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
  - ClueWeb12 - 870 million English web pages
- 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

\[\text{\url{http://lemurproject.org/clueweb09/}}\]
\[\text{\url{http://lemurproject.org/clueweb12/}}\]
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]
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]
Web IE Methods and Systems

- Snowball [Agichtein and Gravano, 2000]
- KnowItAll [Etzioni et al., 2004]
- OpenIE/TextRunner [Banko et al., 2007]
- ReVerb [Fader et al., 2011]
- NELL [Carlson et al., 2010]
- PROSPERA [Nakashole et al., 2011]
- Probase [Wu et al., 2012]
- LODIE [Ciravegna et al., 2012]
Web IE Methods and Systems

Snowball [Agichtein and Gravano, 2000]

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

**KnowItAll** [Etzioni et al., 2004]

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

**OpenIE/TextRunner** [Banko et al., 2007]

- Learn syntactic patterns to extract any relation instances from any domains (open IE)
- Completely unsupervised, no need for seeds
  - Input: corpus C, parser phase on a portion of C
  - pattern generation from parsed documents, t: <e₁, r, e₂>
  - 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

Diagram showing the development of IE methods from 2000 to 2013:
- Snowball (2000)
- KnowItAll (2003)
- OpenIE/Text-Runner (2005)
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](image-url)
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
Tables

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

Outline
Web IE methods

- Entity centric web pages
  - Wrapper Induction

- Tables
  - Table Interpretation

- Conclusions
Outline

1. Overview
2. Wrapper Induction
3. Table Interpretation
4. Conclusions
Wrapper Induction: definition of the task

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

- Traditional wrapper induction task
  - schema
  - set of pages output from a single script
  - training data are given as input, and a wrapper is inferred that recovers data from the pages according to the schema.

- Web-scale wrapper induction task
  - large number of sites
  - each site comprising the output of an unknown number of scripts, along with a schema
  - per-site training examples can no longer be given

[Gulhane et al., 2011]
Wrapper Induction: example

Extracting book attributes on e-commerce websites

- **Title**: Eat Pray Love: One Woman’s Search for Everything Across Italy, India and Indonesia
- **Author**: Elizabeth Gilbert
- **Date**: 2006

Anna Lisa Gentile, Ziqi Zhang
Extraction lifecycle

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

Robust methods

[Gulhane et al., 2011]
Website Clustering Problem

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

Clustering approaches:

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

Further reading survey [Gottron, 2008]
Scalable Clustering

Structurally clustering webpages for extraction

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

[Blanco et al., 2011]
Scalable Clustering

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

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

[Blanco et al., 2011]
Scalable Clustering

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

**Principle of Minimum Description Length (MDL)**

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

---

[Blanco et al., 2011]
Learning Extraction rules: Characteristics

- **Languages**
  - Grammars
  - Xpath
  - OXpath
  - Xstring [Grigalis, 2013]

- **Techniques**
  - contextual rules (boundaries detection)
  - html-aware
  - visual features
  - hybrid approaches [Zhai and Liu, 2005, Zhao et al., 2005, Grigalis, 2013]

- **Approaches**
  - supervised
  - unsupervised

- **Extraction dimensions**
  - attribute-value pairs from tables
  - record level extractor (lists) [Álvarez et al., 2008, Zhai and Liu, 2005, Zhao et al., 2005]
  - detail page extractor
Supervised methods

Training data

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

Web site specific

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

- Multi-view learner [Hao et al., 2011]
- Vertex! [Gulhane et al., 2011]
Supervised methods: Multi-view learners

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

[Hao et al., 2011]
Supervised methods: Multi-view learners

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

- Weak features
  - general across attributes, verticals and websites
  - identify a large amount of candidate attribute values
    - likely to contain noise

- Strong features
  - site-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)

[Hao et al., 2011]
Supervised methods: Multi-view learners

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

[Hao et al., 2011]
Supervised methods: Vertex!

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

---

[Gulhane et al., 2011]
Unsupervised methods

- Do not require training data

**BUT**

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

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

Schema finding problem

a. Source Dataset

<table>
<thead>
<tr>
<th>Name</th>
<th>Email</th>
<th>Title</th>
<th>Descr.</th>
<th>Books</th>
<th>Editions</th>
<th>Details</th>
<th>Year</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Smith</td>
<td>smith@..</td>
<td>DB Primer</td>
<td>This book..</td>
<td>1st Ed., P.back</td>
<td>1998</td>
<td>208</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2nd Ed., H. Cover</td>
<td>2000</td>
<td>308</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computer S.</td>
<td></td>
<td>An undergrad..</td>
<td></td>
<td>1st Ed., P.back</td>
<td>1995</td>
<td>408</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paul Jones</td>
<td>null</td>
<td>XML at..</td>
<td>A compr..</td>
<td>1st Ed., P.back</td>
<td>1999</td>
<td>308</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>HTML..</td>
<td>A useful..</td>
<td>null</td>
<td>1993</td>
<td>308</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2nd Ed., H. Cover</td>
<td>1999</td>
<td>458</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>JavaScript</td>
<td>A must in..</td>
<td>null</td>
<td>2000</td>
<td>508</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b. HTML Pages

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

[Crescenzi and Mecca, 2004]
Unsupervised methods: RoadRunner

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

[Crescenzi and Mecca, 2004]
Unsupervised methods: Yahoo!

