On Supporting Effective Web Extraction

Jinsoo Lee
Database Lab.
Kyungpook National University

Milestones

- **IEEE ICDE 2010 (데이터베이스 분야 3대 Conf.)**
  - Han, W., Kwak, W., and Yu, H., "On Supporting Effective Web Extraction," In ICDE 2010.
- **VLDB 2010 (데이터베이스 분야 3대 Conf.)**
  - Han, W., Lee, J., Duc, P., and Yu, J., "iGraph: A Framework for Comparisons of Disk-based Graph Indexing Techniques," In VLDB 2010.
- **DTMBIO 2010 (invited to Bioinformatics Journal, IF:3.4)**
  - Lee, J., Pham, M., Lee, J., Han, W., Cho, H., Yu, H., and Lee, J., "Processing SPARQL Queries with Regular Expressions in RDF Databases," In DTMBIO 2010.

Tuple Extraction from Web Pages

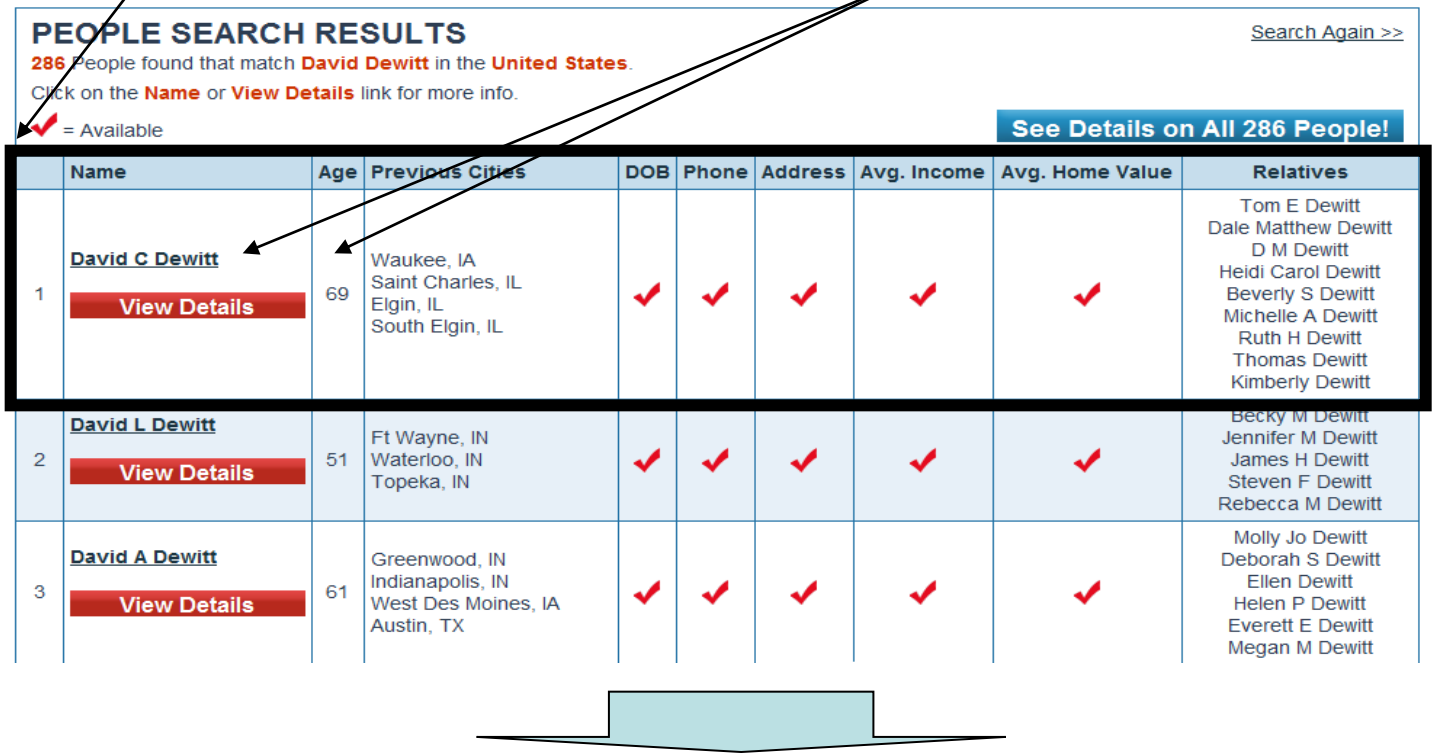
- Various web applications such as web data integration, e-commerce, market monitoring, and mashups

