Information Extraction, Content Curation, and Machine Learning

Raghu Ramakrishnan

Yahoo! Fellow

Chief Scientist, Audience and Cloud Computing

(Many slides courtesy of others at Yahoo!)

berts - Yahoo! Search Results - Mozilla Firefox

View History Bookmarks Yahoo! Tools Help

http://search.yahoo.com/search?p=julia+roberts&ei=UTF-8&fr=moz35

ar sunnyvale - Yahoo! Search Re... | julia roberts - Yahoo! Search Res...

YAHOO!

julia roberts

Search

julia roberts twins
julia roberts movies
lyle julia roberts
julia roberts babies
julia roberts henry daniel moder

Explore related concepts:
actor
Search for ju obrts actor
Pretty
Be ctress

julia roberts photos
Julia Roberts News
julia roberts biography
Julia Fiona Roberts

Search Pad

SearchScan - On

40,500,000 results for
julia roberts

Related People

Scarlett Johansson

Emma Robe

Hila Swank

Lindsay Lohan

Tom Hanks

Halle Berry

Julia Roberts - Image Resul

News & Photos | Videos | Twitter

more Julia Roberts photos...

ates News:

Hindus concerned about Julia Roberts' "Eat, Pray, Love" - New Kerala - 6 hours ago

Trailer For 'Eat Pray Love' Starring Julia Roberts - KPBS San Diego - Mar 19 03:18pm

Link Party: Julia Roberts' New Movie Will Teach You How to Live - E! Online - Mar 18 05:48pm

more Julia Roberts news...

Julia Roberts - Wikipedia

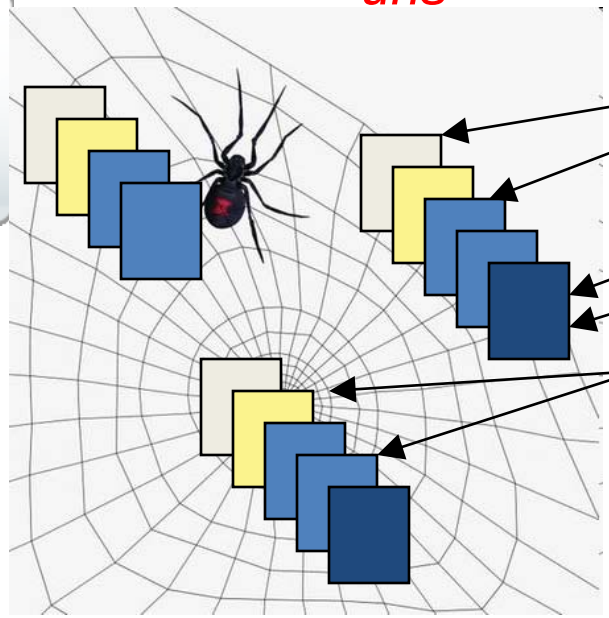
Early life | Career | Influence | Personal life

Julia Fiona **Roberts** is an American actress. She is known for starring in the romantic comedy Pretty Woman opposite Richard Gere, which grossed $463 million worldwide. After receiving...