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

---

[Dalvi et al., 2011]
Unsupervised methods: SKES

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

[He et al., 2013]
SKES - Page Representation

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

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

Page representation:

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

[He et al., 2013]
SKES - Page Representation example

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

[He et al., 2013]
SKES - wrapper extraction

INPUT: a set of HTML pages and a support threshold
OUTPUT: induced template
IDEA: counting the support of:
- tag paths
  - checking the presence of the tag path on the set of pages
- text tag paths
  - repetitive text tag paths are likely to be attributes indicators
  - unique text tag paths are likely to be data region
SKES - pros and cons

- **pros**
  - completely unsupervised
  - no knowledge required (in the form of a schema)

- **cons**
  - no semantics for the attributes

[He et al., 2013]
Unsupervised methods: LODIE Wrapper Induction

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

[Gentile et al., 2013]
LODIE Wrapper Induction: task definition

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

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

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

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

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

---

[Gentile et al., 2013]
LODIE Wrapper Induction: Dictionary Generation

- User Information Need formalisation
  - translate the concept and attributes of interest to the vocabularies used within the *Linked Data*

- given a SPARQL endpoint, query the exposed *Linked Data* to identify the relevant concepts

- select the most appropriate class and properties that describe the attributes of interest

- using the SPARQL endpoint, query the *Linked Data* to retrieve instances of the properties of interest

[Gentile et al., 2013]
LODIE Wrapper Induction: Dictionary Generation example

Find all concepts matching the keyword “university”

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

Identify all properties defined with this concept

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

Extract all available values of this attribute

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

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

Get all annotations for attribute $a_{i,n}$, as ($\langle \text{xpath}, \text{value}_{i,k} \rangle$) pairs

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

[Gentile et al., 2013]
LODIE Wrapper Induction: XPath identification

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

---

[Gentile et al., 2013]
LODIE - pros and cons

- **pros**
  - unsupervised
  - information need driven approach
    - search space for the wrappers is limited to possibly relevant portions of pages

- **cons**
  - user need definition still manual
    - BUT concept and attribute semantic is defined only once and valid to all websites of same domain

---

[Gentile et al., 2013]
Robust methods

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

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

- **Motivation**
- Problem definition
- Table Interpretation - Methods
A widely used structure for (relational) information

**Table 1. Classification results on 81 symbolic columns**

<table>
<thead>
<tr>
<th>Technique</th>
<th>Value often</th>
<th>Value occasionally</th>
<th>Technique</th>
<th>Value often</th>
<th>Value occasionally</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjusted yield approach</td>
<td>4.8% (1)</td>
<td>23.8% (10)</td>
<td>Adjusted yield approach</td>
<td>17.0% (9)</td>
<td>43.3% (16)</td>
</tr>
<tr>
<td>Cash flow</td>
<td></td>
<td></td>
<td>Cash flow</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(term and reversion approach)</td>
<td>38.1% (8)</td>
<td>19.1% (8)</td>
<td>(term and reversion approach)</td>
<td>22.6% (12)</td>
<td>189% (7)</td>
</tr>
<tr>
<td>Shortfall approach</td>
<td>52.4% (11)</td>
<td>47.6% (20)</td>
<td>Shortfall approach</td>
<td>56.6% (20)</td>
<td>27.0% (10)</td>
</tr>
<tr>
<td>Layer approach</td>
<td>4.7% (1)</td>
<td>95.4% (4)</td>
<td>Layer approach</td>
<td>3.8% (2)</td>
<td>108% (4)</td>
</tr>
<tr>
<td>Test of independence: chi square</td>
<td>5.374 with DF 3</td>
<td>11.46 with DF3 (highly dependent)</td>
<td>Test of independence: chi square</td>
<td>5.374 with DF 3</td>
<td>11.46 with DF3 (highly dependent)</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Museum</th>
<th>City</th>
<th>Country</th>
<th>Visitor count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Musée du Louvre</td>
<td>Paris</td>
<td>France</td>
<td>8,880,000</td>
</tr>
<tr>
<td>Metropolitan Museum of Art</td>
<td>New York City</td>
<td>USA</td>
<td>6,004,254</td>
</tr>
<tr>
<td>British Museum</td>
<td>London</td>
<td>UK</td>
<td>5,848,534</td>
</tr>
<tr>
<td>National Gallery</td>
<td>London</td>
<td>UK</td>
<td>5,253,216</td>
</tr>
<tr>
<td>Tate Modern</td>
<td>London</td>
<td>UK</td>
<td>4,802,287</td>
</tr>
<tr>
<td>National Gallery of Art</td>
<td>Washington, D.C.</td>
<td>USA</td>
<td>4,392,252</td>
</tr>
</tbody>
</table>
Table Interpretation – Motivation

