

Lecture 14: Some Topics in Web Search Research

The Presidents of the United States of America

President	Party	Term as President	Vice-President
George Washington (1775-1799)	None, Federalist	1789-1797	John Adams
John Adams (1797-1801)	Federalist	1797-1801	Thomas Jefferson
Thomas Jefferson (1801-1809)	Democratic-Republican	1801-1809	Aaron Burr, George Clinton
James Madison (1809-1817)	Democratic-Republican	1809-1817	Congressional Union, Elbridge Gerry
James Monroe (1817-1825)	Democratic-Republican	1817-1825	David Thompson
John Quincy Adams (1825-1849)	Democratic-Republican	1825-1829	John Calhoun, Martin Van Buren
Andrew Jackson (1829-1837)	Democratic	1829-1837	Richard Johnson
W. H. Harrison (1841-1845)	Whig	1841	John Tyler
J. K. Polk (1845-1849)	Democrat	1845-1849	George Dallas
J. A. Tyler (1849-1850)	Whig	1849-1850	Millard Fillmore
Z. Taylor (1850-1852)	Whig	1850-1852	
J. C. Breckinridge (1852-1857)	Whig	1852-1857	John Breckinridge
Franklin Pierce (1857-1861)	Democratic	1857-1861	William King
James Buchanan (1861-1865)	Democratic	1861-1865	John Breckinridge

Administration

- Some changes to the syllabus
 - PA6 goes out today; **due March 15**
 - PA7 goes out March 15; **due March 22**
 - Final Project goes out March 17 (as before)
 - FP Proposals **due March 24** (earlier!)
 - Midterm #2 on March 29 (later!)
- Upcoming lectures:
 - Topics in Web Search Research (today!)
 - Auctions
 - Recommendation systems
 - Maybe some surprises

2

Topics in Web Research

- There's a lot going on
 - Text Mining, Search Log Mining
 - Opinion Extraction
 - Recommender Systems
 - Social Search, Mobile Search
 - Search Engine Architecture
 - Developing World Access to the Web
 - All Things Wikipedia
 - Browser Security
- Even XML !!!

3

Topics in Web Research (2)

- Rather than talk about 10 things for 3 slides each, we'll discuss a few in depth
- My work deals w/*information extraction*
 - Getting structured, queryable information from Web text
 - Perhaps slightly more database-centric than a lot of Web work

4

The Modern Web

- Web pages contain structure that is obvious to humans, though not machines
- Search engines are blind to it
 - Search interfaces sort URLs by relevance
 - That's it

5

The Presidents of the USA - EnchantedLearning.com - Mozilla Firefox

As a thank-you bonus, click [here](#) for a banner-free version of the site, with print-friendly pages.

Already a member? [Click here.](#)

EnchantedLearning.com

U.S. History

President	Party	Term as President	Vice-President
1. George Washington (1775-1799)	None, Federalist	1789-1797	John Adams
2. John Adams (1797-1801)	Federalist	1797-1801	Thomas Jefferson
3. Thomas Jefferson (1801-1809)	Democratic-Republican	1801-1809	Aaron Burr, George Clinton
4. James Madison (1809-1817)	Democratic-Republican	1809-1817	Elbridge Gerry
5. James Monroe (1817-1825)	Democratic-Republican	1817-1825	David Thompson
6. John Quincy Adams (1825-1849)	Democratic-Republican	1825-1829	John Calhoun
7. Andrew Jackson (1829-1845)	Democratic	1829-1837	John Calhoun, Martin Van Buren
8. Martin Van Buren (1837-1841)	Democratic	1837-1841	Walter Johnson
9. William H. Harrison (1837-1841)	Whig	1841	John Tyler
10. John Tyler (1841-1845)	Whig	1841-1845	
11. James K. Polk (1845-1849)	Democrat	1845-1849	George Dallas
12. Zachary Taylor (1849-1850)	Whig	1849-1850	
13. Millard Fillmore (1850-1857)	Whig	1850-1857	
14. Franklin Pierce (1857-1861)	Democratic	1857-1857	William King
15. James Buchanan (1857-1861)	Democratic	1857-1861	John Breckinridge

1

Albert Einstein was born in Germany on March 14 1879.

