Record-Boundary Discovery in Web Documents

D. W. Embley

Y. Jiang

Y. -K. Ng

Brigham Young University

ACM SIGMOD Conference, 1999.

"Ontology-based Extraction and Structuring of Information from Data-rich Unstructured Documents"

ACM CIKM Conference, 1998.

Outline

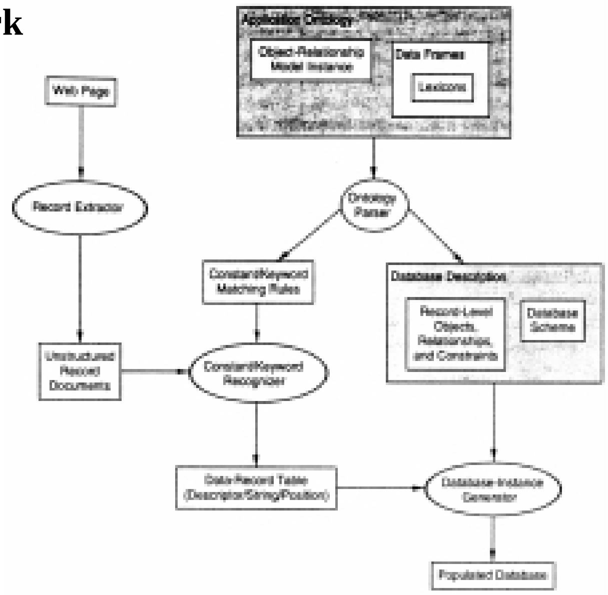
- **Introduction**
- **Record-Boundary Discovery**
- **Individual Heuristics**
- **Combined Heuristic**
- **Experiment**
- **Extraction and Structuring**
- **Conclusion**

Introduction

- **Extract and structure the web data**
  - ◆ develop an ontological model instance for a domain of interest (Application Ontology)
  - ◆ parse this ontology (Ontology Parser)
    - ☞ **generate the rules for matching constants and keywords**
    - ☞ **generate a database scheme**
  - ◆ separate a web page into individual record-size chunks (Record Extractor)
  - ◆ extract objects and relationships (Recognizer)
  - ◆ populate the database instance (Generator)

Introduction

- **Framework**



Introduction

- **Assumptions**
  - a web page has multiple records (data-rich)
    - **car advertisements, job listing, obituaries**
  - a web page contains at least one record-separator tag
    - **discover boundaries of records**

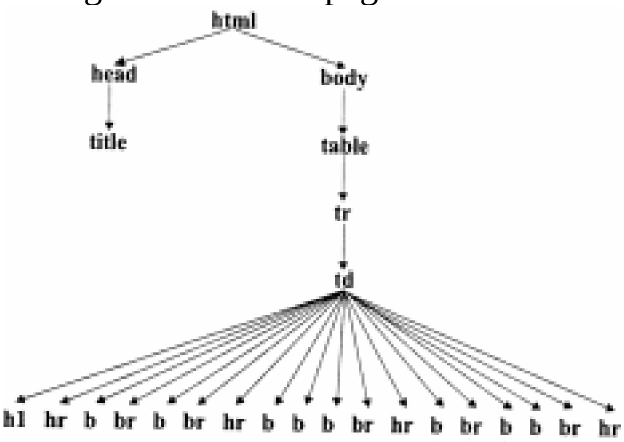
Introduction

- **Sample web page**

```
<html><head><title>Classifieds</title></head>
<body bgcolor="#FFFFFF">
<table><tr><td>
<h1 align="left">Funeral Notices - </h1> October 1, 1998
<hr>
<b>Lemar K. Adamson</b><br> died on September 30, 1998. Lemar was born on September 5, 1913
...
church. ... <b>MEMORIAL CHAPEL</b>, ... <br>
<hr>
Our beloved <b>Brian Fielding Frost</b>, age 41, passed away on September 30, 1998, ...
...
held at ... in the <b>Howard Stake Center</b>,
<b>Carrillo's Tucson Mortuary</b>, ...
Holy Hope Cemetery<br>, ...
<hr>
<b>Leonard Kenneth Gunther</b><br> passed away on September 30, 1998. ...
...
... at <b>HEATHER MORTUARY</b>, ...
... at 11:00 a.m. at <b>HEATHER MORTUARY</b>, on
Tuesday, October 6, 1998. ... <br>
<hr>
</td></tr></table>
All material is copyrighted.
</body>
</html>
```

