WebRelate: Integrating Web Data with Spreadsheets using Examples

Jeevana Inala  
MIT

Rishabh Singh  
Microsoft Research
<table>
<thead>
<tr>
<th>Company</th>
<th>Stock value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSFT</td>
<td></td>
</tr>
<tr>
<td>AMZN</td>
<td></td>
</tr>
<tr>
<td>AAPL</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Company</td>
<td>Stock value</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>MSFT</td>
<td>$57.90</td>
</tr>
<tr>
<td>AMZN</td>
<td></td>
</tr>
<tr>
<td>AAPL</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Company</td>
<td>Stock value</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>MSFT</td>
<td>57.90</td>
</tr>
<tr>
<td>AMZN</td>
<td></td>
</tr>
<tr>
<td>AAPL</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td></td>
</tr>
</tbody>
</table>
Code:

Public Sub Import_Yahoo_Finance_Historical()

    Dim URL As String
    Dim dateParams As String

    ' Date ranges from default earliest Yahoo start date (m/d/y) to current date (m/d/y)
    dateParams = "&a=0&b=3&c=1977&d=1 & " & &e=" & Day(Date) & "&f=" & Year(Date)

    ' Daily prices
    URL = "http://ichart.finance.yahoo.com/table.csv?s=" & Sheets("Analysis").Range("C2").Value & dateParams & "&g=d&ignore=.csv"

    With Worksheets("Input")
        With .QueryTables.Add(Connection:="TEXT;" & URL, Destination:=.Range("A1"))
            .QueryFileStartRow = 1
            .TextFileParseType = xlDelimited
            .TextFileCommaDelimiter = True
            .Refresh BackgroundQuery:=False
        End With
        .QueryTables(1).Delete
    End With

    ' Dividends only

    With Worksheets("Input")
        With .QueryTables.Add(Connection:="TEXT;" & URL, Destination:=.Range("I1"))
            .TextFileStartRow = 1
        End With
    End With

End Sub
### Company | Stock value
--- | ---
MSFT | 57.90
AMZN | 759.48
AAPL | 108.51
T | 40.91
S | 6.04
Demo
Overview

MSFT  AMZN  AAPL  TWTR

Spreadsheet

URL examples

URL synthesizer

http://.../msft
http://.../amzn
http://.../aapl

List of URLs

Selected data nodes

57.90  756.90  107.90  ...

Desired data

Selected data

Web-Data Synthesizer

\( P_u \)  \( P_w \)
Related Work

- DataXFormer [Abedjan et.al., CIDR 2015]
- WebCombine [Chaisins et.al., WWW 2015]
- Vegemite [Lin et.al., IUI 2009]

The diagram illustrates a trade-off between user effort and expressiveness. Points are plotted along a two-dimensional plane where:

- More expressive is on the vertical axis.
- Less user effort is on the horizontal axis.
- More user effort is on the horizontal axis.

Points are placed according to their characteristics:

- DataXFormer: Less user effort, less expressive.
- WebCombine: More user effort, more expressive.
- Vegemite: More user effort, less expressive.
Related Work

- DataXFormer [Abedjan et.al., CIDR 2015]
- WebRelate
- WebCombine [Chaisins et.al., WWW 2015]
- Vegemite [Lin et.al., IUI 2009]
Related Work

• Learning string transformations
  • Flash Fill (Gulwani, POPL 11), Blink Fill (Singh, VLDB 16)

Rishabh Singh  R.S.
Related Work

• Learning string transformations
  • Flash Fill (Gulwani, POPL 11), Blink Fill (Singh, VLDB 16)

Rishabh Singh \[\rightarrow\] R.S.

0 \[\rightarrow\] 1 \[\rightarrow\] 2 \[\rightarrow\] 3 \[\rightarrow\] 4

Version space algebra
Related Work

• Learning string transformations
  • Flash Fill (Gulwani, POPL 11), Blink Fill (Singh, VLDB 16)

Rishabh Singh → R.S. (0 1 4)

Constant (R.S.)