- After tuples are extracted from web pages, they can be easily transformed to different structures

Example

1. select the boundary of the first tuple

2. select elements to extract



PEOPLE SEARCH RESULTS

Search Again >>

286 People found that match **David Dewitt** in the **United States**.

Click on the **Name** or **View Details** link for more info.

✔ = Available

See Details on All 286 People!

	Name	Age	Previous Cities	DOB	Phone	Address	Avg. Income	Avg. Home Value	Relatives
1	David C Dewitt View Details	69	Waukee, IA Saint Charles, IL Elgin, IL South Elgin, IL	✔	✔	✔	✔	✔	Tom E Dewitt Dale Matthew Dewitt D M Dewitt Heidi Carol Dewitt Beverly S Dewitt Michelle A Dewitt Ruth H Dewitt Thomas Dewitt Kimberly Dewitt
2	David L Dewitt View Details	51	Ft Wayne, IN Waterloo, IN Topeka, IN	✔	✔	✔	✔	✔	Becky M Dewitt Jennifer M Dewitt James H Dewitt Steven F Dewitt Rebecca M Dewitt
3	David A Dewitt View Details	61	Greenwood, IN Indianapolis, IN West Des Moines, IA Austin, TX	✔	✔	✔	✔	✔	Molly Jo Dewitt Deborah S Dewitt Ellen Dewitt Helen P Dewitt Everett E Dewitt Megan M Dewitt

Extracted tuples:

T1: (name="David C. Dewitt", age=69)
T2: (name="David L. Dewitt", age=51)
T3: (name="David A. Dewitt", age=61)

Tuple extraction process

- select the boundary element of the first tuple to extract
  - /HTML/BASE/BODY/DIV[2]/TABLE/TBODY/TR[**3**]
- select the corresponding tag for each attribute of the tuple
  - Name: ./TD[2]/DIV[1]/A[1]/B[1]
  - Age: ./TD[3]
- increase the index of the TR tag to extract the second tuple.
  - /HTML/BASE/BODY/DIV[2]/TABLE/TBODY/TR[**4**]

Motivation example

- What if the email column is added just before the age column?
  - Existing systems fail to extract ages
- To correctly extract ages, the XPATH must be changed accordingly
  - *Age*: ./TD[**3**] → ./TD[**4**]
- If we save the HTML file using Microsoft word, the whole structure of the file is changed!