Albert Einstein had two public personas. One was his work: he was a dedicated and ground-breaking scientist. The other was peace, to which he was committed all his life. Both were based on the same principles: justice, equality, and that all people's health and well-being must, one of the great modern problems, be ensured. How responsible are scientists for the consequences of their discoveries?

A child like Albert could be easily intimidated, threatening and after. He loathed the 'dull, mechanical method of teaching'; he didn't fit in, he didn't work, and was thought 'precious and insolent'; at 15 he took seriously the suggestion of an

What about databases?

- Import interfaces assume data is formally-structured and defined
- Bad assumptions about scale

8

Vision: The Structured Web

- What if we had an automatically-extracted structured version of all information on the Web?
- A "Database of Everything"?

9

Highly Expressive Web Search

- "List some scientists, their inventions, and their years of birth"
 - `q(?scientist, ?invention, ?dob):-
invented(?scientist, ?invention),
born-in(?scientist, <year>?dob)`
[CIDR07, "Structured Querying...", Cafarella et al]

scientist	invention	dob	prob
Kepler	log tables	1571	.8
Heisenberg	matrix mechanics	1901	.8
Galileo	telescope	1564	.7
Newton	calculus	1643	.7

- Vision: Combine breadth of search engine with expressiveness of relational database

10

Rank(1)	City / Urban area(5)	Country	Population(1)
1/1/1	Tokyo-Yokohama	Japan	36,000,000(3.6E7)
2/2/2	New York	USA	17,800,000(1.78E7)
3/3/3	Los Angeles	USA	13,800,000(1.38E7)
4/4/4	Santiago	Chile	13,700,000(1.37E7)
5/5/5	Mexico City	Mexico	18,000,000(1.8E7)
6/6/6	Delhi	India	14,000,000(1.4E7)
7/7/7	Shanghai	China	14,000,000(1.4E7)
8/8/8	Montreal	Canada	3,600,000(3.6E6)
9/9/9	London	UK	8,900,000(8.9E6)
10/10/10	Paris	France	2,200,000(2.2E6)
11/11/11	Istanbul	Turkey	14,000,000(1.4E7)
12/12/12	Lagos	Nigeria	10,000,000(1.0E7)
13/13/13	Madrid	Spain	3,400,000(3.4E6)
14/14/14	Caracas	Venezuela	3,200,000(3.2E6)
15/15/15	Los Angeles	USA	11,700,000(1.17E7)
16/16/16	Beijing	China	21,000,000(2.1E7)
17/17/17	Shenzhen	China	14,000,000(1.4E7)
18/18/18	Taipei	Taiwan	2,800,000(2.8E6)
19/19/19	Yokohama	Japan	33,200,000(3.32E7)
20/20/20	New York Metro	USA	17,800,000(1.78E7)
21/21/21	Sao Paulo	Brazil	17,700,000(1.77E7)
22/22/22	Seoul/Gyeonggi	South Korea	17,500,000(1.75E7)

11

Web Data Integration

- "Compile database of all VLDB PC members"
 - [Under Submission, "Data Integration...", Cafarella et al]
- Vision: Easy reuse of Web's wealth of data (whether or not in XML)

12

Outline

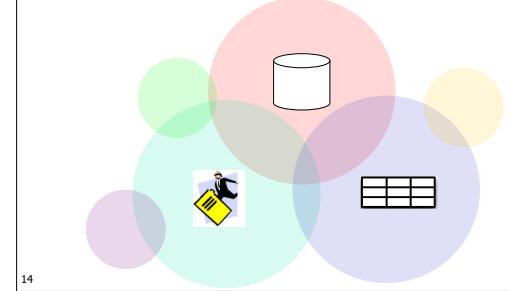
- Introduction
- The Structured Web
 - Tables
 - Multiple Tables
 - Text
- Future Work
- Conclusions



