From One Tree to a Forest –
A Unified Solution for Structured Web Data Extraction

Qiang Hao†‡, Rui Cai†, Yanwei Pang‡, Lei Zhang†
† Microsoft Research Asia
‡ Tianjin University
Outline

• Motivation & Challenges
• Our Solution
  – Main Idea & Framework Overview
  – Feature Extraction
  – Vertical Knowledge Learning
  – Vertical Knowledge Adaptation
• Experimental Results
• Summary
What’s Structured Data Extraction

• Extracting structured data records from web pages = identifying values of attributes

<table>
<thead>
<tr>
<th>Title</th>
<th>Author</th>
<th>Publish Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Kite Runner</td>
<td>Khaled Hosseini</td>
<td>April 2004</td>
</tr>
<tr>
<td>Mercy</td>
<td>Toni Morrison</td>
<td>2008</td>
</tr>
<tr>
<td>The Time Machine</td>
<td>H. G. Wells</td>
<td>June 30, 2004</td>
</tr>
</tbody>
</table>

a data record = attribute values of an entity
We Need Structured Data

- A **vertical** is a category of entities associated with similar **attributes** (e.g., each book has title/author/…)

**verticals**

**Books**
- Title
- Author
- Publisher
- Publish Date

**Restaurants**
- Name
- Cuisine
- Address
- Phone

**Autos**
- Model
- Price
- Engine
- Fuel Economy

**websites**

**Structured data is highly desired**

vertical search engines

knowledge bases
Challenges

- Example: Vertical = *Book*, Attribute = *Pub. Date*

The same entity

Noisy page contents

Attribute value variations across sites
Challenges (contd.)

Page layout variations across sites

Various verticals & attributes

Restaurants
- Name
- Cuisine
- Address
- Phone

Autos
- Model
- Price
- Engine
- Fuel Economy

Books
- Title
- Author
- Publisher
- Publish Date

Jobs
- Title
- Company
- Location
- Date

Movies
- Title
- Director
- Genre
- Rating

Universities
- Name

Various verticals & attributes

New Moon (The Twilight Saga, Book 2)
Stephanie Meyer (Author)

Product Summary
Format: Paperback
Publisher: Harper Collins
ISBN: 9780061852893
Buy.com Sales Rank: 9240
Dimensions: (in Inches) 8.25H x 5.5L x 1T
Pages: 206
Age Range: NA

A Journey
Tony Blair
132 132 Free Delivery
RRP: £12.99 You save: £2.99 (24%)
In Stock | Usually dispatched within 24 hours
Also available New & Used from £12.00

Tomb With a View (Paperback)
Robert Goddard

$7.00
List price: $12.99 (24%)

The Bodies Left Behind
by Jeffrey Deaver
(Hardcover - Bargain)

Reader Rating: 4 of 5 stars
Read Customer Reviews
Wanted
- Pub Date: November 2008
- Sales Rank: 211
Existing Solutions

**Manual solutions**
- **Pros:** highly accurate
- **Cons:** labor-intensive; difficult to scale up

**Semi-automatic solutions**
- **Pros:** automatically locate data in templates
- **Cons:** need to annotate semantics manually

**Automatic solutions**
- **Pros:** extract data with specified semantics
- **Cons:** need strong features and/or abundant training data

Kushmerick (PhD thesis '97)
Muslea et al. (AGENTS '99)
Soderland (Mach. Learn. '99)
Zheng et al. (KDD '07)

... Crescenzi et al. (VLDB '01)
Arasu et al. (SIGMOD '03)
Liu et al. (KDD '03)
Zhai et al. (WWW '05)

... Zhu et al. (ICML '05, KDD '06)
Carlson et al. (ECML '08)
Wong et al. (SIGIR '08 & '09)
Yang et al. (WWW '09)
...
Our Goal

• A unified solution for extracting structured data with:

**Minimal human effort**

- Label **one** seed site for each vertical → many unseen sites

**Books**
- Title
- Author
- Publisher
- Publish Date

**Autos**
- Model
- Price
- Engine
- Fuel Economy

**Jobs**
- Title
- Company
- Location
- Date

**Restaurants**
- Name
- Cuisine
- Address
- Phone

**Flexibility for verticals**

- Handle **various** verticals & attributes without redesign

**Books**
- Title
- Author
- Publisher
- Publish Date

**Jobs**
- Title
- Company
- Location
- Date
Outline

• Motivation & Challenges
• Our Solution
  – Main Idea & Framework Overview
  – Feature Extraction
  – Vertical Knowledge Learning
  – Vertical Knowledge Adaptation
• Experimental Results
• Summary
Our Solution: Main Idea

