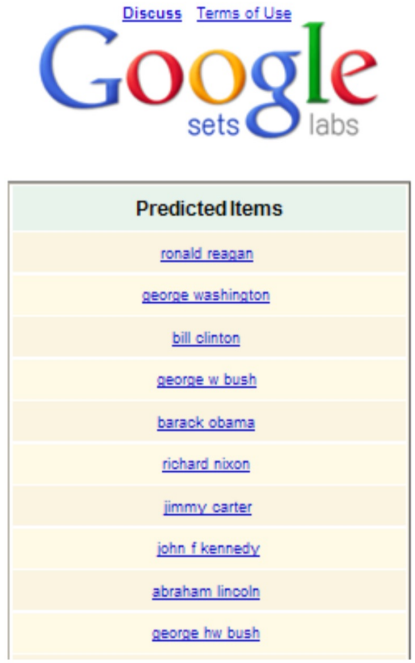
Set of T-uples Expansion by Example

A. Sanjaya, T. Abdessalem, S. Bressan

November 23, 2016

Motivation



- Google introduced Googlet Set.
- Given <*George Washington*>, <*Richard Nixon*> ➔ returned other US presidents.

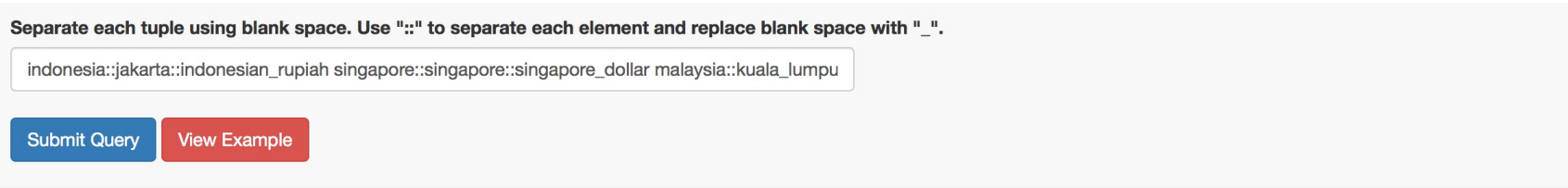
Only considered **ATOMIC** values!

Related Works

- Set Expansion
  - DIPRE [1]
    - ★ Extract attribute-value pairs.
    - ★ Few examples → find occurrences → generate pattern → new books.
  - SEAL [2],
    - ★ Generate pattern for each document.
    - ★ Introduce ranking of candidates.

Set of T-uples Expansion

- We extend to the general case of composite seeds and **n-ary** relations.
- Given <*Indonesia*, *Jakarta*, *Indonesian Rupiah*>, <*Singapore*, *Singapore*, *Singapore Dollar*>, <*Malaysia*, *Kuala Lumpur*, *Malaysian Ringgit*>



No.	Weight	Attr1	Attr2	Attr3
1	0.001988416331633271	Benin	Porto-Novo	CFA Franc BCEAO
2	0.001988416331633271	Slovenia	Ljubljana	Euro
3	0.001988416331633271	Marshall Islands	Majuro	US Dollar
4	0.001988416331633271	Hungary	Budapest	Hungarian Forint
5	0.001988416331633271	Malaysia	Kuala Lumpur	Malaysian Ringgit

- The approach consists of crawling, wrapper generation, candidate extraction, ranking.