en.wikipedia.org/wiki/Julia_Roberts - 122k - Cached

Concept-centric information aggregation

Web of Pages



urls

token

julia roberts

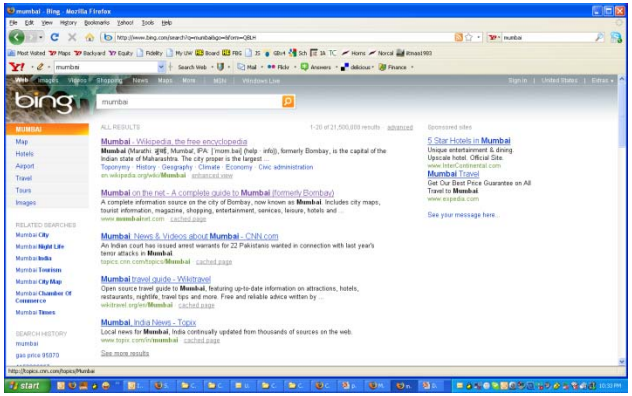
mumbai

restaurant

chinese

san jose

• • •



WEB

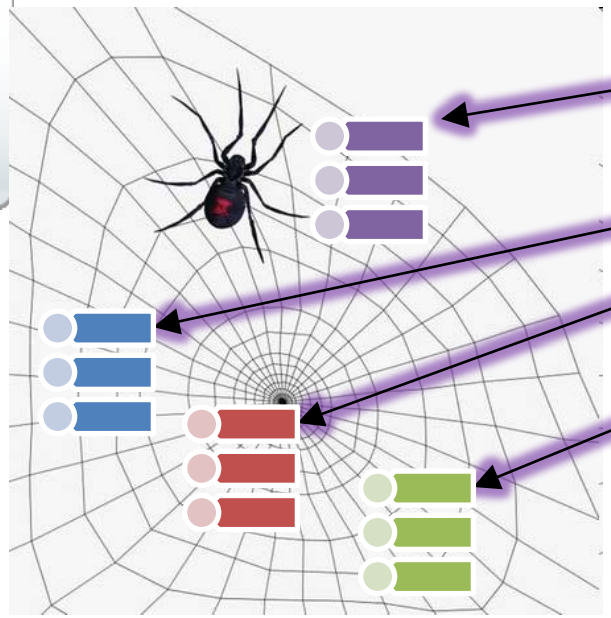
INDEX

SERP

Web of Concepts

structured data

concept



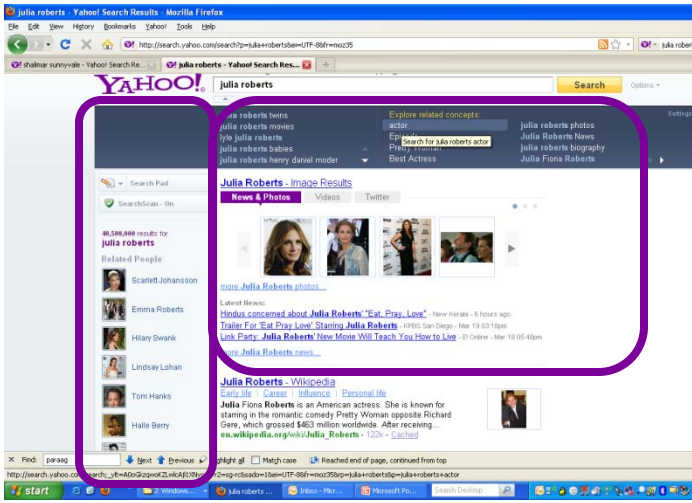
julia roberts

mumbai

restaurant

san jose

• • •



Aggregated KB **INDEX** **SERP**

The “index” is keyed by concept instance, and organizes all relevant information, wherever it is drawn from, in semantically meaningful ways

DBLife: Superhomepage of Gerhard Weikum - Mozilla Firefox

File Edit View Go Bookmarks Yahoo! Tools Help

DBLife weikum Search

Login | Help | The Cimple Project | Wiki

Gerhard Weikum

Check all *non-related* images and click the submit button. Show All Images






0 votes | 0 votes | 0 votes | 0 votes | 0 votes | 0 votes | 0 votes

Submit Check All Clear All

Mentions **1** - **10** out of **137**

Tuesday Oct 31, 2006

PC member of SIGMOD 2007

...Jeff UllmanStanford University Patrick ValduriezINRIA Wei WangUniversity of North Carolina Gerhard WeikumMax-Planck Institute of Computer Sc. Till WestmannBEA Kyu Young WhangKAIST Jun YangDuke University...

http://sigmod07.riit.tsinghua.edu.cn/program.shtml Cached Annotated

Monday Oct 30, 2006

"PUBLICATIONS LIST:" page, Patrick O'Neil's homepage

...Invited paper to the IEEE Bulletin on Data Engineering, March 1994, 45-50, Vol. 17, No. 1. With Gerhard Weikum, "A Log-Structured History Data Access Method (LHAM)".Presented at the Fifth International...