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

A widely used structure for (relational) information

<table>
<thead>
<tr>
<th>Museum</th>
<th>City</th>
<th>Country</th>
<th>Visitor count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Musée du Louvre</td>
<td>Paris</td>
<td>France</td>
<td>8,880,000</td>
</tr>
<tr>
<td>Metropolitan Museum of Art</td>
<td>New York City</td>
<td>USA</td>
<td>6,004,254</td>
</tr>
<tr>
<td>British Museum</td>
<td>London</td>
<td>UK</td>
<td>5,848,534</td>
</tr>
<tr>
<td>National Gallery</td>
<td>London</td>
<td>UK</td>
<td>5,253,216</td>
</tr>
<tr>
<td>Tate Modern</td>
<td>London</td>
<td>UK</td>
<td>4,802,287</td>
</tr>
<tr>
<td>National Gallery of Art</td>
<td>Washington, D.C.</td>
<td>USA</td>
<td>4,392,252</td>
</tr>
</tbody>
</table>
Table Interpretation – Motivation

A widely used structure
for (relational)
information

... contains extractable, interpretable ...

Enormous data source ....

... 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

KB:
- ns:Musee du Louvre
- ns:inCity
- ns:Paris
- ns:NYC
- ns:Museum
- ns:typeOf
- ns:City
- ns:typeOf
- ns:Visitor count
- ns:Museum
- ns:typeOf
- ns:City
- ns:typeOf

Overview
Wrapper Induction
Table Interpretation
Conclusions
Table Interpretation – Definition

Input: tables

Goal: link

KB

Input tables:

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

Goal: link

ns:Museum

ns:typeOf

ns:City

ns:Musee du Louvre

ns:typeOf

ns:inCity

ns:NYC

ns:Paris
# Table Interpretation – Definition

**Input**

<table>
<thead>
<tr>
<th>Museum</th>
<th>City</th>
<th>Country</th>
<th>Visitor count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Musee du Louvre</td>
<td>Paris</td>
<td>France</td>
<td>8,880,000</td>
</tr>
<tr>
<td>Metropolitan Museum of Art</td>
<td>New York City</td>
<td>USA</td>
<td>6,004,254</td>
</tr>
<tr>
<td>British Museum</td>
<td>London</td>
<td>UK</td>
<td>5,848,534</td>
</tr>
<tr>
<td>National Gallery</td>
<td>London</td>
<td>UK</td>
<td>5,253,216</td>
</tr>
<tr>
<td>Tate Modern</td>
<td>London</td>
<td>UK</td>
<td>4,802,287</td>
</tr>
<tr>
<td>National Gallery of Art</td>
<td>Washington, D.C.</td>
<td>USA</td>
<td>4,392,252</td>
</tr>
</tbody>
</table>

**Goal:** link

KB

ns:Museum

ns:City

ns:Musee du Louvre

ns:inCity

ns:NYC

ns:Paris

ns:typeOf

ns:typeOf
Table Interpretation – Definition

Input

Goal: link

KB

ns:Museum
ns:typeOf
ns:Musee du Louvre
ns:typeOf
ns:inCity
ns:City
ns:typeOf
ns:NYC
ns:Paris
ns:inCity

Goal: link

A.L. Gentile, Z. Zhang
Table Interpretation – Definition

Overview

Wrapper Induction

Table Interpretation

Conclusions

Anna Lisa Gentile, Ziqi Zhang
Table Interpretation – Definition

Input

tables

Goal: enrich

KB

ns:Museum
ns:City
ns:Paris
ns:NYC
ns:Metropolitan MoA
ns:inCity
ns:typeOf
ns:City
ns:typeOf
ns:inCity
**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 [Cafarrella et al., 2008a]
- Relational tables
  - tables containing relational data
Table Interpretation – Challenges

Noise and diversity
Table Interpretation – Challenges

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

<table>
<thead>
<tr>
<th>TITLE</th>
<th>DESCRIPTION</th>
<th>ADDRESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Karioji Temple</td>
<td>The temple has approximately 600 years history on…</td>
<td>Uraso 257, Minamisatsuma-shi</td>
</tr>
<tr>
<td>Uonto Temple</td>
<td>Uonto is a temple of the Soto school of Zen Buddhism…</td>
<td>860 Uonto, Minamisatsuma, Nagata</td>
</tr>
<tr>
<td>Bishamun Temple</td>
<td>Fukui is a “designated cultural asset” by the city…</td>
<td>Fukui, Toyama Prefecture</td>
</tr>
<tr>
<td>Gyouyuji Palace</td>
<td>RYUGON is A TRADITIONAL JAPANESE STYLE HOTEL…</td>
<td>73, Sakado, MINAMISATSUMA-SH</td>
</tr>
</tbody>
</table>
Table Interpretation – Challenges