Record-Boundary Discovery

- **Tag tree**
  - ◆ represent the nested structure of a web page
  - ◆ a node ≡ a region in the web page



Record-Boundary Discovery

- **Locating groups of records**
  - ◆ choose the subtree whose root has the highest fan-out
  - ◆ count the number of appearances for each tag in the immediate child nodes
    - ☞ **irrelevant tag (< 10%): <h1>**
    - ☞ **candidate tag: <hr> <b>
**
- **Ranking the candidate tags**
  - ◆ individual heuristics
  - ◆ combined heuristics

Individual Heuristics

- **Highest-count Tags: HT**
  - sort by the number of appearances in the highest fan-out subtree
    - **b:8 → br:5 → hr:4**
- **Identifiable "separator" Tags: IT**
  - ordered list for common use by observation
    - **hr tr td a table p br h4 h1 strong b i**
    - **75 15 15 14 10 10 10 10 10 10 6 5 (%)**
- **Standard Deviation: SD**
  - compute the standard deviation of the interval between each occurrence of a tag
    - **hr:0.57 → b:0.89 → br:1.25**

Individual Heuristics

- **Repeating-tag Pattern: RP**
  - count the number of occurrences for all pairs of candidate tags without intervening text
  - calculate the differences between counts
    - $x_{xy}=|c_{xy} - c_x|$, $y_{xy}=|c_{xy} - c_y|$
    - $x \rightarrow y$ **if** $\min(x_{xi}) < \min(y_{xj})$ **for all possible i, j**
    - **hr:1 → br:2 → b:5** $[c_{br,hr}=3, c_{hr,b}=3]$
- **Ontology Matching: OM**
  - estimate the number of records by applying the record-identifying fields
    - **Funeral:4, Birth Date: 2, Death Date: 3 ⇒ 3**
    - **hr:1 → br:2 → b:5**

Combined Heuristic

- **Certainty measure**
  - define a confidence measure by using Stanford certainty theory
    - **two evidences E_1 and E_2 come from the same observation B**
    - **$CF(E_1)$ and $CF(E_2)$ are certainty factors (CF)**
      - compound CF = $CF(E_1)+CF(E_2)-CF(E_1)\times CF(E_2)$
  - example
    - **<hr>: HT ⇒ 88%, IT ⇒ 74%, SD ⇒ 66%**
      - 88%+74%+66% - 88%×74% - 74%×66% - 66%×88% + 88%×74%×66% = 98.93%

Combined Heuristic

- **CF of individual heuristics [training]**

Obituaries

Heuristic \ Ranking	*1*	*2*	*3*	*4*
OM	83%	17%	0%	0%
RP	83%	7%	10%	0%
SD	59%	27%	14%	0%
IT	92%	8%	0%	0%
HT	58%	23%	17%	2%

Car advertisements

Heuristic \ Ranking	*1*	*2*	*3*	*4*
OM	86%	8%	4%	2%
RP	72%	18%	8%	2%
SD	72%	18%	10%	0%
IT	100%	0%	0%	0%
HT	40%	42%	16%	2%

selected CF (average) ⇒

Heuristic \ Ranking	*1*	*2*	*3*	*4*
OM	84.5%	12.5%	2.0%	1.0%
RP	77.5%	12.5%	9.0%	1.0%
SD	65.5%	22.5%	12.0%	0.0%
IT	96.0%	4.0%	0.0%	0.0%
HT	49.0%	32.5%	16.5%	2.0%

Combined Heuristic

- **Compound heuristic**
  - apply individual heuristics to ranking
    - **ranks of <hr>: HT ⇒ 3, IT ⇒ 1, SD ⇒ 1**
      - 2%+96%+65.5% - ... + 2%×96%×65.5%
- **Algorithm**
  - construct a tag tree
  - locate the highest fan-out subtree
  - extract the candidate tags
  - apply individual heuristics
  - apply Stanford certainty theory
  - choose the tag with the highest compound CF