http://www.cs.umb.edu/~poneil/publist.html Cached Annotated

Thursday Oct 26, 2006

Posts message "CIDR 2007 Accepted Papers and Demos" on Dbworld

...25-Oct-2006 conf. ann. Gerhard Weikum CIDR 2007 Accepted Papers and Demos 7-Jan-2007 web page...

http://www.cs.wisc.edu/dbworld/browse.html Cached Annotated

Saturday Oct 21, 2006

PC member of SMDB 2007

Page monitored for first time

...Canada) Ken Salem (University of Waterloo, Canada) Dennis Shasha (New York University, USA) Gerhard Weikum (Max-Planck-Institut fur

137 total mentions occuring in 90 pages.
0 new mentions found in the last 24 hours.

Related People

- Vagelis Hristidis
- Surajit Chaudhuri
- Gustavo Alonso
- Nick Koudas

more

Related Topics

- xml
- optimization
- performance
- algorithms

more

Services

- ICDE 2007 (PC member)
- SIGMOD 2007 (PC member)
- SMDB 2007 (PC member)
- WWW 2007 (PC member)

more

Publications

- Efficient and Decentralized PageRank Approximation in a Peer-to-Peer Web Search Network
- IO-Top-k: Index-access Optimized Top-k Query Processing
- Foundations of Automated Database Tuning
- Counting at Large: Efficient Cardinality Estimation in Internet-Scale Data Networks

more (filtered)

Co-Authors

santa clara - Yahoo! Search Results - Mozilla Firefox

History Bookmarks Yahoo! Tools Help

http://search.yahoo.com/search?p=shower+head+santa+clara&ei=UTF-8&fr=moz35

shower head santa clara

vale - Yahoo! Search Re... | shower head santa clara - Yahoo!...

YAHOO!

shower head santa clara

Search

Options

Start typing to see suggestions.

Explore related concepts:

pool
Detached Single Family
Tub
faucets
tile
MLSListings
toilets
Real Estate MLS

Settings

Search Pad

SearchScan - On

577,000 results for
shower head santa cl...:

Show All

Google Sites

Yahoo! Local

Metacafe

Video Sites

Sponsored Results

Shower Head
High Performance Showerheads for Full Body Coverage. Learn More.
Moen.com/Showerheads

Shower Head stores near Santa Clara, CA

Nearby City

All (26) | San Jose (8) | Mountain View (3) | Los Gatos (2)
Santa Clara (2) | Fremont (3) | Campbell (3) | Sunnyvale (1)

1. Conleff Plumbing Supply ★★★★★ (5)
conleff.com
(408) 988-8005 - 2301 Lafayette St, Santa Clara, CA
5 Reviews | Overview | 1 Photo | Directions

2. Home Depot ★★★☆☆ (9)
homedepot.com
(408) 492-9600 - 2435 Lafayette St, Santa Clara, CA
5 Reviews | Overview | Directions

3. Kitchen & Bath Showplace ★★★★★ (1)
kbshowplace.com
(408) 249-9880 - 1200 Campbell Ave, San Jose, CA
Overview | 1 Photo | Directions

23 More Local Results...

Sponsored Results

Shower Heads
Find **Shower Heads**. Find Out More at Guide2Faucets.
Guide2Faucetse.com/Faucets

Shower Heads for Sale
Your Personal Guide to **shower heads** for sale
www.aashowerheads.info

See your message here...

Next | Previous | Highlight all | Match case | Reached end of page, continued from top

2 Windows... | shower hea... | Inbox - Micr... | Microsoft Po... | Search Desktop

Bookmarks Yahoo! Tools Help

http://search.yahoo.com/search;_ylt=A0oGklHEA_RLAyYBgxJXNyoA?p=eggplant+parmigiana+baltimore&fr2=sb-top&fr=moz35&sao=1

hummus

na baltimore - Yah...

YAHOO!