Noise and diversity

Very difficult to generalise one universal solution

Table 1. Classification results on 81 symbolic columns

<table>
<thead>
<tr>
<th>Method</th>
<th>Food</th>
<th>Micro</th>
<th>Resp.</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td>34</td>
<td>0</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>Computed</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>SMO</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

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

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

Table 1. Classification results on 81 symbolic columns

<table>
<thead>
<tr>
<th></th>
<th>our method using the ontology</th>
<th>SMO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Food</td>
<td>Micro.</td>
</tr>
<tr>
<td>computed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>manual</td>
<td>34</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
Table Interpretation – Challenges

Noise and diversity
Ignore “entity centric page” (consult wrapper induction)

<table>
<thead>
<tr>
<th>Museum</th>
<th>City</th>
<th>Country</th>
<th>Visitor count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Musée du Louvre</td>
<td>Paris</td>
<td>France</td>
<td>8,880,000</td>
</tr>
<tr>
<td>Metropolitan Museum of Art</td>
<td>New York City</td>
<td>USA</td>
<td>5,004,254</td>
</tr>
<tr>
<td>British Museum</td>
<td>London</td>
<td>UK</td>
<td>5,848,534</td>
</tr>
</tbody>
</table>

**Table 1. Classification**

<table>
<thead>
<tr>
<th></th>
<th>computed</th>
<th>manual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food</td>
<td>34</td>
<td>1</td>
</tr>
<tr>
<td>Micro.</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Response</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

*Note: The figures in brackets are the actual number of responses in each category. All chi-square tests are based on numbers and not percents.*
Table Interpretation – Challenges

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

Anna Lisa Gentile, Ziqi Zhang
Noise and diversity
Ignore long texts (see free text IE)
Table Interpretation – Challenges

**Noise and diversity**

Ignore numeric tables – those with many numeric values

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

Figure 1: Tables feature properties that favour the extraction of information.
Table Interpretation – Challenges

Noise and diversity

Many studies also ignore vertical tables
Table Interpretation – Challenges

Noise and diversity

Typical tables addressed in the literature

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

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

Figure 1: Tables feature properties that favour the extraction of information.
Table Interpretation – Challenges

A multi-task problem

What type?

What entities/attributes?

What relations?

Figure 1: Tables feature properties that favour the extraction of information.
A multi-task problem

Classification/Typing

Relation Extraction

Disambiguation

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

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]
Table Interpretation

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

General workflow

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

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

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

General workflow

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

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...

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

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

We will focus on
Table Interpretation - Methods

Interpretation – a closer look

<table>
<thead>
<tr>
<th>Concept</th>
<th>Relation</th>
<th>Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept_001_cityInTheUK</td>
<td>Relation_a01_capitalOf</td>
<td>Entity_1_London_UK</td>
</tr>
<tr>
<td>Concept_023_cityInTheUS</td>
<td>Relation_a87_cityOf</td>
<td>Entity_2_London_USA</td>
</tr>
<tr>
<td>Concept_125_city</td>
<td>Relation_a91_locatedIn</td>
<td></td>
</tr>
</tbody>
</table>

?’ (not found in KB)
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

Entity_1_London_UK
Entity_2_London_USA
Entity_101_MMoA
Entity_2_London_USA
A couple of highly influential work in table information extraction in general
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
**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)

<table>
<thead>
<tr>
<th>Country (exonym)</th>
<th>Capital (exonym)</th>
<th>Country (endonym)</th>
<th>Capital (endonym)</th>
<th>Official or native</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saint Barthélemy</td>
<td>Gustavia</td>
<td>Saint-Barthélemy</td>
<td>Gustavia</td>
<td>French</td>
</tr>
<tr>
<td>Saint Helena, Ascension</td>
<td>Jamestown</td>
<td></td>
<td></td>
<td>English</td>
</tr>
<tr>
<td>Saint Kitts and</td>
<td>Basseterre</td>
<td>Saint-Martin</td>
<td>Marigot</td>
<td>English</td>
</tr>
<tr>
<td>Saint Martin</td>
<td>Marigot</td>
<td>Saint-Martin</td>
<td>Marigot</td>
<td>French</td>
</tr>
<tr>
<td>Saint Lucia</td>
<td>Castries</td>
<td></td>
<td></td>
<td>English</td>
</tr>
</tbody>
</table>

List of countries and dependencies and their capitals in native...
Table Interpretation - Methods
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.
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
Table Interpretation - Methods

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

Custom query considers “context” of each value

<table>
<thead>
<tr>
<th>Museum</th>
<th>City</th>
<th>Country</th>
<th>Visitor count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Musée du Louvre</td>
<td>Paris</td>
<td>France</td>
<td>8,880,000</td>
</tr>
<tr>
<td>Metropolitan Museum of Art</td>
<td>New York City</td>
<td>USA</td>
<td>6,004,254</td>
</tr>
</tbody>
</table>