Combined Heuristic

- **Combinations of five heuristics [training]**
  - 26 choices
  - success rate

Compound Heuristic	Success Rate	Compound Heuristic	Success Rate
OR	85.83%	OSI	95.00%
OS	88.00%	OSH	87.50%
OI	95.00%	OIH	95.00%
OH	79.00%	RSI	95.00%
RS	79.50%	RSH	85.50%
RI	95.00%	RIH	95.00%
RH	76.33%	SIH	95.00%
SI	95.00%	ORSI	100.00%
SH	69.50%	ORSH	82.50%
IH	95.00%	ORIH	100.00%
ORS	81.50%	OSIH	95.00%
ORI	93.33%	RSIH	100.00%
ORH	84.83%	ORSIH	100.00%

HT	**H**
IT	**I**
SD	**S**
RP	**R**
OM	**O**

Combined Heuristic

- **Example**
  - individual heuristics
    - **HT: [(b,1), (br,2), (hr,3)]**
    - **IT: [(hr,1), (br,2), (b,3)]**
    - **SD: [(hr,1), (b,2), (br,3)]**
    - **RP: [(hr,1), (br,2), (b,3)]**
    - **OM: [(hr,1), (br,2), (b,3)]**
  - compound heuristic
    - **ORSIH: [(hr,99.96%), (br,64.75%), (b,56.34%)]**

Experiment

- **Test results**
  - 4 test sets
  - 20 web pages

On-line Newspaper URL	*OM*	*RP*	*SD*	*IT*	*HT*	*A*
Alameda Newspaper www.adone.com/alameda	1	1	1	1	1	1
Idaho State Journal www.journalnet.com	1	1	2	1	2	1
	1	1	1	1	1	1
	1	1	1	1	1	1
	1	1	1	1	2	1

On-line Newspaper URL	*OM*	*RP*	*SD*	*IT*	*HT*	*A*
Arkansas Democrat - Gazette www.ardemgaz.com	1	1	1	1	2	1
Sioux City Journal www.siouxcityjournal.com	1	2	2	1	4	1
Knoxville News www.knoxnews.com	1	1	1	1	1	1
Lincoln Journal Star www.nebweb.com	1	1	1	1	1	1
Reno Gazette - Journal www.nevadanet.com/renogazette	3	3	1	1	3	1

rank number of correct tag

Experiment

- **Success rates**

Heuristic	*Success Rate*
OM	80%
RP	75%
SD	65%
IT	95%
HT	45%
ORSIH	100%

```
'94 CHEV Monte Carlo Z34, loaded, bright Red
35,000 actual miles!  A great buy at $14,990.
$750 to 1000 down.  MURDOCK CHEVROLET 298-8090

#####

'94 CHEV Corsica, 88,281 miles.  Ask for #16. $4,900.
Government Surplus533-5885

#####

'89 AUDI 80, red, auto., p/w, p/l, sunroof, loaded, 128K,
new trans., new diff.  Runs perfect, must sell. $3300 obo.
gcall Nate, 554-4414
```

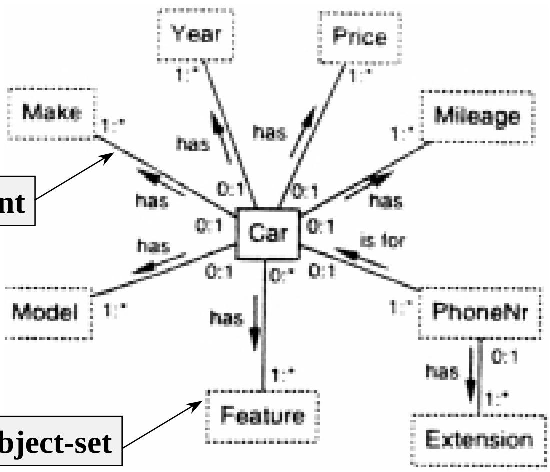

Extraction and Structuring

- **Main processes**
  - ontology parser
    - **constant/keyword matching rules**
      - a list of regular expressions for each object-set
    - **SQL create-table statements**
      - object-set name ⇒ attribute name
  - constant/keyword recognizer
    - **name/string/position table**
  - database-instance generator
    - **tuples with a sequence of (attribute,value) pairs**