Version space algebra
Related Work

- Learning string transformations
  - Flash Fill (Gulwani, POPL 11), Blink Fill (Singh, VLDB 16)
Overview

- Output-constrained PBE
- Layered Version Space Algebra
- Input-dependent Web Extraction
Overview

• **Output-constrained PBE**

• Layered Version Space Algebra

• Input-dependent Web Extraction
Traditional PBE

Ana Trujillo 357 21th Place SE, Redmond, WA

Output-constrained PBE

Ana Trujillo 357 21th Place SE, Redmond, WA

Charlie Gunaja 732 Memorial Drive, Cambridge, MA
Traditional PBE

Ana Trujillo 357 21th Place
SE, Redmond, WA

Output-constrained PBE (O-PBE)

Ana Trujillo 357 21th Place
SE, Redmond, WA

Charlie Gunaja 732 Memorial Drive, Cambridge, MA

Temperature: 42°
Output-constrained PBE (O-PBE)

• Generalization constraint
  • \( \forall \text{ in } \in \text{ inputs. } \quad p(\text{in}) \in \text{ list of possible of outputs for in} \)

• Uniqueness constraint
  • \( \forall (\text{in, out}) \in \text{ input-output examples. } \quad p(\text{in}) = \text{out} \)
Richer class of problems

Ana Trujillo 357 21th Place SE, Redmond, WA

Richer class of problems

Ana Trujillo 357 21th Place
SE, Redmond, WA


Richer class of problems

Ana Trujillo 357 21th Place
SE, Redmond, WA


Richer class of problems

Ana Trujillo 357 21th Place
SE, Redmond, WA

Richer class of problems

Ana Trujillo 357 21th Place
SE, Redmond, WA


It is sufficient to learn a program that uniquely identifies the desired output.
Better ranking

Ana Trujillo 357 21th Place SE, Redmond, WA

....redmond-wa/98052/weather-forecast/341347

How do we know which part of the URL should be .*?
Better ranking

Ana Trujillo 357 21th Place SE, Redmond, WA


....redmond-wa/98052/weather-forecast/341347
Better ranking

Ana Trujillo 357 21th Place SE, Redmond, WA


....redmond-wa/98052/weather-forecast/341347
Better ranking

Ana Trujillo 357 21th Place SE, Redmond, WA

Charlie Gunaja 732 Memorial Drive, Cambridge, MA


….redmond-wa/98052/weather-forecast/341347
Overview

- Output-constrained PBE
- **Layered Version Space Algebra**
- Input-dependent Web Extraction
URLs are long!

- Complexity of VSA used in Flash Fill
  - Quadratic in length of output
  - Exponential in number of examples

Input
Ana Trujillo 357 21th Place
SE, Redmond, WA

Output
redmond-wa/98052
0 1 7 8 10 11 16
Layered Version Space Algebra

• Performs search over increasingly expressive sub-languages
  \[ L_1 \subseteq L_2 \subseteq \ldots \subseteq L_k \]

Ana Trujillo 357 21th Place
SE, Redmond, WA

Layer 1

Layer 2

Layer 3

Layer 4 – Full dag

redmond-wa/98052
0 1 7 8 9 10 11 12 13 14 15 16
Overview

• Output-constrained PBE

• Layered Version Space Algebra

• **Input-dependent Web Extraction**
Web-Extraction

Can be expressed in Xpath query language
# Web-Extraction

<table>
<thead>
<tr>
<th>Cur1</th>
<th>Cur2</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>USD</td>
<td>INR</td>
<td>28, December, 2017</td>
</tr>
<tr>
<td>EUR</td>
<td>GBP</td>
<td>03, January, 2018</td>
</tr>
<tr>
<td>USD</td>
<td>CHF</td>
<td>05, January, 2018</td>
</tr>
</tbody>
</table>

## USD/INR

<table>
<thead>
<tr>
<th>USD/INR</th>
<th>Price</th>
<th>Change (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>63.275</td>
<td>-0.095 (-0.15%)</td>
<td></td>
</tr>
</tbody>
</table>