Flexible for various verticals & attributes

Robust to variations across websites

General Features + Loose Classifiers

Site-Level Constraints

Web pages are generated by site-level templates

Recall↑

Combine

Precision↑
Framework Overview

(a) Feature Extraction
- Layout
- Content
- Context

(b) Vertical Knowledge Learning
- Attribute-Specific Semantics Learning
- Inter-Attribute Layout Learning

(c) Vertical Knowledge Adaptation
- Page-Level Semantic Prediction
- Inter-Page Aggregation
- Inter-Attribute Re-ranking

A labeled seed site

A new unseen site

Data Extraction

Structured Data

Wrappers
Feature Extraction

- Layout
- Content
- Context
General Features of Web Pages

- Extract features from **text nodes** in **DOM trees** of web pages

![Web Page Diagram]

Three types of features: *layout, content, context*
Layout Features

• Goal: characterize the **position** of a text node

Visual Position = (24, 798)

**Visual Position**

position in a rendered page
= coordinates to the top left

**DOM Path**

position in a DOM tree
= root-to-leaf tag path

DOM Path = /html/body/div/div/div/div/div/h1/em/a/text

The Kite Runner
by Khaled Hosseini
(Paperback - Reprint)

Reader Rating: ★★★★★ (1377 ratings)
> Read customer reviews  Write a Review

- Pub. Date: April 2004
- 400pp
- Sales Rank: 644

Web Page
Content Features

- Derived from the **value** contained in a text node

  - **Unigram**: Set of unique tokens

  - **Length**: Number of tokens/characters

  - **Character Type**: Proportion of letters/digits/symbols

  - **Page Redundancy**: Proportion of pages containing text node with the same value

  **Extracted at page level**

  **Site-level statistics**

Example:
- Compared with specific features like (for **Price**) **Contains** (‘$’, ‘.’, ‘0-9’), more flexible for any mixture of characters

Example:
- In **Restaurant** vertical, **Cuisines** are much redundant than **Names**

**General** enough to characterize various attributes
Context Features

- **Motivation**
  - Surrounding text indicates semantics of text nodes
  - Text nodes with identical context $\rightarrow$ similar semantics

- **Preceding Text**
  - Extracted at page level
  - Values of visually preceding text nodes

- **Prefix & Suffix**
  - Site-level statistics
  - Static (across pages)
  - Sub-strings of text node values

Illustrative example of prefix/suffix extraction
Illustrative example of prefix/suffix extraction

Motivation

- Surrounding text indicates semantics of text nodes
- Text nodes with identical context → similar semantics
Vertical Knowledge Learning

- Attribute-Specific Semantics
- Inter-Attribute Layout
Features to Vertical Knowledge

• Goal
  – Learn knowledge from a labeled seed site based on features extracted from text nodes
  – Guide data extraction from unseen sites

• Two types of vertical knowledge
Attribute-Specific Semantics

- **Content** features → *classifiers* (e.g., SVMs)
- **Context** features → (token-score) *lookup tables*

Text nodes for attribute $a_i$

Content Features

train

Classifier $C(a_i)$

lookup tables

Context Features

build

Lookup Table $T(a_i)$

Semantics of attribute $a_i$

Predict semantic relevance to $a_i$ for new text nodes
Inter-Attribute Layout

- Construct a $K \times K$ layout matrix from layout features
- Encode pairwise distances between $K$ attributes

Example: layout matrices from 5 websites in Book vertical

Tend to be close:
- title
- author
- ISBN-13
- publisher
- publish-date

Help verify combinations of attributes

$K \times K$ layout matrix
(darker $\rightarrow$ closer)
Vertical Knowledge Adaptation

- Page-Level Semantic Prediction
- Inter-Page Aggregation
- Inter-Attribute Re-ranking
Page-Level Semantic Prediction

Attribute-specific semantics $\rightarrow$ page-level candidates

Attributes: A B C

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Page 1</th>
<th>Page 2</th>
<th>Page 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
</tr>
<tr>
<td>B</td>
<td>0.9</td>
<td>0.9</td>
<td>0.7</td>
</tr>
<tr>
<td>C</td>
<td>0.4</td>
<td>0.4</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Page 1: 0.3
Page 2: 0.5
Page 3: 0.5
Inter-Page Aggregation

- For each attribute: multiple candidates per page →

<table>
<thead>
<tr>
<th>Page 1</th>
<th>Page 2</th>
<th>Page 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔️ 0.7</td>
<td>✔️ 0.8</td>
<td>✔️ 0.9</td>
</tr>
<tr>
<td>✗ 0.9</td>
<td>✗ 0.1</td>
<td>✗ 0.2</td>
</tr>
</tbody>
</table>

Ground-truth

False alarms

Site-level “page”

align & aggregate

- True attribute occurrences
- Infrequent noise
- Occasional false prediction
Inter-Attribute Re-ranking

- Multiple possible solutions (attribute combinations)