Extraction and Structuring

- **Ontology for car advertisements**
  - graphical
  - textual

participation constraint

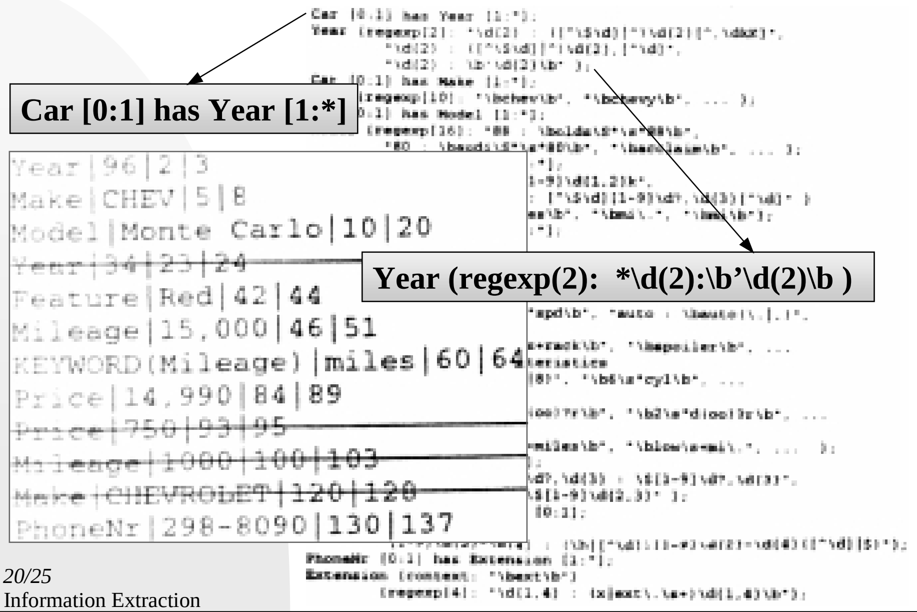


object-set

Extraction and Structuring

Car [0:1] has Year [1:*]

Year (regexp(2): *\d(2):\b'\d(2)\b)



Extraction and Structuring

- **Heuristics for selection of constants**
  - keyword proximity: one-to-many
    - **(Mileage,15000)**
  - functional relationship: one-to-one
    - **(Model,Carlo), (PhoneNr,298-8090)**
  - nonfunctional relationship: many-to-many
    - **(Feature,Red)**
  - first occurrence without constraint violation
    - **(Year,96), (Make,CHEV), (Price,14990)**

Extraction and Structuring

- **Database instances**
  - tuples
    - **Car (1001, “96”, “CHEV”, “Monte Carlo”, “15000”, “14990”, “298-8090”)**
    - **Car-Feature(“1001”, ”Red”)**
  - table

Year	Make	Model	Price
94	DODGE		4,995
94	DODGE	Intrepid	10,000
91	FORD	Taurus	3,500
90	FORD	Probe	
88	FORD	Escort	1000

Extraction and Structuring

- **Experiments**
  - N: number of facts in the source
  - C: number of facts declared correctly
  - I: number of facts declared incorrectly

precision

recall

	N	C	I	$\frac{C}{N}$	$\frac{C}{C+I}$
Car	116	116	0	1.00	1.00
Year	116	116	0	1.00	1.00
Make	116	113	0	0.97	1.00
Model	114	93	0	0.82	1.00
Mileage	31	28	0	0.90	1.00
Price	103	103	0	1.00	1.00
PhoneNr	116	109	0	0.94	1.00
Extension	2	1	0	0.50	1.00
Feature	289	264	1	0.91	0.996
All Attributes	1003	943	1	0.94	0.9989

Conclusion

- **Contributions**
  - an ontology-based framework for extracting and structuring information in web pages
  - a heuristic approach to discovering record boundaries in web pages
  - a heuristic approach to recognizing facts contained in web pages
- **Difficulties**
  - misidentification of attributes: I-15566-2441
  - variations in patterns: Wind95⇒Win95
  - typographical mistakes: Chrystler⇒Chrysler

Conclusion

- **Issues**
  - make the ontological descriptions richer
    - ☞ **generalization/specialization hierarchies, aggregation, n-ary relationship sets, …**
  - improve schema generation and database population
    - ☞ **better data type: 55000 vs. 55k**