13

A Note

- Structured Web overlaps w/Deep Web...
- But *not* the same thing



14

Research Methodology

- Information Extraction is venerable field, but doesn't scale to the Web. We need:
 - Domain-independence
 - Scalable computer systems
- Draw on databases, AI, systems

15

Outline

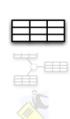
- Introduction
- The Structured Web
 - Tables
 - Multiple Tables
 - Text
- Future Work
- Conclusions



16

Outline

- Introduction
- The Structured Web
 - Tables
 - Multiple Tables
 - Text
- Future Work
- Conclusions



17

President	Party	Term as President	Vice-President
George Washington (1789-1797)	Federalist	1789-1797	John Adams
John Adams (1797-1801)	Federalist	1797-1801	Tamason Jefferson
Thomas Jefferson (1801-1809)	Democratic-Republican	1801-1809	
James Madison (1809-1817)	Democratic-Republican	1809-1817	George Clinton, Elbridge Gerry
James Monroe (1817-1825)	Democratic-Republican	1817-1825	Daniel Tompkins
John Quincy Adams (1825-1849)	Democratic-Republican	1825-1829	John Calhoun
Andrew Jackson (1829-1845)	Democratic	1829-1837	John Calhoun, Martin Van Buren
K. Martin Van Buren (1837-1841)	Democratic	1837-1841	Richard Johnson
W. H. Harrison (1841)	Whig	(1841)	John Tyler
J. Tyler (1790-1862)	Whig	1841-1845	
John Tyler (1790-1862)	Whig	1841-1845	
Jane K. Polk (1795-1849)	Democrat	1845-1849	George Dallas
Zachary Taylor (1794-1850)	Whig	1849-1850	Millard Fillmore
Millard Fillmore (1850-1852)	Whig	1850-1852	
Franklin Pierce (1853-1859)	Democrat	1853-1857	William King
J. F. Buchanan (1791-1865)	Democrat	1857-1861	John Breckinridge

This page contains 16 distinct HTML tables, but only one relational database

Each relational database has its own schema, usually with labeled columns.

WebTables

- WebTables system automatically extracts dbs from web crawl
[WebDB08, "Uncovering... ", Cafarella et al]
[VLDB08, "WebTables: Exploring... ", Cafarella et al]

Raw crawled pages → Raw HTML Tables → Recovered Relations → Schema Statistics → Applications

- An extracted relation is one table plus labeled columns
- Estimate that our crawl of 14.1B raw HTML tables contains ~154M good relational dbs

21

Relation Recovery

Step 1. Relational Filtering
Recall 81%, Precision 41%

Step 2. Metadata Detection
Recall 85%, Precision 89%

- Output**
 - 271M databases, about 125M are good
 - Five orders of magnitude larger than previous largest corpus [WWW02, "A Machine Learning... ", Wang & Hu]
 - 2.6M unique relational schemas
- What can we do with those schemas?**
[VLDB08, "WebTables: Exploring... ", Cafarella et al]

22

Schema Statistics

Recovered Relations

make	model	year
Toyota	Camry	1984

make	model	year
Mazda	Protégé	2003
Chevrolet	Impala	1979

make	model	year	color
Chrysler	Volare	1974	yellow
Nissan	Sentra	1994	red

name	addr	city	state	zip
Dan S	16 Elm St	Bethel	WA	98195
Alon H	129 Elm St	Belmont	CA	94011

Schema Stats useful for computing attribute probabilities

• $p(\text{make})$, $p(\text{model})$, $p(\text{zipcode})$

• $p(\text{name} | \text{make}, \text{model})$, $p(\text{make} | \text{zipcode})$

Readme.txt 182 Apr 26, 2005
cac.xml 813 Jul 23, 2008

23

App #1: Schema Autocomplete

- Input:** topic attribute (e.g., make)
- Output:** relevant schema
 $\{\text{make, model, year, price}\}$
 - "tab-complete" for your database
- For input set I , output S , threshold t**
 - while $p(S - I | I) > t$
 - $\text{newAttr} = \max p(\text{newAttr}, S - I | I)$
 - $S = S \cup \text{newAttr}$
 - emit newAttr

24

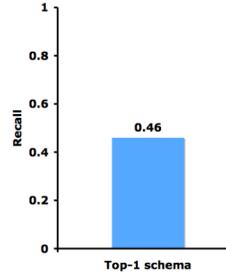
App #1: Schema Autocomplete

name	name, size, last-modified, type
instructor	instructor, time, title, days, room, course
elected	elected, party, district, incumbent, status, ...
ab	ab, h, r, bb, so, rbi, avg, lob, hr, pos, batters
sqft	sqft, price, baths, beds, year, type, lot-sqft, ...