Query: \texttt{title}^1=“Paris” \& \texttt{redirects}^1=“Paris” \& \texttt{firstSent}^1=“City” \& \texttt{linkedConcepts}^1={“Musee du Louvre”, “France”, “8,880,000”} \& \texttt{infoboxPropertyValues} ={“Musee du Louvre”, “France”, “8,880,000”}

Result: \textit{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
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
Syed et al. [2010], Mulwad et al. [2010, 2011]

Custom query considers “context” of each value

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

<table>
<thead>
<tr>
<th>Museum</th>
<th>City</th>
<th>Country</th>
<th>Visitor count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Musée du Louvre</td>
<td>Paris</td>
<td>France</td>
<td>8,880,000</td>
</tr>
<tr>
<td>Metropolitan Museum of Art</td>
<td>New York City</td>
<td>USA</td>
<td>6,004,254</td>
</tr>
</tbody>
</table>
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
Venetis et al. [2011]

A sequential model:

Step 1) Label columns

- A maximum likelihood model – the best class label $l(A)$ is the one maximising the probability of the values given the class label for the column:

  $l_i$ – candidate label; $v_n$ – values in the cells of a column (except header)

$$l(A) = \arg \max_{l_i} \{ \Pr [v_1, \ldots, v_n \mid l_i] \} = \prod_j \frac{\Pr [l_i \mid v_j] \times \Pr [v_j]}{\Pr [l_i]} \propto \prod_j \frac{\Pr [l_i \mid v_j]}{\Pr [l_i]}.$$
Venetis et al. [2011]

A sequential model:

Step 1) Label columns

- A maximum likelihood model – the best class label $l(A)$ is the one maximising the probability of the values given the class label for the column:

  $l_i$ – candidate label; $v_n$ – values in the cells of a column (except header)

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

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

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

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

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

Frequency of the entity being mentioned by label $l_j$ in Wikipedia
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)

\[
LQ(r_{i,j}) = \alpha \cdot P_{pr}(r_{i,j}) + (1 - \alpha) \cdot Coh(r_{i,j})
\]

\[
LQ(M) = \alpha \cdot \sum_{s=1}^{L} P_{pr}(m_{s}) + (1 - \alpha) \cdot \sum_{s=1}^{L} Coh(m_{s})
\]

- 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)
MN – a primer
A graphical representation of dependency between variables

(based on https://class.coursera.org/pgm/lecture/preview)
MN – a primer
Factors ($\phi$) to encode “compatibility” between variables

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

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

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
Limaye et al. [2010]

Inference

\[
\max_{e,t,b} \prod_{c,c'} \phi_4(b_{cc'}, t_c, t_{c'}) \prod_r \phi_5(b_{cc'}, e_{rc}, e_{rc'}) \\
\prod_c \phi_2(c, t_c) \prod_r \phi_1(r, c, e_{rc}) \phi_3(t_c, e_{rc}).
\]

- Implementation: belief propagation
Table Interpretation - Methods

One potential limitation of these methods: candidate search

<table>
<thead>
<tr>
<th>Name</th>
<th>Birthdate</th>
<th>Political Party</th>
<th>Assumed Office</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barack Obama</td>
<td>4 Aug 1961</td>
<td>Democratic</td>
<td>2009</td>
<td>6'1</td>
</tr>
<tr>
<td>Arnold Schwarzenegger</td>
<td>30 Jul 1947</td>
<td>Republican</td>
<td>2003</td>
<td>6'2</td>
</tr>
<tr>
<td>Hillary Clinton</td>
<td>26 Oct 1947</td>
<td>Democratic</td>
<td>2009</td>
<td>5'8</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Name</th>
<th>Birthdate</th>
<th>Political Party</th>
<th>Assumed Office</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barack Obama</td>
<td>4 Aug 1961</td>
<td>Democratic</td>
<td>2009</td>
<td>6'1</td>
</tr>
<tr>
<td>Arnold Schwarzenegger</td>
<td>30 Jul 1947</td>
<td>Republican</td>
<td>2003</td>
<td>6'2</td>
</tr>
<tr>
<td>Hillary Clinton</td>
<td>26 Oct 1947</td>
<td>Democratic</td>
<td>2009</td>
<td>5'8</td>
</tr>
</tbody>
</table>

(US presidents, \{Birthday, Political Party, Assumed Office\}, 0.90)
(politicians, \{Birthday, Political Party, Assumed Office\}, 0.88)
(NBA players, \{Birthday, Height\}, 0.65)
Table Interpretation - Methods

Wang et al. [2010]

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

Wang et al. [2010]

The *row* of headers describes a specific concept

<table>
<thead>
<tr>
<th>Name</th>
<th>Birthdate</th>
<th>Political Party</th>
<th>Assumed Office</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barack Obama</td>
<td>4 Aug 1961</td>
<td>Democratic</td>
<td>2009</td>
</tr>
<tr>
<td>Arnold Schwarzenegger</td>
<td>30 Jul 1947</td>
<td>Republican</td>
<td>2003</td>
</tr>
<tr>
<td>Hillary Clinton</td>
<td>26 Oct 1947</td>
<td>Democratic</td>
<td>2009</td>
</tr>
</tbody>
</table>
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
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.