Crawling

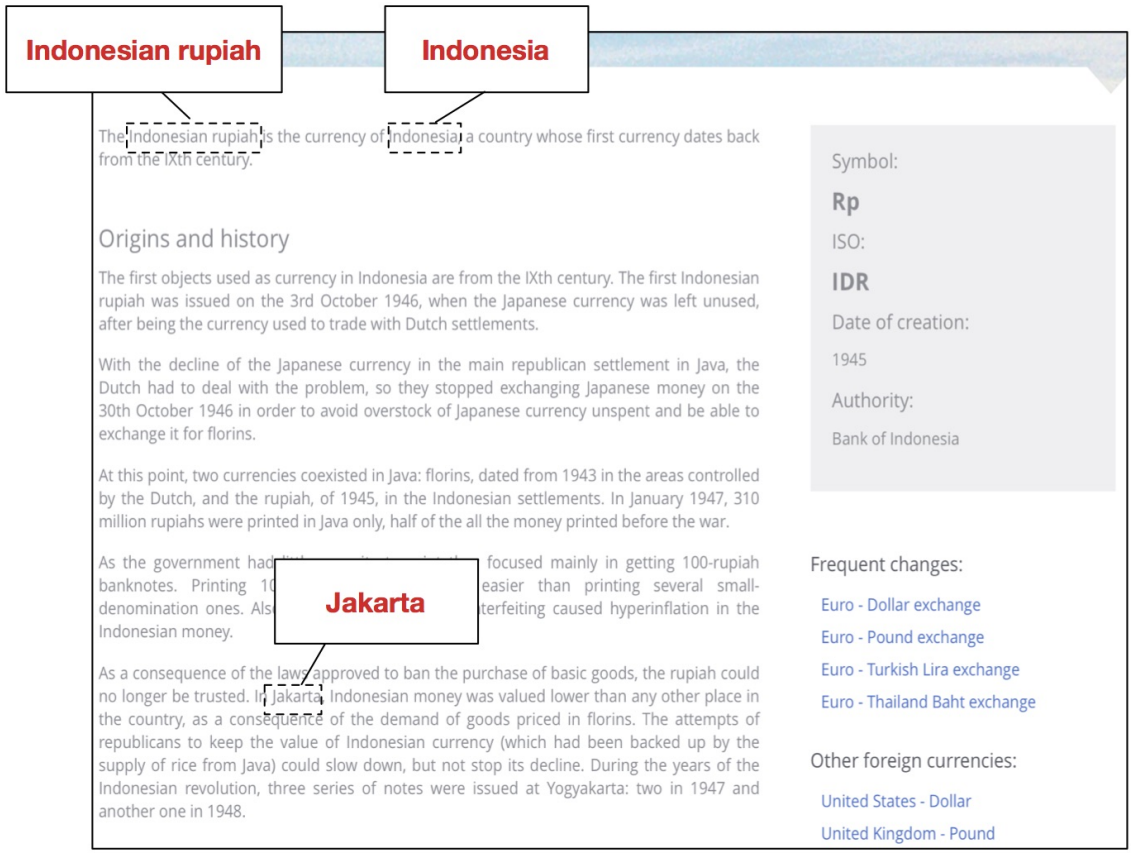
- We rely on Google search engine to collect web pages.
- The search query is the concatenation of the sets of examples given by the user.
- For the set of seeds <*IDR*, *Indonesia*, *Jakarta*>, <*CYN*, *China*, *Beijing*>, the input query for Google is '"IDR" + "Indonesia" + "Jakarta" + "CYN" + "China" + "Beijing"'.

Wrapper Generation

- Input: set of t-uple seeds T, each with $\boldsymbol{n}$ elements and set of documents $\boldsymbol{D}$.
- For each Web page $\boldsymbol{w}$ in $\boldsymbol{D}$:
  - For each t-uple t in T:
    - ★ Find the occurrences in w.
    - ★ Generate left, right and middle context for each occurrence.
  - For pairs of left and right context:
    - ★ Do character wise comparison for pairs of left and right context.
  - For pairs of middle context:
    - ★ Induce common regular expression for pairs of middle context.
- Wrapper = Left longest common string + *n-1* common regular expressions + Right longest common string

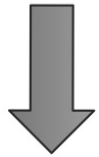
Permutation of Elements in a T-uple

- Given seed <*Indonesia*, *Jakarta*, *Indonesian Rupiah*>
- Also consider finding the occurrence of its permutation.
  - <*Indonesian Rupiah*, *Indonesia*, *Jakarta*>
  - <*Indonesia*, *Indonesian Rupiah*, *Jakarta*>

