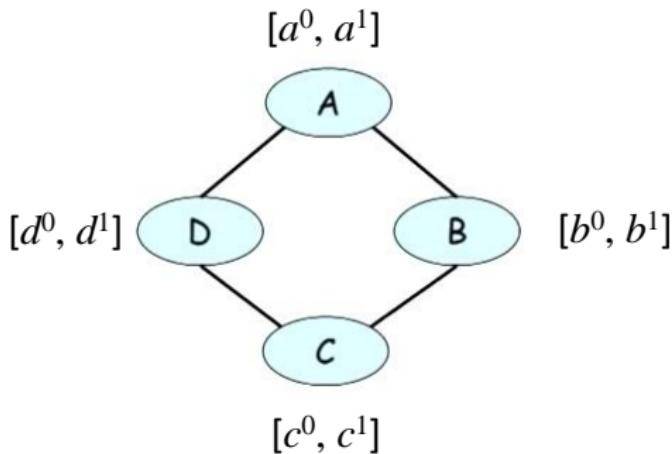


Table Interpretation - Methods

MN – a primer

A graphical representation of dependency between variables

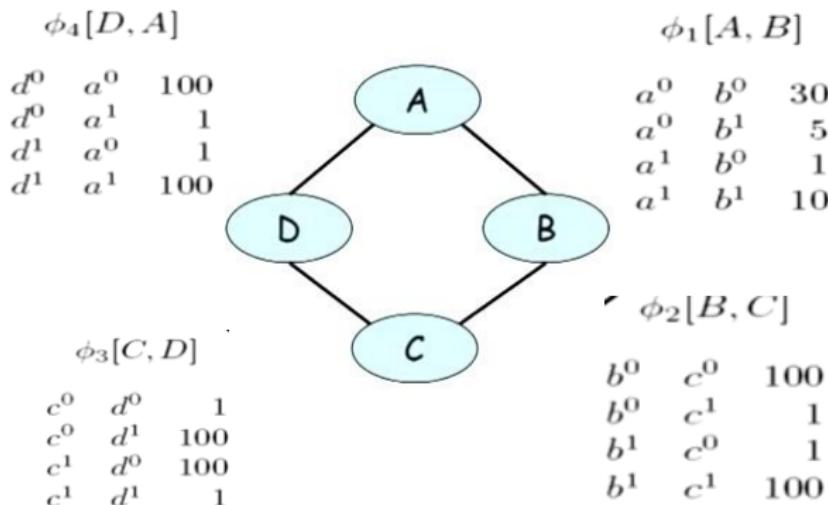


(based on <https://class.coursera.org/pgm/lecture/preview>)

Table Interpretation - Methods

MN – a primer

Factors (ϕ) to encode “compatibility” between variables

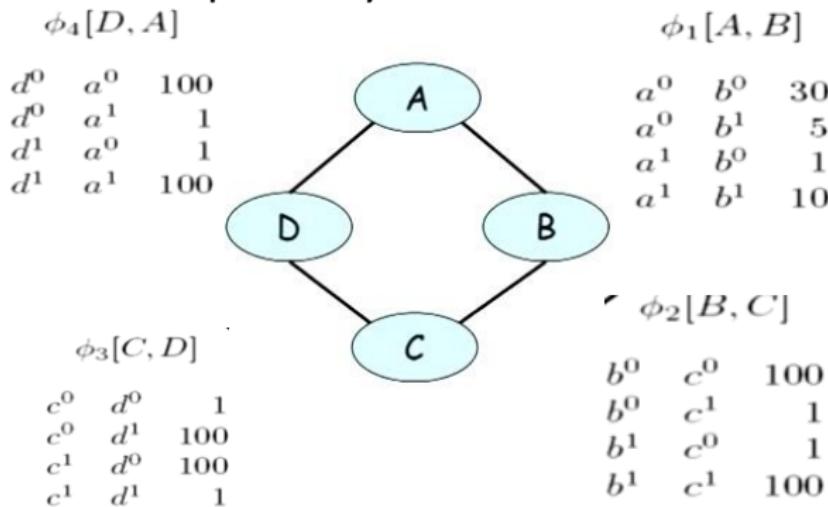


(based on <https://class.coursera.org/pgm/lecture/preview>)