<table>
<thead>
<tr>
<th>Name</th>
<th>Birthdate</th>
<th>Political Party</th>
<th>Assumed Office</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barack Obama</td>
<td>4 Aug 1961</td>
<td>Democratic</td>
<td>2009</td>
</tr>
<tr>
<td>Arnold Schwarzenegger</td>
<td>30 Jul 1947</td>
<td>Republican</td>
<td>2003</td>
</tr>
<tr>
<td>Hillary Clinton</td>
<td>26 Oct 1947</td>
<td>Democratic</td>
<td>2009</td>
</tr>
</tbody>
</table>
Wang et al. [2010]

\[ SC_A = f(A^s) \]

\((c, a, \text{prob}), a \in A\)

\[\{\begin{array}{|l|l|l|l|}
\hline
\text{Name} & \text{Birthdate} & \text{Political Party} & \text{Assumed Office} \\
\hline
\text{Barack Obama} & 4 \text{ Aug 1961} & \text{Democratic} & 2009 \\
\text{Arnold Schwarzenegger} & 30 \text{ Jul 1947} & \text{Republican} & 2003 \\
\text{Hillary Clinton} & 26 \text{ Oct 1947} & \text{Democratic} & 2009 \\
\hline
\end{array}\}\]

1 Simplified for explanation. Consult Wang et al. for full details
Table Interpretation - Methods

Wang et al. [2010]

\[
SC_A = f(A^s)^1 \\
SC_E = g(E^{col})^1
\]

(\(c, a, \text{prob}\), \(a \in A\))

(\(c, e, \text{prob}\), \(e \in E\))

\[\text{Simplified for explanation. Consult Wang et al. for full details}\]
Table Interpretation - Methods

**Wang et al. [2010]**

\[
SC_A = f(A^s) \quad (c, a, \text{prob}), a \in A \\
SC_E = g(E^{col}) \quad (c, e, \text{prob}), e \in E
\]

<table>
<thead>
<tr>
<th>Name</th>
<th>Birthdate</th>
<th>Political Party</th>
<th>Assumed Office</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barack Obama</td>
<td>4 Aug 1961</td>
<td>Democratic</td>
<td>2009</td>
</tr>
<tr>
<td>Arnold Schwarzenegger</td>
<td>30 Jul 1947</td>
<td>Republican</td>
<td>2003</td>
</tr>
<tr>
<td>Hillary Clinton</td>
<td>26 Oct 1947</td>
<td>Democratic</td>
<td>2009</td>
</tr>
</tbody>
</table>

\[
h(s, \text{col}) = \max \left\{ sa_i \cdot se_j \mid (c_i, A^s_i, sa_i) \in SC_A, (c_j, E^{col}_j, se_j) \in SC_E, c_i = c_j \right\}
\]

\[
(f \text{inal schema, entity column}) = \arg\max_{s, \text{col}} h(s, \text{col})
\]

\(^1\) Simplified for explanation. Consult Wang et al. for full details
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 | Real 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
Table Interpretation - Methods

Guo et al. [2011]
Mapping between a tuple \((t)\) and an entity \((e)\)

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

Based on string similarity

Column index
Predicate index

<table>
<thead>
<tr>
<th>Lionel Messi</th>
<th>Argentina</th>
<th>Barcelona</th>
<th>1.69m</th>
<th>24</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zlatan Ibrahimovic</td>
<td>Sweden</td>
<td>AC Milan</td>
<td>1.95m</td>
<td>31</td>
</tr>
<tr>
<td>Cristiano Ronaldo</td>
<td>Portugal</td>
<td>Real Madrid</td>
<td>1.86m</td>
<td>51</td>
</tr>
</tbody>
</table>
Table Interpretation - Methods

Guo et al. [2011]
Optimal mapping between a tuple \( (t) \) and an entity \( (e) \)

\[ \text{score}(M_1) = 0.8 + 0.2 + 0.9 = 1.9 \]
Table Interpretation - Methods

Guo et al. [2011]
Optimal mapping between a tuple \((t)\) and an entity \((e)\)

\[
score(M_2) = 0.8 + 0.5 + 0.8 = 2.1
\]
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