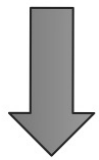


Candidate Extraction

<tr><td>(.*?)</td><td>(.*?)</td><td>(.*?)</td><td>



```
<h2><a name=A />A</a></h2>
<table class="db  tablesorter" cellspacing="1" cellpadding="2" >
<thead> <tr><th>Country</th><th>Capital</th><th>Currency Name</th><th> Currency Code</th></tr> </thead>
<tbody>
<tr><td>Afghanistan</td><td>Kabul</td><td>Afghanistan Afghani</td><td>AFN</td></tr>
<tr><td>Albania</td><td>Tirana</td><td>Albanian Lek</td><td>ALL</td></tr>
<tr><td>Algeria</td><td>Algiers</td><td>Algerian Dinar</td><td>DZD</td></tr>
<tr><td>American Samoa</td><td>Pago Pago</td><td>US Dollar</td><td>USD</td></tr>
<tr><td>Andorra</td><td>Andorra</td><td>Euro</td><td>EUR</td></tr>
<tr><td>Angola</td><td>Luanda</td><td>Angolan Kwanza</td><td>AOA</td></tr>
<tr><td>Anguilla</td><td>The Valley</td><td>East Caribbean Dollar</td><td>XCD</td></tr>
<tr><td>Antarctica</td><td>None</td><td>East Caribbean Dollar</td><td>XCD</td></tr>
<tr><td>Antigua and Barbuda</td><td>St. Johns</td><td>East Caribbean Dollar</td><td>XCD</td></tr>
<tr><td>Argentina</td><td>Buenos Aires</td><td>Argentine Peso</td><td>ARS</td></tr>
<tr><td>Armenia</td><td>Yerevan</td><td>Armenian Dram</td><td>AMD</td></tr>
<tr><td>Aruba</td><td>Oranjestad</td><td>Aruban Guilder</td><td>AWG</td></tr>
<tr><td>Australia</td><td>Canberra</td><td>Australian Dollar</td><td>AUD</td></tr>
<tr><td>Austria</td><td>Vienna</td><td>Euro</td><td>EUR</td></tr>
<tr><td>Azerbaijan</td><td>Baku</td><td>Azerbaijan New Manat</td><td>AZN</td></tr>
</tbody> </table>
```



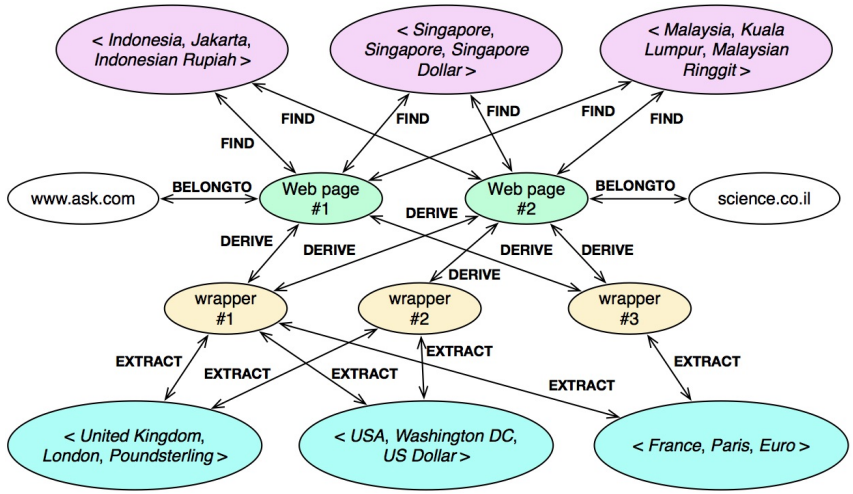
<Afghanistan, Kabul, Afghanistan Afghani>

...

<Azerbaijan, Baku, Azerbaijan New Manat>

Ranking Mechanism

Source Node	Relation	Target Node
seeds	*find*	Web page
Web page	*derive* *belongto* $find^{-1}$	wrapper domain seeds
wrapper	*extract* $derive^{-1}$	candidate Web page
candidate	$extract^{-1}$	wrapper
domain	$belongto^{-1}$	Web page



- Define entities and relations between them.
- Build graph and do random walk on graph.

Can produce a ranking list of entities.

Performance Evaluation

- 11 topics for performance evaluation, 2 to 4 seeds for each topic.
- We manually construct ground truth from Google and Google Tables.
- Exclude Web pages used to contruct ground truth in the experiment.