Solution 1

Solution 2

Solution 3

Inter-attribute layout learnt from the seed site

Re-rank candidates by measuring similarity
Summary: Flowchart of Features

Content Features → Attribute-Specific Semantics → Page-Level Semantic Prediction

Context Features → Inter-Page Alignment → Inter-Page Aggregation

Layout Features → Inter-Attribute Layout → Inter-Attribute Re-ranking
Outline

• Motivation & Challenges
• Our Solution
  – Main Idea & Framework Overview
  – Feature Extraction
  – Vertical Knowledge Learning
  – Vertical Knowledge Adaptation
• Experimental Results
• Summary
A Large-Scale Dataset

(Publicly available at [http://swde.codeplex.com](http://swde.codeplex.com))

- **8 verticals** with diverse semantics
- **80 websites** (10 per vertical)
- **124,291 pages** (200~2,000 per website)
- **32 attributes** (3~5 per vertical) with labeled ground-truth

<table>
<thead>
<tr>
<th>Vertical</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Autos</strong></td>
<td>model, price, engine, fuel_economy</td>
</tr>
<tr>
<td><strong>Books</strong></td>
<td>title, author, ISBN, publisher, pub_date</td>
</tr>
<tr>
<td><strong>Cameras</strong></td>
<td>model, price, manufacturer</td>
</tr>
<tr>
<td><strong>Jobs</strong></td>
<td>title, company, location, date</td>
</tr>
<tr>
<td><strong>Movies</strong></td>
<td>title, director, genre, rating</td>
</tr>
<tr>
<td><strong>NBA players</strong></td>
<td>name, team, height, weight</td>
</tr>
<tr>
<td><strong>Restaurants</strong></td>
<td>name, address, phone, cuisine</td>
</tr>
<tr>
<td><strong>Universities</strong></td>
<td>name, phone, website, type</td>
</tr>
</tbody>
</table>
Experimental Settings

- Methods
  <Baselines>
  1. **SSM** (Stacked Skews Model) *Carlson et al. ECML’08*
  2. **PL** (page-level semantic prediction)
  3. **PL + IP** (inter-page aggregation)
  <Full solution>
  4. **PL + IP + IA** (inter-attribute re-ranking)

- One seed site (by turns), test on other sites

- Performance metrics: precision & recall
Performance

![Graphs showing performance metrics for Autos, Books, Movies, and NBA Players.](image)
Performance (contd.)

- **Cameras**
  - Precision: SSM 0.53, PL 0.77, PL+IP 0.94, PL+IP+IA 0.96
  - Recall: SSM 0.80, PL 0.89, PL+IP 0.90, PL+IP+IA 0.94

- **Jobs**
  - Precision: SSM 0.65, PL 0.72, PL+IP 0.84, PL+IP+IA 0.87
  - Recall: SSM 0.51, PL 0.71, PL+IP 0.81, PL+IP+IA 0.84

- **Restaurants**
  - Precision: SSM 0.69, PL 0.83, PL+IP 0.95, PL+IP+IA 0.97
  - Recall: SSM 0.69, PL 0.83, PL+IP 0.92, PL+IP+IA 0.94

- **Universities**
  - Precision: SSM 0.73, PL 0.74, PL+IP 0.85, PL+IP+IA 0.86
  - Recall: SSM 0.73, PL 0.73, PL+IP 0.80, PL+IP+IA 0.81
Performance: Multiple Seed Sites

- Our solution with multiple seed sites
  - Take the solution with highest confidence score
- Our solution with (one seed + bootstrapping seeds)
- SSM with multiple seed sites

Table 3: Average F-scores of the proposed solution based on multiple seed sites.

<table>
<thead>
<tr>
<th>#Seeds</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Our Solution</strong></td>
<td>0.843</td>
<td>0.860</td>
<td>0.868</td>
<td>0.884</td>
<td>0.886</td>
</tr>
<tr>
<td><strong>Our Solution (Bootstrap)</strong></td>
<td>0.843</td>
<td>0.856</td>
<td>0.861</td>
<td>0.859</td>
<td>0.865</td>
</tr>
<tr>
<td><strong>SSM</strong></td>
<td>0.630</td>
<td>0.645</td>
<td>0.692</td>
<td>0.719</td>
<td>0.741</td>
</tr>
</tbody>
</table>
Summary

• A unified solution for structured data extraction
  – Minimal human effort: labeling one site per vertical
  – Flexible for various verticals & attributes

• A large-scale dataset (has been published online)
  – 124K pages from 80 websites in 8 verticals

• Promising performance
  – Precision ≥ 80%, Recall ≥ 80% for most verticals

• Future work
  – Bootstrapping: accumulate vertical knowledge incrementally
Thank you!

Dataset available at
http://swde.codeplex.com
Dataset available at http://swde.codeplex.com