<table>
<thead>
<tr>
<th>Player</th>
<th>Nationality</th>
<th>Team</th>
<th>Height (m)</th>
<th>Birthdate</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lionel Messi</td>
<td>Argentina</td>
<td>Barcelona</td>
<td>1.69m</td>
<td>24 June 1987</td>
<td>99</td>
</tr>
<tr>
<td>Zlatan Ibrahimovic</td>
<td>Sweden</td>
<td>AC Milan</td>
<td>1.95m</td>
<td>3 October 1981</td>
<td>80</td>
</tr>
<tr>
<td>Cristiano Ronaldo</td>
<td>Portugal</td>
<td>Real Madrid</td>
<td>1.86m</td>
<td>5 February 1985</td>
<td>90</td>
</tr>
<tr>
<td>Hao Junmin</td>
<td>China</td>
<td>Schalke 04</td>
<td>1.78m</td>
<td>24 March 1987</td>
<td>60</td>
</tr>
<tr>
<td>Shinji Kagawa</td>
<td>Japan</td>
<td>Dortmund</td>
<td>1.73m</td>
<td>7 March 1989</td>
<td>88</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Lionel Messi</th>
<th>Argentina</th>
<th>Barcelona</th>
<th>1.69m</th>
<th>24 June 1987</th>
<th>99</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zlatan Ibrahimovic</td>
<td>Sweden</td>
<td>AC Milan</td>
<td>1.95m</td>
<td>3 October 1981</td>
<td>80</td>
</tr>
<tr>
<td>Cristinano Ronaldo</td>
<td>Portugal</td>
<td>Real Madrid</td>
<td>1.86m</td>
<td>5 February 1985</td>
<td>90</td>
</tr>
<tr>
<td>Hao Junmin</td>
<td>China</td>
<td>Schalke 04</td>
<td>1.78m</td>
<td>24 March 1987</td>
<td>60</td>
</tr>
<tr>
<td>Shinji Kagawa</td>
<td>Japan</td>
<td>Dortmund</td>
<td>1.73m</td>
<td>17 March 1989</td>
<td>88</td>
</tr>
</tbody>
</table>
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]

\[
\begin{align*}
\text{w1} &= \text{score}(M1) = 2.20 \\
 &= 0.80 \\
&= 0.50 \\
&= 0.90 \\
\text{w2} &= \text{score}(M2) = 1.90 \\
&= 0.60 \\
&= 0.70 \\
&= 0.60 \\
\text{w3} &= \text{score}(M3) = 1.60 \\
\text{w4} &= \text{score}(M4) = 1.30 \\
\text{w5} &= \text{score}(M5) = 3.80 \\
\text{w6} &= \text{score}(M6) = 2.30
\end{align*}
\]
Table Interpretation - Methods

Guo et al. [2011]
## Table Interpretation - Methods

### Evaluation and comparison

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Evaluation methods</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Limaye</strong> Wiki manual, Web manual, Web relation, Wiki link (≈6500 tables)</td>
<td>Annotation accuracy (% of correct annotation)</td>
<td></td>
</tr>
<tr>
<td><strong>Venetis</strong> Wiki manual, Web manual, additional crawled corpus</td>
<td>Annotation accuracy, table search</td>
<td>Column labelling outperforms Limaye</td>
</tr>
<tr>
<td><strong>Shen</strong> Wiki manual, Web manual, Web list (≈ 660 lists)</td>
<td>Annotation accuracy</td>
<td>Entity linking outperforms Limaye (even using only a basic model)</td>
</tr>
</tbody>
</table>
### Table Interpretation - Methods

#### Evaluation and comparison

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Evaluation methods</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wang 69 million tables filtered from the Web</td>
<td>Table search, taxonomy expansion</td>
<td></td>
</tr>
<tr>
<td>Guo 100 Google Fusion Tables</td>
<td>Annotation accuracy</td>
<td></td>
</tr>
<tr>
<td>Syed 5 tables</td>
<td>Annotation accuracy</td>
<td></td>
</tr>
</tbody>
</table>
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

**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

DEMO: [https://d5gate.ag5.mpi-sb.mpg.de/webaida/](https://d5gate.ag5.mpi-sb.mpg.de/webaida/)

Anna Lisa Gentile, Ziqi Zhang
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]

<table>
<thead>
<tr>
<th>Airport Code</th>
<th>Organization</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>LHR</td>
<td>England</td>
<td>123.2</td>
</tr>
<tr>
<td>LGA</td>
<td>United States</td>
<td>232.12</td>
</tr>
<tr>
<td>HUU</td>
<td>Peru</td>
<td>321.7</td>
</tr>
<tr>
<td>DBO</td>
<td>Australia</td>
<td>354.64</td>
</tr>
<tr>
<td>BGY</td>
<td>Italy</td>
<td>243.8</td>
</tr>
</tbody>
</table>

Table 1. Source Table

<table>
<thead>
<tr>
<th>Airport</th>
<th>Pays</th>
<th>OR_lbl</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>LaGuardia</td>
<td>Estados Unidos</td>
<td>MS</td>
<td>201.41</td>
</tr>
<tr>
<td>Heathrow</td>
<td>Angleterre</td>
<td>Yahoo</td>
<td>90.5</td>
</tr>
<tr>
<td>٣‌٢‌٧‌٥‌٥</td>
<td>الأردن</td>
<td>Samsung</td>
<td>198</td>
</tr>
<tr>
<td>Prestwick</td>
<td>Scozia</td>
<td>GOOG</td>
<td>211.27</td>
</tr>
<tr>
<td>Beauvais</td>
<td>Frankreich</td>
<td>HP</td>
<td>55.99</td>
</tr>
</tbody>
</table>

Table 2. Target Table
Overview
Wrapper Induction
Table Interpretation
Conclusions