List of Topics

Topic Name	Seeds
D1 - Airports	<*London Heathrow Airport*, *London*> <*Charles De Gaulle International Airport*, *Paris*> <*Schipol Airport*, *Amsterdam*>
D2 - Universities	<*Massachusetts Institute of Technology (MIT)*, *United States*> <*Stanford University*, *United States*> <*University of Cambridge*, *United Kingdom*>
D3 - Car brands	<*Chevrolet*, *USA*> <*Daihatsu*, *Japan*> <*Kia*, *Korea*>
D4 - US agencies	<*ARB*, *Administrative Review Board*> <*VOA*, *Voice of America*>
D5 - Rock bands	<*Creep*, *Radiohead*> <*Black Hole Sun*, *Soundgarden*> <*In Bloom*, *Nirvana*>
D6 - MLM	<*mary kay*, *usa*> <*herbalife*, *usa*> <*amway*, *usa*>
D7 - Olympic	<*1896*, *Athens*, *Greece*> <*1900*, *Paris*, *France*> <*1904*, *St Louis*, *USA*>
D8 - FIFA player	<*2015*, *Lionel Messi*, *Argentina*> <*2014*, *Cristiano Ronaldo*, *Portugal*> <*2007*, *Kaka*, *Brazil*> <*1992*, *Marco van Basten*, *Netherlands*>
D9 - US governor	<*Rick Scott*, *Florida*, *Republican*> <*Andrew Cuomo*, *New York*, *Democratic*>
D10 - Currency	<*China*, *Beijing*, *Yuan Renminbi*> <*Canada*, *Ottawa*, *Canadian Dollar*> <*Iceland*, *Reykjavik*, *Iceland Krona*>
D11 - Formula 1	<*1990*, *Ayrton Senna*, *McLaren*> <*2000*, *Michael Schumacher*, *Ferrari*> <*2010*, *Sebastian Vettel*, *Red Bull*>

Metrics

- Precision and recall for the top-k results.
- Let R be the result lists of the system and G is the ground truth:

$$p = \frac{\sum_{i=1}^{|R|} Entity(i)}{|R|}; r = \frac{\sum_{i=1}^{|R|} Entity(i)}{|G|} \quad (1)$$

- *Entity(i)* is a binary function.

Precision and Recall

- Topic D1 (Airports), D3 (Car brands), D4 (US Agencies), D10 (Currency) have a minimum precision of 0.78, while other topics receive low score due to various reasons (different spelling, incomplete reference, ambiguous seeds).
- The general recall is more than 0.5 except for topic D2 (Universities), D4 (US agencies), D5 (Rock bands) because lack of Web pages returned by search engine, heterogeneous ground truth.

Discussion

- Challenges:
  - Different spelling.
  - Incomplete or heterogeneous ground truth.
  - Multifaceted seeds.
- Elements permutation in t-uple seeds for wrapper generation has little affect on the precision and recall of the system.
- Not excluding Web pages used as ground truth does not greatly increase the precision and recall of the system.

Conclusion and Future works

- The system is efficient, effective and practical.
- How to leverage ontological information.
- Additional semantics in the form of integrity constraints, such as candidate keys, admissible values and ranges, and dependencies.

References

1. S. Brin. Extracting patterns and relations from the world wide web. In Selected Papers from the International Workshop on The World Wide Web and Databases, WebDB '98, pages 172 - 183, London, UK, UK, 1999. SpringerVerlag.
2. R. C. Wang and W. W. Cohen. Language-independent set expansion of named entities using the web. In Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, ICDM '07, pages 342 - 350, Washington, DC, USA, 2007. IEEE Computer Society.