Table Interpretation - Methods

MN – a primer

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



(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

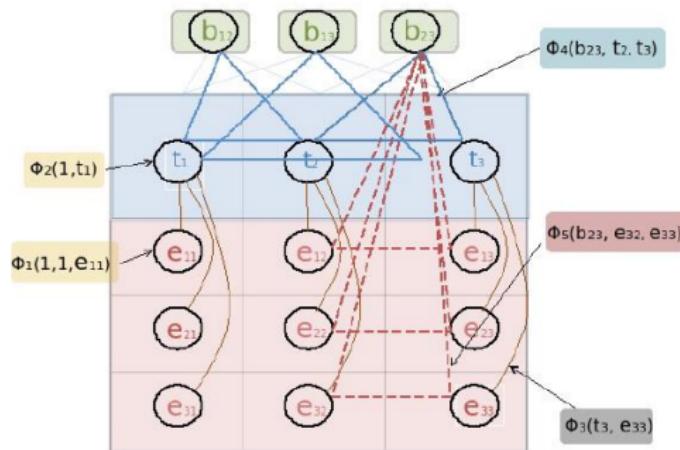
Table Interpretation - Methods

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

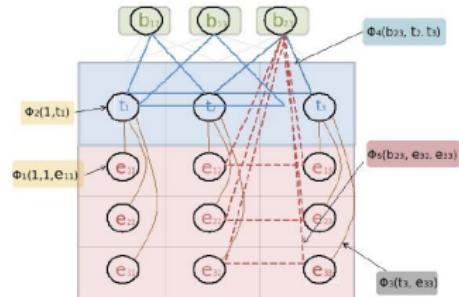
- Each column type, cell entity, pairwise column-column relation becomes a variable
- Retrieve candidates (variable values) from YAGO
- Modelling compatibility
 - Cell text and entity label
 - Column header text and type label
 - Column type and cell entity
 - Relation and pair of column types
 - Relation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

Table Interpretation - Methods

Limaye et al. [2010]

Inference

$$\max_{e, t, b} \underbrace{\prod_{c, c'} \phi_4(b_{cc'}, t_c, t_{c'}) \prod_r \phi_5(b_{cc'}, e_{rc}, e_{rc'})}_{\text{relation}} \\ \underbrace{\prod_c \phi_2(c, t_c) \prod_r \phi_1(r, c, e_{rc}) \phi_3(t_c, e_{rc})}_{\text{columns}} \underbrace{.}_{\text{cells}}$$



- 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)

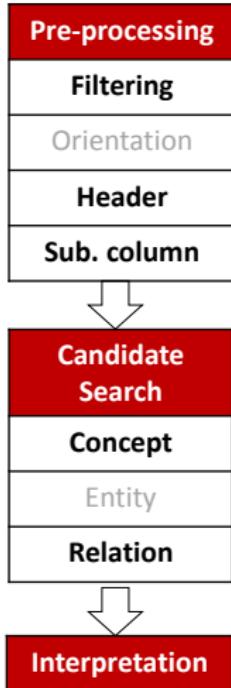


Table Interpretation - Methods

Wang et al. [2010]

KB:

- Probbase – 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 $(c, e, prob), \ e \in E$

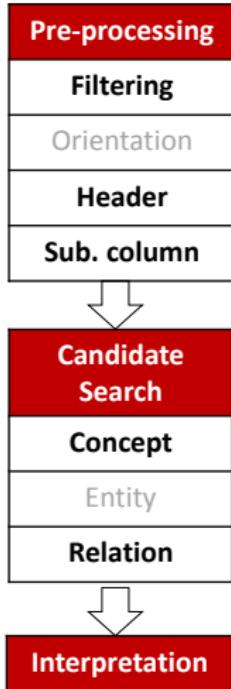


Table Interpretation - Methods

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**

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]

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

↓

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]

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

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

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
Cristiano Ronaldo	Portugal	R.Madrid	1.86m	5 February 1985	90

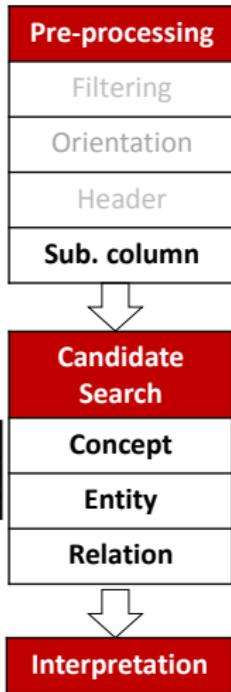


Table Interpretation - Methods

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	30
Cristiano Ronaldo	Portugal	Real Madrid	1.86m	51