25

App #1: Experiments

- Asked experts for schemas in 10 areas
- What was autocomplete's recall?



26

App #2: Synonym Discovery

- Input: topic attribute (e.g., address)
- Output: relevant synonym pairs (telephone = tel-#)
 - Used for schema matching [VLDB01, "Generic Schema Matching...", Madhavan et al.]
 - Linguistic thesauri are incomplete; hand-made thesauri are burdensome
- For attributes a, b and input domain C , when $p(a,b)=0$

$$syn(a, b) = \frac{p(a)p(b)}{\epsilon + \sum_{z \in A} (p(z|a, C) - p(z|b, C))^2}$$

27

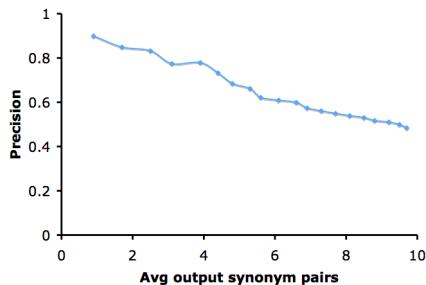
App #2: Synonym Discovery

name	e-mail email, phone telephone, e-mail_address email_address, date last_modified
instructor	course-title title, day days, course course-#, course-name course-title
elected	candidate name, presiding-officer speaker
ab	k so, h hits, avg ba, name player
sqft	bath baths, list list-price, bed beds, price rent

28

App #2: Experiments

- For each input attr, repeatedly emit best synonym pair (until min threshold reached)



29

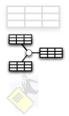
WebTables Contributions

- Largest collection of databases and schemas, by far
- Large-scale extracted schema data for first time; enables novel applications

30

Outline

- Introduction
- The Structured Web
 - Tables
 - Multiple Tables
 - Text
- Future Work
- Conclusions



31

Multiple Tables

- Can we combine tables to create new data sources?
- Data integration for the Structured Web
- Many existing "mashup" tools, which ignore realities of Web data
 - A lot of useful data is not in XML
 - User cannot know all sources in advance
 - Transient integrations



32

Integration Challenge

- Try to create a database of all "VLDB program committee members"

The screenshot shows the VLDB '08 conference website. The main header reads "Very Large Data Bases" and "VLDB '08". Below it, there's a sidebar with links like "Introduction", "Meetings", "PC Chair Message", etc. The main content area is titled "Program Committees" and lists several committees with their chairs:

Committee	Chair
Local Database Program Committee	Beng Chin Ooi (National University of Singapore)
Infrastructure for Information Systems Program Committee	Daniel Abadi (Yale University)
Interactive, Applications, and Experience Program Committee	Shivnath Balu (Duke University)
Education Program Committee	Elisa Bertino (University of North Carolina)
Experiments and Analytics Program Committee	Nico Bruno (Microsoft Research)
Fifth Workshop Program Committee	David J. DeWitt (Cornell University)
VLDB 10-Year Best Paper Award Committee	Elisa Bertino (University of North Carolina)

33

Octopus

- Provides "workbench" of data integration operators to build target database
- Most operators are not correct/incorrect, but high/low quality (like search)
- Also, prosaic traditional operators

34

Walkthrough - Operator #1

- **SEARCH("VLDB program committee members")**

The screenshot shows the VLDB 2005 conference website. The main header is "VLDB 2005". The left sidebar has links for "Core Database Technology Program Committee", "Committee Chair", "Committee Members", "Participants", and "Organization". The main content area is titled "Core Database Technology Program Committee" and lists the "Committee Chair" as Martin Kersten, CWI, Netherlands. Below that is a table of "Committee Members" with names like serge abiteboul, michael adiba, antonio albano, etc. At the bottom, there's a search result table for "SEARCH('VLDB program committee members')".