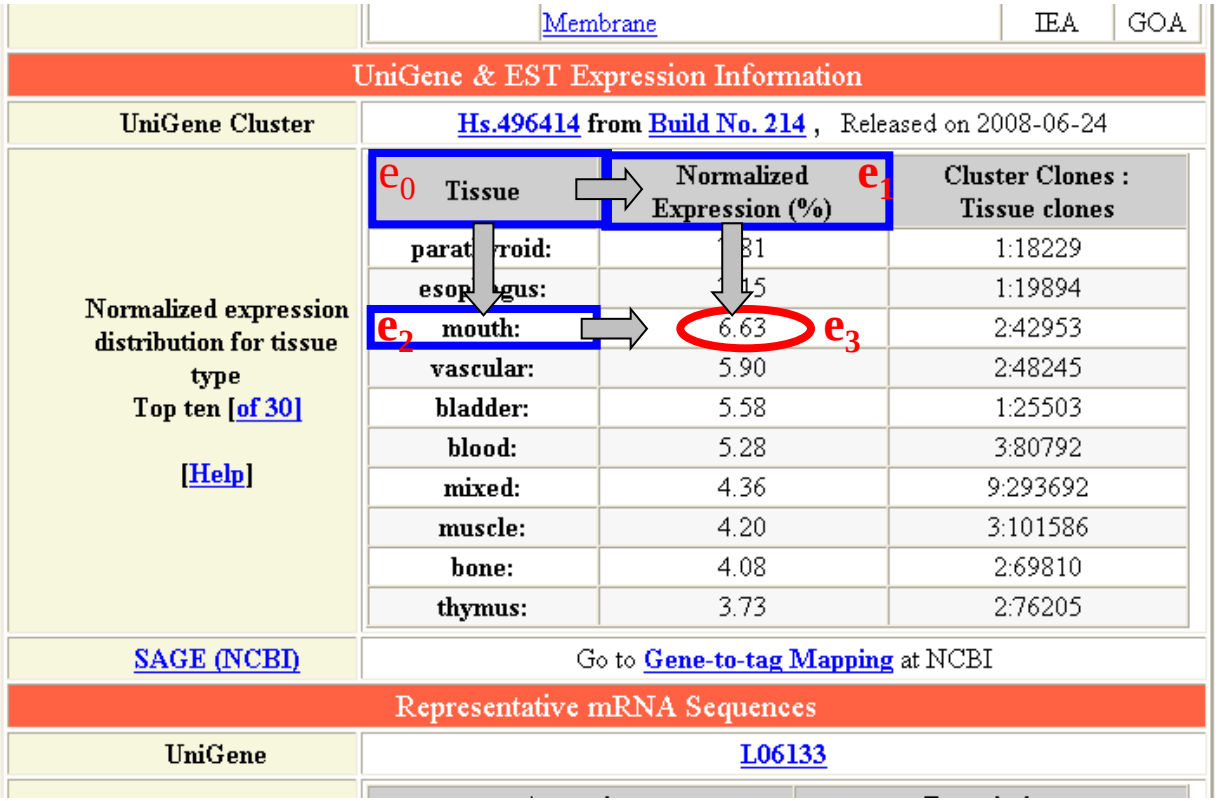
How do humans identify an element in a web page?

- They do not care the underlying HTML tree structures!
- They first seek **reference elements** relevant to the target elements in the web browser
- Then, identify target elements by **relative spatial location** from the reference elements!

Example

Q: Find the normalized expression of the mouth of ATP7A

: reference element ⇨ : spatial relationship ⬭ : extracted element



	Membrane	IEA	GOA

UniGene & EST Expression Information

UniGene Cluster	Hs.496414 from Build No. 214 , Released on 2008-06-24

Normalized expression distribution for tissue type
Top ten [of 30]
[Help]

e_0 Tissue	e_1 Normalized Expression (%)	Cluster Clones : Tissue clones
parat roid:	31	1:18229
esop gus:	15	1:19894
e_2 mouth:	6.63 e_3	2:42953
vascular:	5.90	2:48245
bladder:	5.58	1:25503
blood:	5.28	3:80792
mixed:	4.36	9:293692
muscle:	4.20	3:101586
bone:	4.08	2:69810
thymus:	3.73	2:76205

SAGE (NCBI)	Go to Gene-to-tag Mapping at NCBI

Representative mRNA Sequences

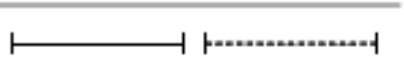
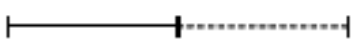
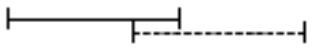
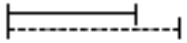
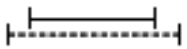
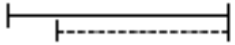
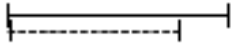
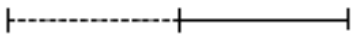
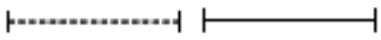
UniGene	L06133