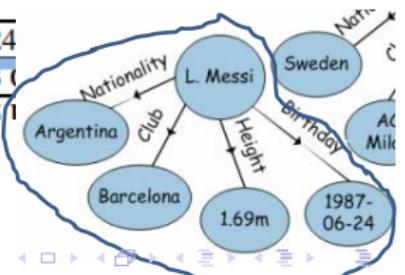


Table Interpretation - Methods

Guo et al. [2011]

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

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

Based on string similarity

Column index

Predicate index

Lionel Messi	Argentina	Barcelona	1.69m	24
Zlatan Ibrahimovic	Sweden	AC Milan	1.95m	30
Cristiano Ronaldo	Portugal	Real Madrid	1.86m	51

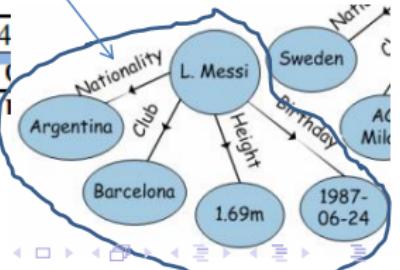


Table Interpretation - Methods

Guo et al. [2011]

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

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

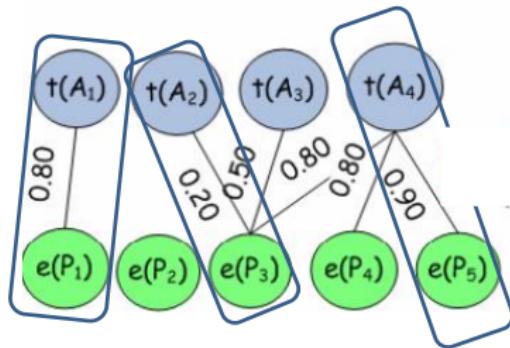


Table Interpretation - Methods

Guo et al. [2011]

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

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

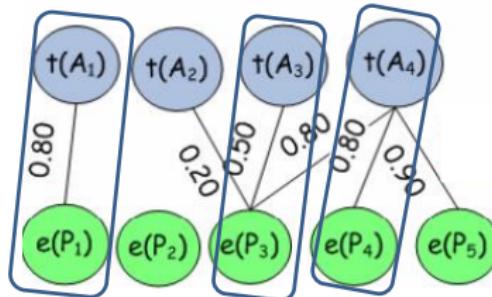


Table Interpretation - Methods

Guo et al. [2011]

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

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

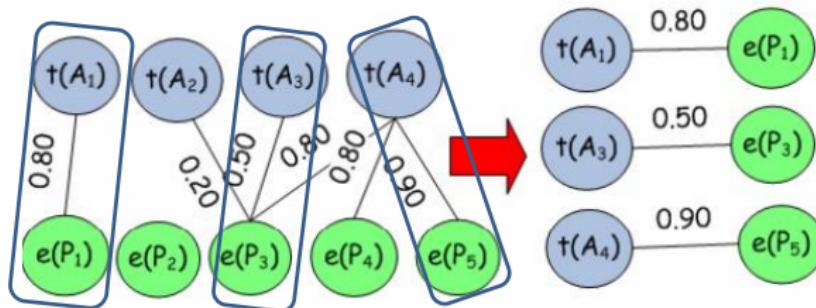


Table Interpretation - Methods

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	7 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