#### USD/INR Historical Data

<table>
<thead>
<tr>
<th>Date</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 05, 2018</td>
<td>63.340</td>
</tr>
<tr>
<td>Jan 04, 2018</td>
<td>63.400</td>
</tr>
<tr>
<td>Jan 03, 2018</td>
<td>63.505</td>
</tr>
<tr>
<td>Jan 02, 2018</td>
<td>63.460</td>
</tr>
<tr>
<td>Jan 01, 2018</td>
<td>63.680</td>
</tr>
<tr>
<td>Dec 31, 2017</td>
<td>63.840</td>
</tr>
<tr>
<td>Dec 28, 2017</td>
<td>64.080</td>
</tr>
<tr>
<td>Dec 27, 2017</td>
<td>64.120</td>
</tr>
</tbody>
</table>

Transform(28, December, 2017)
Input-dependent Web-Extraction

<table>
<thead>
<tr>
<th>Cur1</th>
<th>Cur2</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>USD</td>
<td>INR</td>
<td>28, December, 2017</td>
</tr>
<tr>
<td>EUR</td>
<td>GBP</td>
<td>03, January, 2018</td>
</tr>
<tr>
<td>USD</td>
<td>CHF</td>
<td>05, January, 2018</td>
</tr>
</tbody>
</table>

Transform(28, December, 2017)
Input-dependent Web-Extraction

USD; INR; 28, December, 2017

Input

HTML tree

Find constraints in this tree that satisfy both Uniqueness and Generalization properties of the O-PBE problem
Input-dependent Web-Extraction

USD; INR; 28, December, 2017

Input

Find constraints in this tree that satisfy both Uniqueness and Generalization properties of the O-PBE problem
**Input-dependent** Web-Extraction

USD; INR; 28, December, 2017

**Input**

```
<table>
<thead>
<tr>
<th>text</th>
<th>Dec 28, 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>USD; INR;</td>
<td>28, December, 2017</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>tr</th>
<th>Dec 28, 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>USD; INR;</td>
<td>28, December, 2017</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>td</th>
<th>Dec 28, 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>USD; INR;</td>
<td>28, December, 2017</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>td</th>
<th>Dec 28, 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>USD; INR;</td>
<td>28, December, 2017</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>text</th>
<th>Dec 28, 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>USD; INR;</td>
<td>28, December, 2017</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>tr</th>
<th>Dec 28, 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>USD; INR;</td>
<td>28, December, 2017</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>td</th>
<th>Dec 28, 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>USD; INR;</td>
<td>28, December, 2017</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>td</th>
<th>Dec 28, 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>USD; INR;</td>
<td>28, December, 2017</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>text</th>
<th>Dec 28, 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>USD; INR;</td>
<td>28, December, 2017</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>tr</th>
<th>Dec 28, 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>USD; INR;</td>
<td>28, December, 2017</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>td</th>
<th>Dec 28, 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>USD; INR;</td>
<td>28, December, 2017</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>td</th>
<th>Dec 28, 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>USD; INR;</td>
<td>28, December, 2017</td>
</tr>
</tbody>
</table>
```

Transform(28, December, 2017)
Results
Expressive?

- 88 data integration scenarios
  - Stocks, weather, sports, currency, ....
  - 62 URL learning tasks
  - 88 Web-data extraction tasks
- 5 – 32 number of input rows in the spreadsheet
- Solves all of them correctly
Fast?

URL learning

All tasks take < 1s
Layered search beats VSA

URL learning

- Layered search
- VSA
Benchmarks

Web-extraction learning

Fast?

All tasks take < 10s
Easy to use?

URL learning

85% tasks take only 1 example
Impact of generalization constraint

URL learning

- Output-constrained ranking
- Basic ranking

# examples

Benchmarks

30%
Easy to use?

Web-extraction learning

95% tasks take only 1 example
Impact of generalization constraint

Web-extraction learning

- **Output-constrained**
- **Non output-constrained**

The diagram shows the number of examples (# examples) plotted against the number of benchmarks. The graph compares the performance of output-constrained and non-output-constrained methods.
Summary

• Output-constrained PBE

• Layered Version Space Algebra

• Input-dependent Web Extraction

Thank You!

jinala@mit.edu