Table Interpretation - Methods

**Table (schema) matching and integration**

WikiTable [Bhagavatula et al. 2013]

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

Demo [http://downey-n1.cs.northwestern.edu/public/]
Table (schema) matching and integration

WikiTable [Bhagavatula et al. 2013]

<table>
<thead>
<tr>
<th>City</th>
<th>State</th>
<th>2010 population</th>
<th>Numeric decline from peak population</th>
<th>African American Population</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albany</td>
<td>New York</td>
<td>97,656</td>
<td>-37,139</td>
<td>505,200</td>
<td>43</td>
</tr>
<tr>
<td>Allegheny</td>
<td>Pennsylvania</td>
<td>N/A</td>
<td>--</td>
<td>661,839</td>
<td>30</td>
</tr>
<tr>
<td>Brooklyn</td>
<td>New York</td>
<td>N/A</td>
<td>--</td>
<td>505,200</td>
<td>43</td>
</tr>
<tr>
<td>Camden</td>
<td>New Jersey</td>
<td>77,344</td>
<td>-47,211</td>
<td>46,314</td>
<td>18</td>
</tr>
<tr>
<td>Canton</td>
<td>Ohio</td>
<td>73,007</td>
<td>-43,905</td>
<td>211,672</td>
<td>32</td>
</tr>
<tr>
<td>Dearborn</td>
<td>Michigan</td>
<td>98,163</td>
<td>-13,854</td>
<td>590,226</td>
<td>10</td>
</tr>
</tbody>
</table>
Table Interpretation - Methods

Table from the “hidden” web
[Wang 2003]
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
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

NELL Knowledge Base
CMU Read the Web Project

DBpedia
ProBase

Freebase
PATTY
ReVerb
Open Information Extraction Software

Web Texts

Figure 1. Integration of 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
6

Table Modern
National Gallery of Art
Washington, D.C.

When politicians were debating where to put the headquarters of the new Olympic Nation in use from 1944, one place stood out as a potential artificial capital of the world. New York (already an axis) was then the world's biggest city, and in the early 1940s it was the most successful economy. It was a hotbed of ideas about the future, it was a cultural center, home of movie and stage productions, etc. The idea of a city like New York, as the winter "capital" of the world, was esthetic and treasured. If the athletes were to live in the city, it was said, it would be as if the city itself was a part of the Olympic Games. The city would be as much a part of the games as Athens was to the ancient games. So when the idea of a new Olympic capital was first proposed, it was seen as a step in the right direction.
Further reading I

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

XProj : A Framework for Projected Structural Clustering of XML Documents.
pages 46–55.

Snowball: extracting relations from large plain-text collections.

Improving schema matching with linked data.
In First International Workshop On Open Data.

Finding and Extracting Data Records from Web Pages.

Open information extraction from the web.
In IJCAI, pages 2670–2676.
Further reading II

Highly efficient algorithms for structural clustering of large websites.

Fuzzy web data tables integration guided by an ontological and terminological resource.
*IEEE Transactions on Knowledge and Data Engineering*, 25(4):805–819.

Structured data on the web.
*Communications of the ACM*, 54(2):72–79.

Webtables: exploring the power of tables on the web.

Toward an architecture for never-ending language learning.

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

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

Automatic information extraction from large websites.

Wrapping-oriented classification of web pages.

Robust web extraction: an approach based on a probabilistic tree-edit model.
Proceedings of the 35th SIGMOD international conference on Management of data.

Automatic wrappers for large scale web extraction.

An analysis of structured data on the web.
Further reading IV

Using natural language to integrate, evaluate, and optimize extracted knowledge bases.
In The 3rd Workshop on Knowledge Extraction at CIKM 2013, AKBC ’13.

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

Identifying relations for open information extraction.

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.

Clustering template based web documents.
Advances in Information Retrieval, pages 40–51.
Further reading V

Towards web-scale structured web data extraction.
*Proceedings of the sixth ACM international conference on Web search and data mining - WSDM ’13, page 753.*

Web-scale information extraction with vertex.
*2011 IEEE 27th International Conference on Data Engineering, pages 1209–1220.*

Item: extract and integrate entities from tabular data to rdf knowledge base.

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.*

Scalable and noise tolerant web knowledge extraction for search task simplification.
*Decision Support Systems.*
Further reading VI

An ontology-driven annotation of data tables.

Fuzzy annotation of web data tables driven by a domain ontology.

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

A brief survey of web data extraction tools.
ACM SIGMOD Record, 31(2):84.

Extraction and integration information in html tables.
Further reading VII

Annotating and searching web tables using entities, types and relationships.

An information-theoretic definition of similarity.

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.

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.

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

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.

Yago: a core of semantic knowledge.

Knowledge harvesting in the big-data era.

Wikipedia as an ontology for describing documents.

Exploiting a web of semantic data for interpreting tables.
In *Proceedings of the Second Web Science Conference*. 
Further reading IX


Further reading


Further reading XI