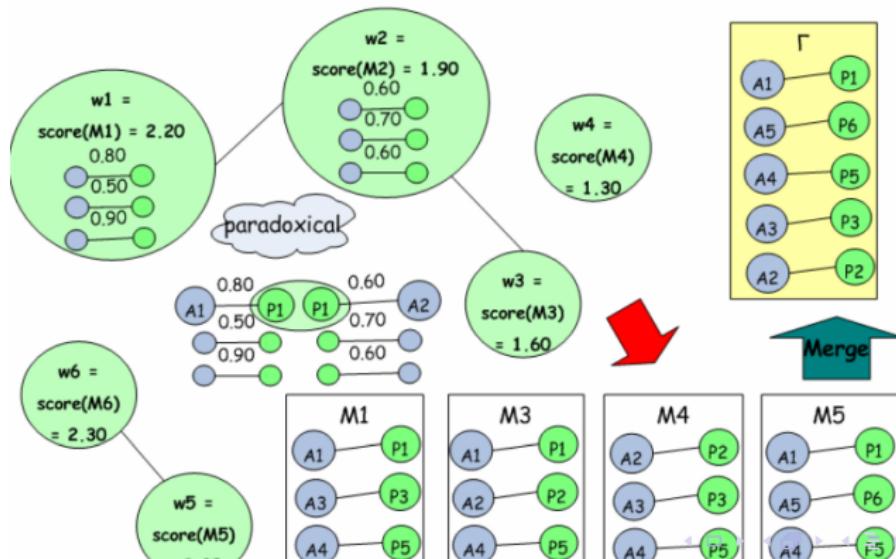


Table Interpretation - Methods

Guo et al. [2011]

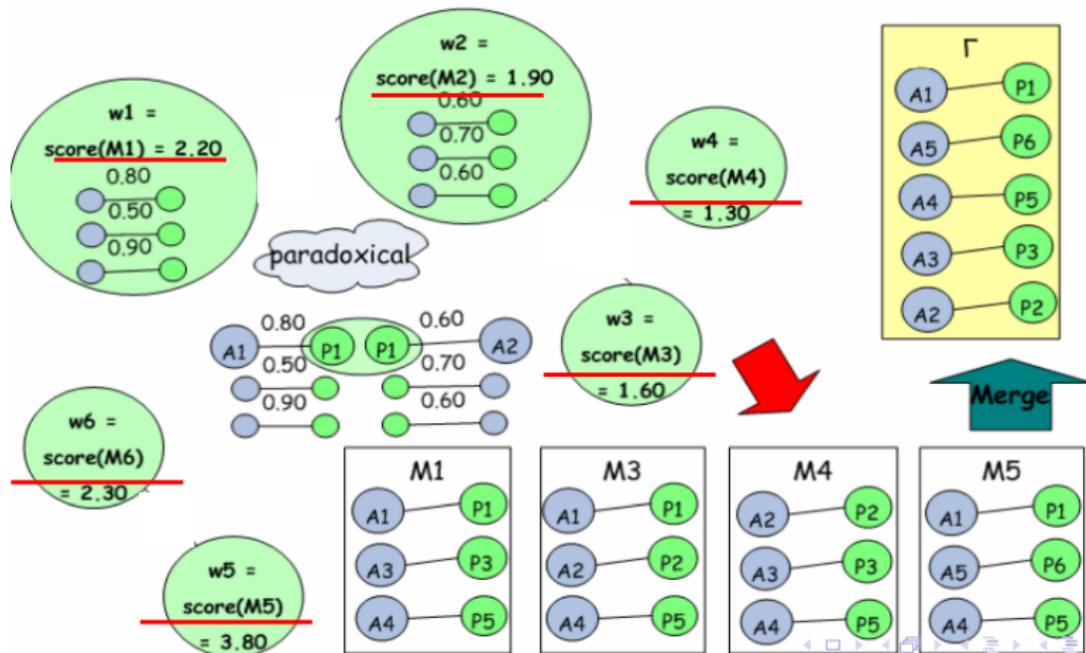


Table Interpretation - Methods

Guo et al. [2011]