Precision

Data		Top-K						
		10	25	50	100	200	300	400
D1 - Airports	OR	1.0	1.0	1.0	0.99	0.985	0.98	0.984 (441)
	PW	1.0	1.0	1.0	0.99	0.98	0.98	0.984 (441)
D2 - Universities	OR	0.7	0.44	0.3	0.24	0.13	0.1	0.08 (473)
	PW	0.7	0.4	0.26	0.23	0.135	0.1	0.07 (542)
D3 - Car brands	OR	0.9	0.84	0.92	0.78 (87)	0.78 (87)	0.78 (87)	0.78 (87)
	PW	0.9	0.84	0.84	0.76	0.75 (102)	0.75 (102)	0.75 (102)
D4 - US agencies	OR	1.0	1.0	0.96	0.97	0.935	0.943	0.945 (332)
	PW	1.0	1.0	0.98	0.94	0.94	0.95	0.945 (332)
D5 - Rock bands	OR	0.2	0.28	0.32	0.32	0.19	0.156	0.156 (319)
	PW	0.2	0.28	0.34	0.3	0.225	0.186	0.133 (1813)
D6 - MLM	OR	0.6	0.52	0.66	0.59	0.365	0.403	0.39 (330)
	PW	0.6	0.44	0.28	0.35	0.36	0.243	0.182 (884)
D7 - Olympic	OR	0.9	0.56	0.44	0.23	0.135	0.135 (200)	0.135 (200)
	PW	0.9	0.64	0.44	0.22	0.11	0.073	0.044 (624)
D8 - FIFA player	OR	0.2	0.24	0.12	0.07	0.075	0.069 (215)	0.069 (215)
	PW	0.3	0.24	0.12	0.1	0.06	0.056 (284)	0.056 (284)
D9 - US governor	OR	0.6	0.68	0.46	0.23	0.125	0.113 (220)	0.113 (220)
	PW	0.5	0.48	0.48	0.24	0.13	0.116 (223)	0.116 (223)
D10 - Currency	OR	1.0	1.0	0.66	0.83	0.91	0.875 (274)	0.875 (274)
	PW	1.0	1.0	0.66	0.83	0.91	0.875 (274)	0.875 (274)
D11 - Formula 1	OR	0.9	0.36	0.18	0.19	0.18	0.152 (289)	0.152 (289)
	PW	0.7	0.48	0.24	0.12	0.11	0.073	0.055 (798)

Recall

Data		Top-K						
		10	25	50	100	200	300	400
D1 - Airports	OR	0.022	0.056	0.1133	0.2244	0.4467	0.66	0.984 (441)
	PW	0.0226	0.056	0.1133	0.2244	0.44	0.66	0.984 (441)
D2 - Universities	OR	0.07	0.11	0.15	0.24	0.26	0.3	0.38 (473)
	PW	0.07	0.1	0.13	0.23	0.27	0.3	0.38 (542)
D3 - Car brands	OR	0.086	0.201	0.442	0.653 (87)	0.653 (87)	0.653 (87)	0.653 (87)
	PW	0.086	0.201	0.403	0.73	0.74 (102)	0.74 (102)	0.74 (102)
D4 - US agencies	OR	0.014	0.035	0.067	0.136	0.262	0.397	0.441 (332)
	PW	0.014	0.035	0.068	0.132	0.264	0.4	0.441 (332)
D5 - Rock bands	OR	0.001	0.0036	0.0083	0.0167	0.0199	0.0246	0.0277 (319)
	PW	0.001	0.0036	0.0089	0.015	0.023	0.029	0.1269 (1813)
D6 - MLM	OR	0.0625	0.135	0.343	0.614	0.76	1.0	1.0 (330)
	PW	0.0625	0.1145	0.1458	0.3645	0.75	0.76	1.0 (884)
D7 - Olympic	OR	0.3	0.46	0.73	0.76	0.9	0.9 (200)	0.9 (200)
	PW	0.3	0.53	0.73	0.73	0.73	0.73	0.93 (624)
D8 - FIFA player	OR	0.08	0.24	0.24	0.28	0.6	0.6 (215)	0.6 (215)
	PW	0.12	0.24	0.24	0.4	0.48	0.64 (284)	0.64 (284)
D9 - US governor	OR	0.12	0.34	0.46	0.46	0.5	0.5 (220)	0.5 (220)
	PW	0.1	0.24	0.48	0.48	0.52	0.52 (223)	0.52 (223)
D10 - Currency	OR	0.04	0.102	0.135	0.34	0.74	0.98 (274)	0.98 (274)
	PW	0.04	0.102	0.135	0.34	0.74	0.98 (274)	0.98 (274)
D11 - Formula 1	OR	0.136	0.136	0.136	0.287	0.54	0.66 (289)	0.66 (289)
	PW	0.106	0.181	0.181	0.181	0.33	0.33	0.66 (798)