Our key idea

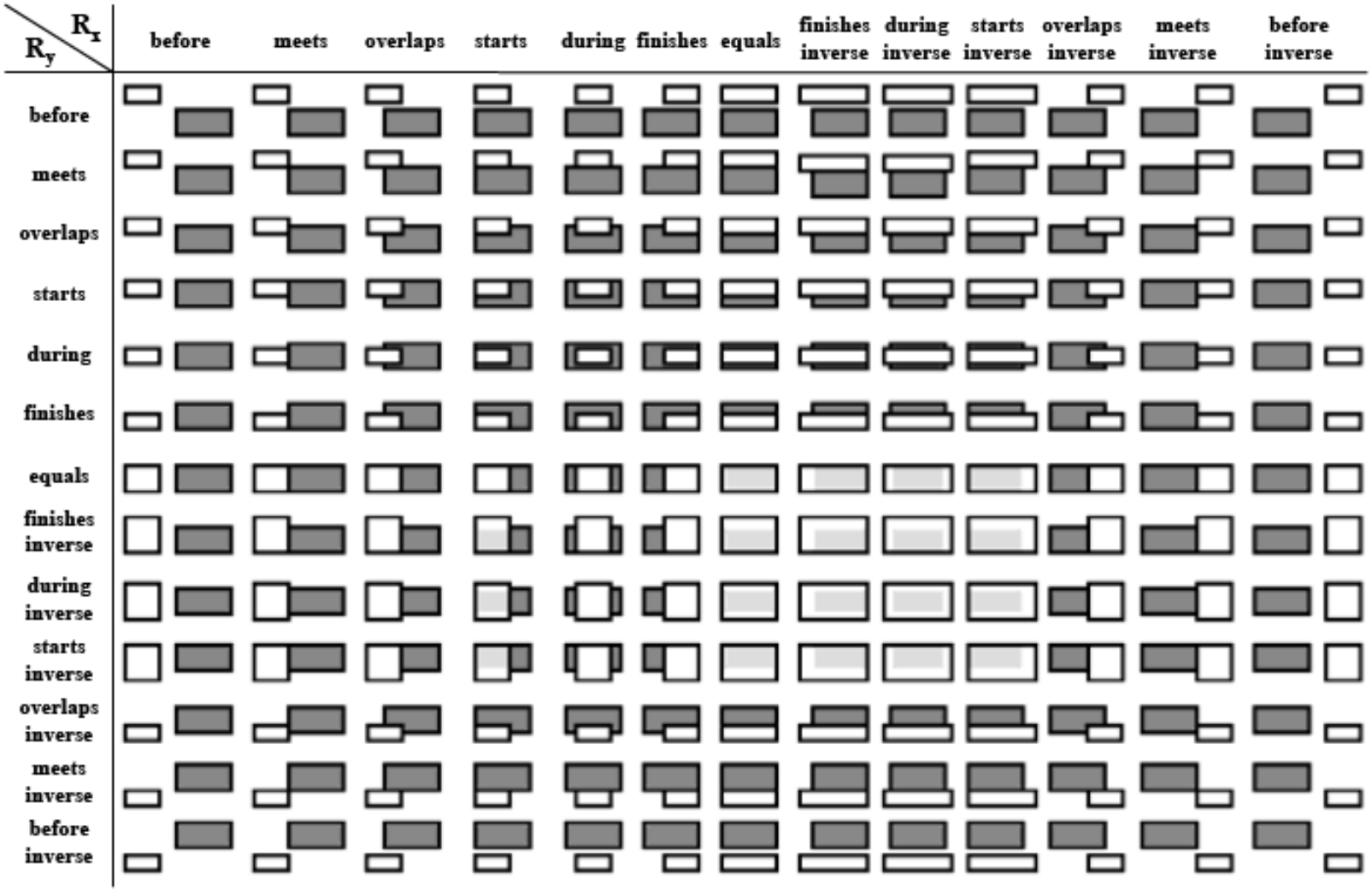
- Regards HTML elements in the rendered page as spatial objects in the 2-D space
- Utilizes spatial relationships among elements rather than the XPath queries
- Executes tailored spatial join to robustly extract tuples from web page

Topological Relationship (1D)

a: b:

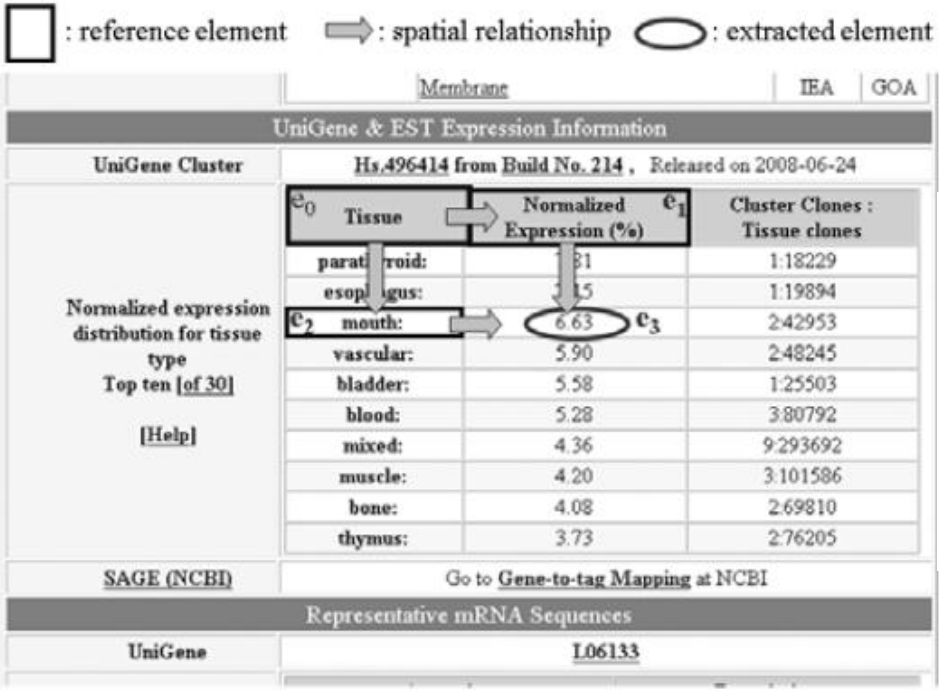
Relation	Semantics
before(a,b)	
meets(a,b)	
overlaps(a,b)	
starts(a,b)	
during(a,b)	
finishes(a,b)	
equals(a,b)	
finishes inverse(a,b)	
during inverse(a,b)	
starts inverse(a,b)	
overlaps inverse(a,b)	
meets inverse(a,b)	
before inverse(a,b)	

Topological Relationship (2D)