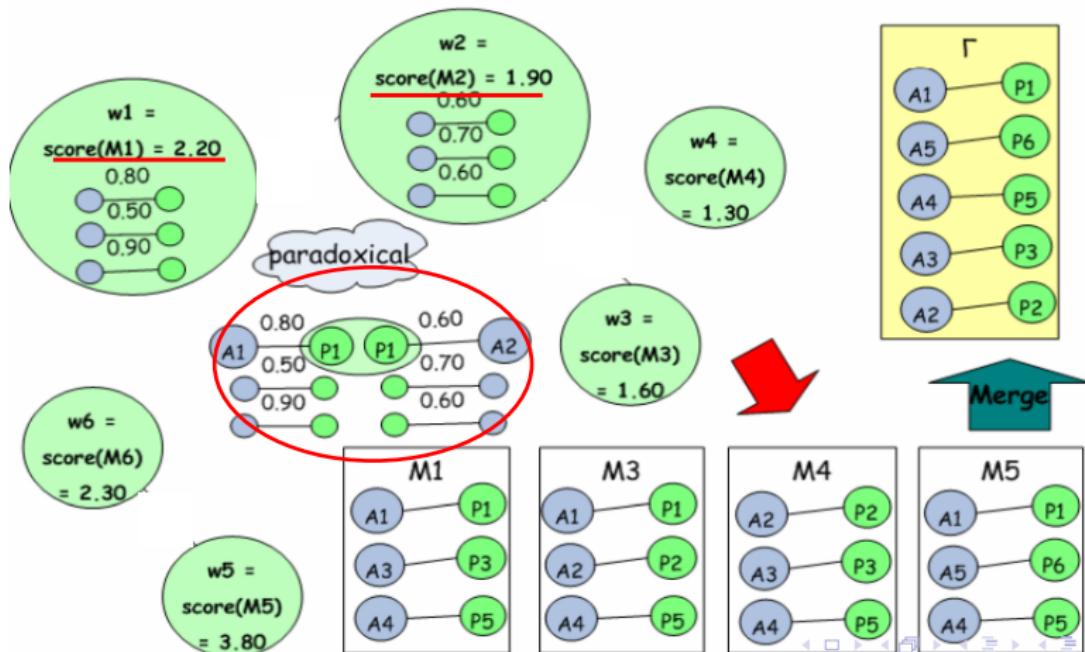


Table Interpretation - Methods

Guo et al. [2011]

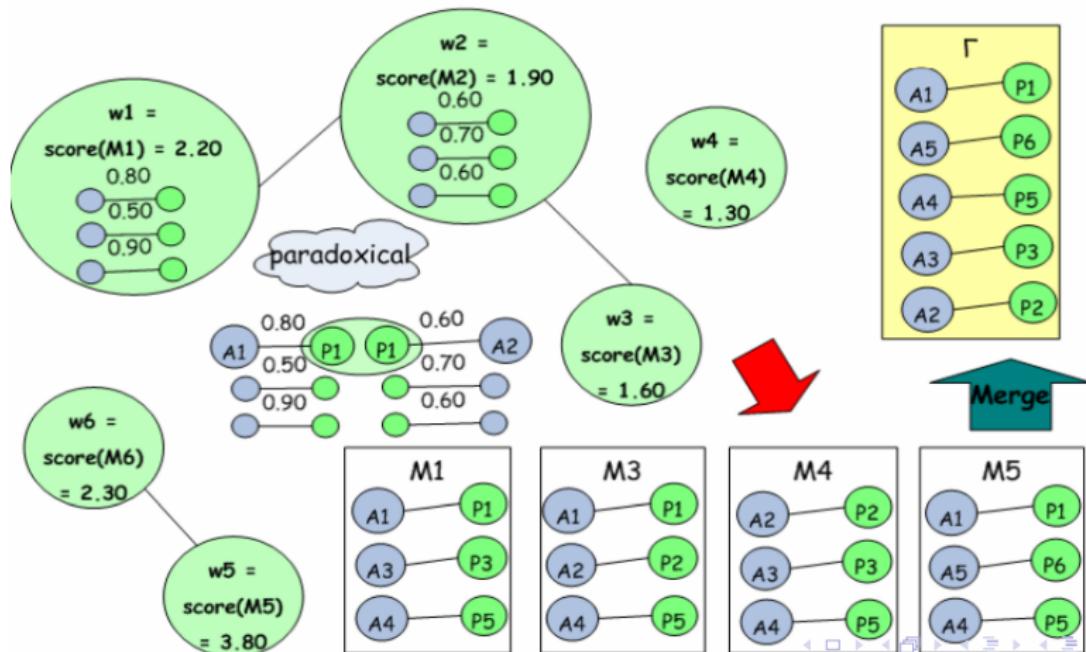


Table Interpretation - Methods

Guo et al. [2011]