Member	Organization
serge abiteboul	inria
michael adiba	...grenoble
antonio albano	...pisa
...	...

Member	Organization
serge abiteboul	inria
anastassia ail...	carnegie...
gustavo alonso	etz zurich
...	...

35

The screenshot shows the VLDB 2005 conference website. The main header is "VLDB 2005". The left sidebar has links for "GENERAL", "PROGRAM", "PARTICIPANTS", and "ORGANIZATION". The main content area is titled "Core Database Technology Program Committee" and lists the "Committee Chair" as Martin Kersten, CWI, Netherlands. Below that is a table of "Committee Members" with names like serge abiteboul, anastassia ail..., gustavo alonso, etc. The entire list of committee members is highlighted with a red box.

Member
Serge Abiteboul, INRIA, France
Anastassia Ailamaki, Carnegie Mellon University, USA
Gustavo Alonso, ETH Zurich, Switzerland
Walid Aref, Purdue University, USA
Lars Arge, Aarhus University, Denmark
Brian Babcock, Stanford University, USA
Mikael Berndtsson, University of Skövde, Sweden
Elisa Bertino, Purdue University, USA

36

Walkthrough - Operator #2

- Recover relevant data

CONTEXT()

serge abiteboul	inria	1996
michael adiba	...grenoble	1996
antonio albano	...pisa	1996
...

CONTEXT()

serge abiteboul	inria	2005
anastassia ail...	carnegie...	2005
gustavo alonso	etz zurich	2005
...

37

Walkthrough - Union

- Combine datasets

Union()

serge abiteboul	inria	1996
michael adiba	...grenoble	1996
antonio albano	...pisa	1996
...
serge abiteboul	inria	2005
gustavo alonso	etz zurich	2005
anastassia ail...	carnegie...	2005
...

38

Walkthrough - Operator #3

- Add column to data
- Similar to "join" but join target is a topic

EXTEND("publications")

serge abiteboul	inria	ser1996	abiteboul	ane iB2P Dist_r996
michael adiba	...grenoble	mi1996	"adiba it"	ngrebonle1996
antonio albano	...pisa	an1996	"albano"	Eample of a_1996
serge abiteboul	inria	se1996	abiteboul	ane iB2P Dist_2005
anastassia ail...	carnegie...	an2005	"stassia"	stassia etz zurich 2005
gustavo alonso	etz zurich	gu2005	"alonso"	etz zurich 2005
...

- User has integrated data sources with little effort
- No wrappers; data was never intended for reuse

39

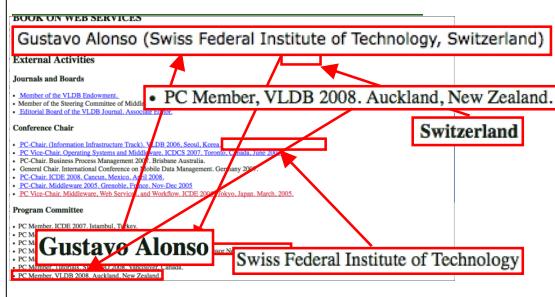
CONTEXT Algorithms

- Input: table and source page
- Output: data values to add to table
- SignificantTerms* sorts terms in source page by "importance" (tf-idf)

40

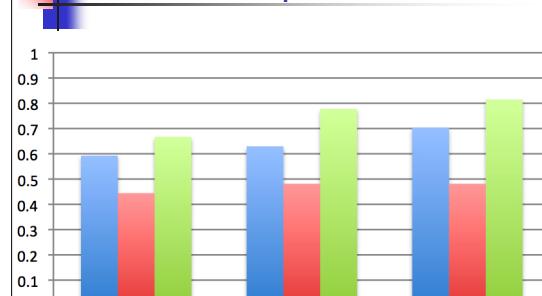
Related View Partners

- Looks for different "views" of same data



41

CONTEXT Experiments



42

Octopus Contributions

- Basic operators that enable Web data integration with very small user burden
- Realistic and useful implementations for all three operators

43

Web Research Topics

- The Structured Web exists in raw form today, but tools largely ignore it
- Traditional databases and Web search not appropriate
- I've described a few useful techniques
- Lots of other interesting Web research problems

44

Extra Material

- For the dedicated reader!