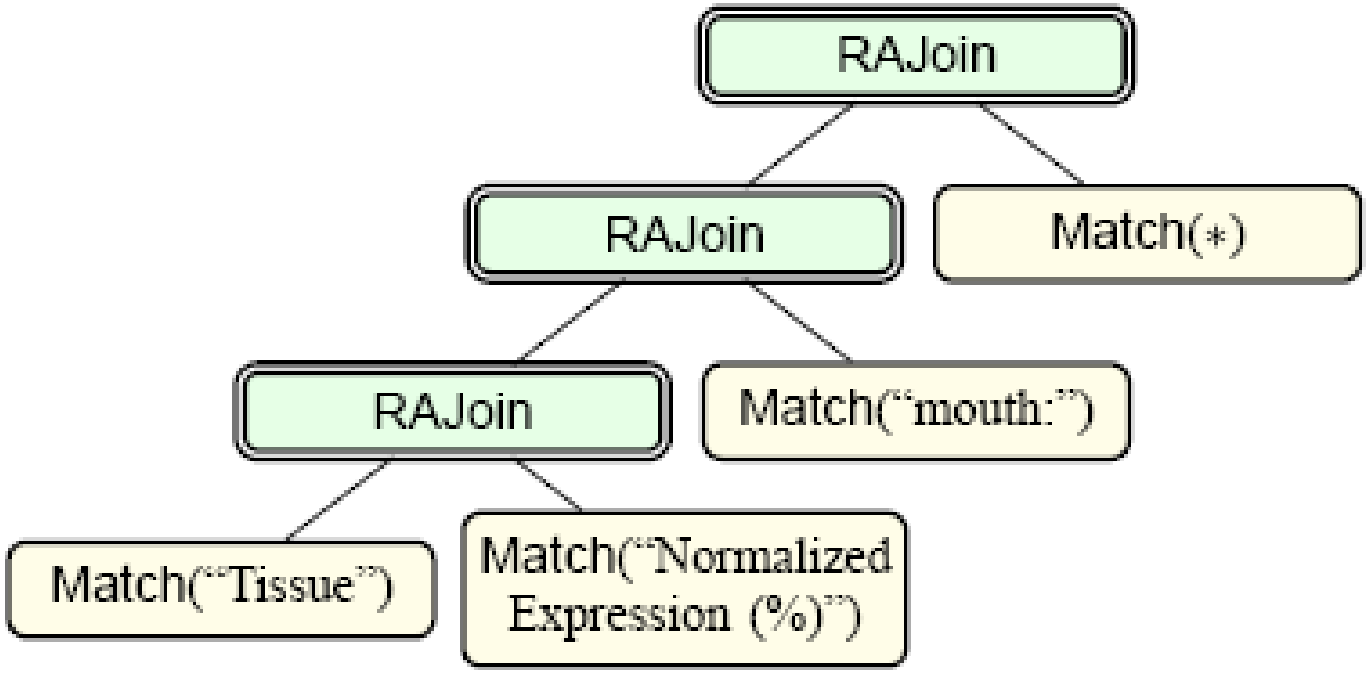
RAQuery example

Query: extract the normalized expression of the mouth of ATP7A




```
for e0 in Match(text="Tissue"),
    e1 in Match(text="Normalized Expression (%)"),
    e2 in Match(text="mouth:"),
    e3 in Match(*)
where  RA(e0, e1, [x:meets_inverse or before_inverse, y:equal]) and
       RA(e0, e2, [x:equal, y:meets_inverse or before_inverse]) and
       RA(e1, e3, [x:equal, y:meets_inverse or before_inverse]) and
       RA(e2, e3, [x:meets_inverse or before_inverse, y:equal])
return e3;
```

Spatial Join for Tuple Extraction



Experiment

- Compared systems
  - Dapper, Robomaker, WebSunDew 2.0, and Web Content Extractor 3.1

- We edit web pages in ten different ways and examine if they still correctly extract target elements

Ten Edit Operations

1. Erase the first column of the table.
2. Erase elements in the table not relevant to the target.
3. Erase elements on the path of target in the HTML source.
4. Split the table into two tables.
5. Insert a similar table around the table of the target.
6. Replace the table tags (e.g., *TABLE, TBODY, TR, TH, TD*) with *DIV* tags.
7. Regroup table elements in the column major order using *DIV*.
8. Insert a new column around the target in the table.
9. Switch the columns of the targets in the table.
10. Save the HTML source using MSWord and Dreamweaver. (The appearance will stay the same but the HTML source can change substantially.)

Table 3: Robustness of the five systems w.r.t the ten edit operations: 'Y' if the system correctly extracts the target elements, and 'N' otherwise. Robo: Robomaker; WebSD: WebSunDew; WebCE: Web Content Extractor.

Edit #	Dapper	Robo	WebSD	WebCE	Ours
	(Yahoo's people search, SOURCE)				
1	(N, N)	(N, Y)	(N, N)	(N, N)	(Y, Y)
2	(Y, Y)	(Y, Y)	(Y, Y)	(Y, Y)	(Y, Y)
3	(Y, N)	(N, N)	(N, N)	(N, N)	(Y, Y)
4	(N, N)	(N, N)	(N, N)	(N, N)	(Y, Y)
5	(N, N)	(N, Y)	(N, N)	(N, N)	(Y, Y)
6	(N, N)	(N, N)	(N, N)	(N, N)	(Y, Y)
7	(N, N)	(N, N)	(N, N)	(N, N)	(Y, Y)
8	(N, N)	(N, N)	(N, N)	(N, N)	(Y, Y)
9	(N, N)	(N, N)	(N, N)	(N, N)	(Y, Y)
10 Word	(N, N)	(N, N)	(N, N)	(N, N)	(Y, Y)
10 Weaver	(Y, Y)	(Y, Y)	(Y, Y)	(Y, Y)	(Y, Y)

Thank You!