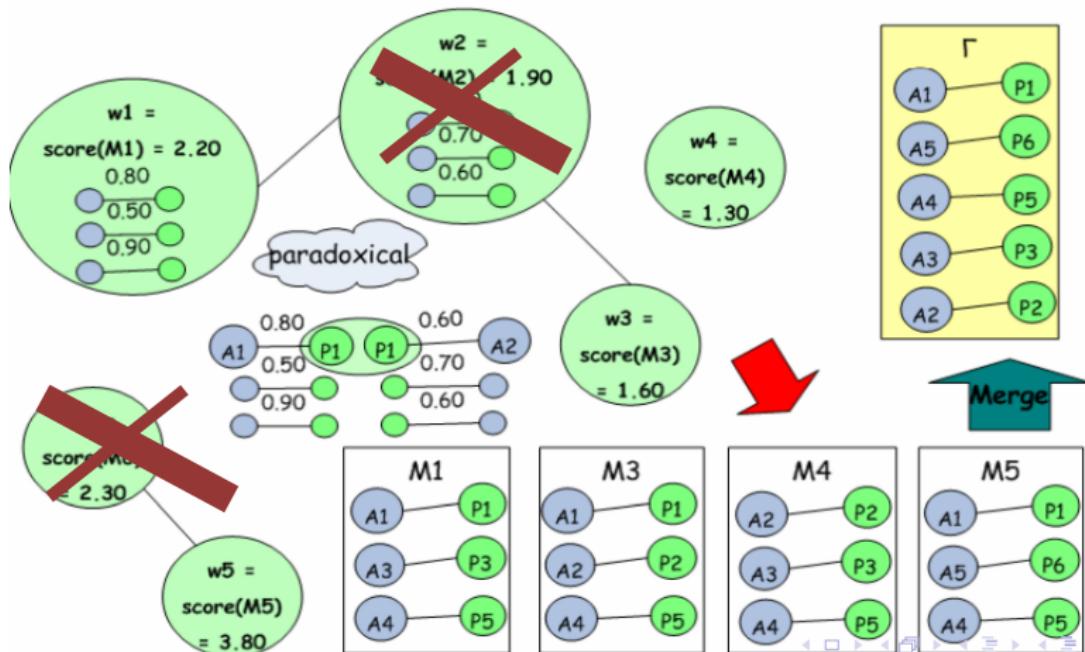


Table Interpretation - Methods

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

Yosef et al. [2011]

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

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

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_Ibl	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

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

WikiTable [Bhagavatula et al. 2013]

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 population

City	State	2010 population	Numeric decline from peak population	African American Population	Rank
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 ▪ 145,065	▪ 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]

The screenshot shows two windows from the mySimon website. The top window is titled 'mySimon: Compare products and prices in Books - Microsoft Internet Explorer' and displays a search interface for books. It includes fields for 'Title', 'Exact Title Match', 'Author', and 'Keywords'. To the right, there's a sidebar with 'books' and 'NY Times Bestsellers' links. The bottom window is titled 'mySimon: Books - Microsoft Internet Explorer' and shows the search results for 'Harry Potter'. The results list three books:

- A Comprehensive Literature Guide to Harry Potter**
Paperback | Jan 2001 | Carson-Dellosa Publishing Company, Incorporated
- Beatrix Potter to Harry Potter : Portraits of Children's Writers**
Julia Eccleshare Hardcover | Sep 2002 | National Portrait Gallery
- A Guide to the Harry Potter Novels**
Julia Eccleshare Hardcover | Apr 2002 | Continuum International Publishing Group, Incorporated

Paperback | Apr 2002 | Continuum International Publishing Group, Incorporated

At the bottom of the results page, it says '107 results total: Page 1 | 2 | 3 | 4 | 5 | 6 | ... | 8 | Next'.

Outline

1 Overview

2 Wrapper Induction

3 Table Interpretation

4 Conclusions

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

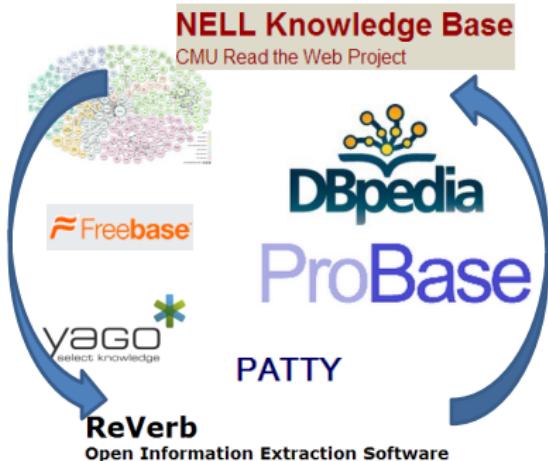
Summary & Conclusion

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



Web Texts



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.
Proceedings. 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.

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., Pașca, 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