45

Outline

- Introduction
- The Structured Web
 - Tables
 - Multiple Tables
 - Text
- Future Work
- Conclusions



46

Text

- "Einstein was born in Germany" ⇒
`(einstein, was-born-in, germany)`

47

Wrapper Induction

[IJCAI97, "Wrapper Induction for...", Kushmerick, Weld, Doorenbos]
[AutAgent01, "Hierarchical Wrapper Induction...", Muslea, Minton, Knoblock]

Home > Seattle > University District >

Is this your restaurant?

University Zoka

(206) 527-0990

University District

2901 NE Blakley St

Seattle, WA 98105

www.zokacoffee.com

[Send to phone](#)

19 people have voted

94% like it

Add your vote

I like it I don't

Add to your wishlist

<h1 class="page-title"> [RESTAURANT NAME] </h1>
 [Phone-Number]

- Site-or-domain-specific training data
- Only for "semistructured" sources, e.g., Amazon

48

Extraction Rule Learning

[WebDB98, "Extracting Patterns and Relations...", Brin]
 [SIGMOD01, "Snowball: A Prototype System...", Agichtein et al]

"Servers at Microsoft's headquarters in Redmond..."
 "The Armonk-based IBM has introduced..."
 "Intel, Santa Clara, cut prices of its Pentium..."

+

Microsoft	Redmond
IBM	Armonk
Intel	Santa Clara

→ [ORGANIZATION]'s headquarters in [LOCATION]
 → [LOCATION]-based [ORGANIZATION]
 → [ORGANIZATION], [LOCATION]

- Again, requires domain-specific training data

49

System Comparison

	Domain independence	Comp. complexity
Wrapper & Extraction-Rule Learning [various]		
WebKB [AIJ00, "Learning to Construct...", Craven et al]		
KnowItAll [AAAI04, "Methods for Domain...", Etzioni et al]		
Shinyama & Sekine '06 [HLT-NAACL06, "Preemptive...", Shinyama & Sekine]		

50

System Comparison

	Domain independence	Comp. complexity
Wrapper & Extraction-Rule Learning [various]	Domain-specific training data	
WebKB [AIJ00, "Learning to Construct...", Craven et al]	Domain-specific training data	
KnowItAll [AAAI04, "Methods for Domain...", Etzioni et al]	Relations must be given in advance	
Shinyama & Sekine '06 [HLT-NAACL06, "Preemptive...", Shinyama & Sekine]	Relations discovered automatically	

51

System Comparison

	Domain independence	Comp. complexity
Wrapper & Extraction-Rule Learning [various]	Domain-specific training data	various
WebKB [AIJ00, "Learning to Construct...", Craven et al]	Domain-specific training data	$O(D * R)$
KnowItAll [AAAI04, "Methods for Domain...", Etzioni et al]	Relations must be given in advance	$O(D * R)$
Shinyama & Sekine '06 [HLT-NAACL06, "Preemptive...", Shinyama & Sekine]	Relations discovered automatically	$O(D^2)$

52

System Comparison

	Domain independence	Comp. complexity
Wrapper & Extraction-Rule Learning [various]	Domain-specific training data	various
WebKB [AIJ00, "Learning to Construct...", Craven et al]	Domain-specific training data	$O(D * R)$
KnowItAll [AAAI04, "Methods for Domain...", Etzioni et al]	Relations must be given in advance	$O(D * R)$
Shinyama & Sekine '06 [HLT-NAACL06, "Preemptive...", Shinyama & Sekine]	Relations discovered automatically	$O(D^2)$
TextRunner [IJCAI07, "Open Information...", Banko, Cafarella, et al]	Relations discovered automatically	$O(D)$

53

Naïve TextRunner

Einstein was born in Germany.

Step 1. Parse each sentence

- Domain-independent
- No training data or extraction rules needed (the corpus itself is the only input)
- Complexity $O(D)$
- Extremely computationally expensive
- Can be brittle on Web text

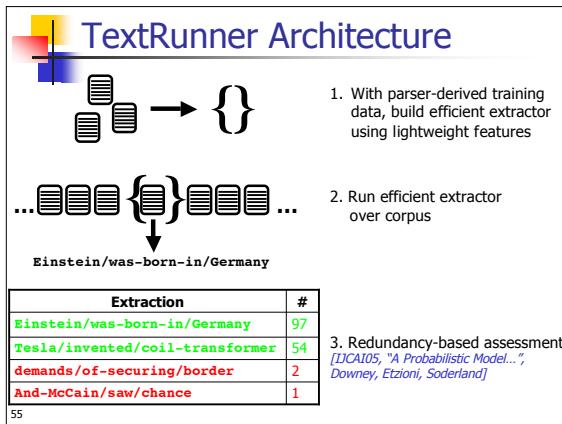
Step 2. Apply extraction patterns

E_1 Verb E_2	X established Y
E_1 NP Prep E_2	X professor of Y
E_1 Verb Prep E_2	X moved to Y

↓

Einstein|was-born-in|Germany

54



Output Quality

- Test corpus of 9M pages
 - 133M sentences
 - 11.3M unique tuples
- On 10 randomly-selected relations...

	Avg error rate	# correct extractions
KnowItAll	18%	11,631
TextRunner		

56

Output Quality

- Test corpus of 9M pages
 - 133M sentences
 - 11.3M unique tuples
- On 10 randomly-selected relations...

	Avg error rate	# correct extractions
KnowItAll	18%	11,631
TextRunner	12%	11,476

57

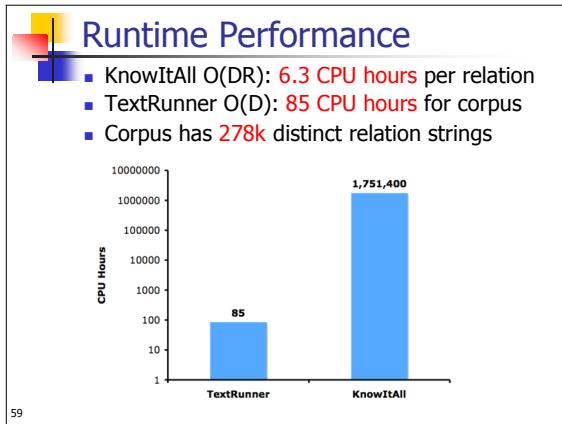
Output Quality

- Test corpus of 9M pages
 - 133M sentences
 - 11.3M unique tuples
- On 10 randomly-selected relations...

	Avg error rate	# correct extractions
KnowItAll	18%	11,631
TextRunner	12%	11,476

- Overall, 79% extractions are “well-formed”, 80% accurate

58



- ## TextRunner Contributions
- TextRunner scales textual information extraction to the entire Web
 - Domain independent (in particular, no relations in advance)
 - Reduces runtime to O(D)
- 60

Future Work

■ WebTables

- Schema autocomplete & synonyms just few of many possible *semantic services*
 - Input: schema; Output: tuples
database autopopulate
 - Input: tuples; Output: schema
schema autogenerate

■ Octopus

- Index support for interactive speeds

61

Future Work (2)

■ "The Database of Everything"

[CIDR09, "Extracting and Querying...", Cafarella]

- Is domain-independence enough?

Text-embedded

be/is
ask/call
arrive/come/go
join/lead
born-in

Table-embedded

<web access log>
<file listing>
<forum posts>
<album listing>
<phone numbers>

- Multi-model, multi-extractor approach probable
- Vast deduplication challenge
- New tools needed for user/extractor interaction

62