Web Images Video Local Shopping News More

eggplant parmigiana baltimore

Search

Options

Start typing to see suggestions.

Explore related concepts:

Pizza
veal
Chicken
Italian Restaurant
mushrooms
tomato sauce
Little Italy
Crab Cake

Settin

Search Pad

SearchScan - On

30,300 results for
eggplant parmigiana ...:

Show All

Los Angeles Times

Local Business Sites

Eggplant Parmigiana Restaurants near Baltimore, MD

Neighborhood

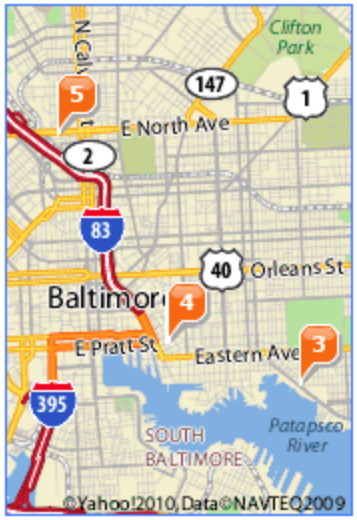
All (36) Canton (2) Charles North (1) Downtown (1)
Abell (1) Central Bal... (4) Chinquapin ... (1) Federal Hill (1)

1 Ciao Bella - Baltimore ★★★★★ (11)
local.yahoo.com
(410) 685-7733 - 236 S High St, Baltimore, MD
Menu: eggplant parmigiana
4 Reviews | Overview | 2 Photos | Directions

2 Amicci's ★★★★★ (20)
amiccis.com
(410) 528-1096 - 231 S High St, Baltimore, MD
Menu: eggplant parmigiana
14 Reviews | Overview | 23 Photos | Directions

3 Pasticcio ★★★★☆ (8)
local.yahoo.com
(410) 522-7700 - 2400 Boston St, Baltimore, MD
Menu: eggplant parmigiana
5 Reviews | Overview | 3 Photos | Directions

4 Caesar's Den ★★★★☆ (7)
caesarsden.com
(410) 547-0820 - 223 S High St, Baltimore, MD
Menu: eggplant parmigiana
4 Reviews | Overview | 11 Photos | Directions



File Edit View History Bookmarks Yahoo! Tools Help

http://search.yahoo.com/search;_ylt=A0oGkIHEA_RLAyYBgxJXNyoA?p=eggplant+parmigiana+baltimore&fr2=

hummus baltimore

eggplant parmigiana baltimore - Yah...

YAHOO!

Web Images Video Local Shopping News More

eggplant parmigiana baltimore

Search

Options

Start typing to see suggestions.

Explore related concepts:

Pizza
veal
Chicken
Italian Restaurant
mushrooms
tomato sauce
Little Italy
Crab Cake

Settin

Search Pad

SearchScan - On

30,300 results for
eggplant parmigiana ...:

Show All

Los Angeles Times

Local Business Sites

Eggplant Parmigiana Restaurants near Baltimore, MD

Neighborhood

All (36) | Canton (2) | Charles North (1) | Downtown (1)
Abell (1) | Central Bal... (4) | Chinquapin ... (1) | Federal Hill (1)

1 Pasticcio ★★★★☆ (8)
local.yahoo.com
(410) 522-7700 - 2400 Boston St, Baltimore, MD
Menu: eggplant parmigiana
5 Reviews | Overview | 3 Photos | Directions

2 Rosina Gourmet
rosinagourmet.com
(410) 675-9300 - 2819 Odonnell St, Baltimore, MD
Menu: chicken parmigiana, grilled eggplant
Overview | Directions

Velleggia's Italian Seafood Restaurant Menu - BALTIMORE, MD ...
Velleggia's Italian Seafood Restaurant: Order Food online from Velleggia's Italian Seafood Restaurant at Baltimore, MD ... Eggplant Parmigiana "Worlds Best" ...
www.campusfood.com/restaurant.asp?campusid=50&mlid=244858 - Cached

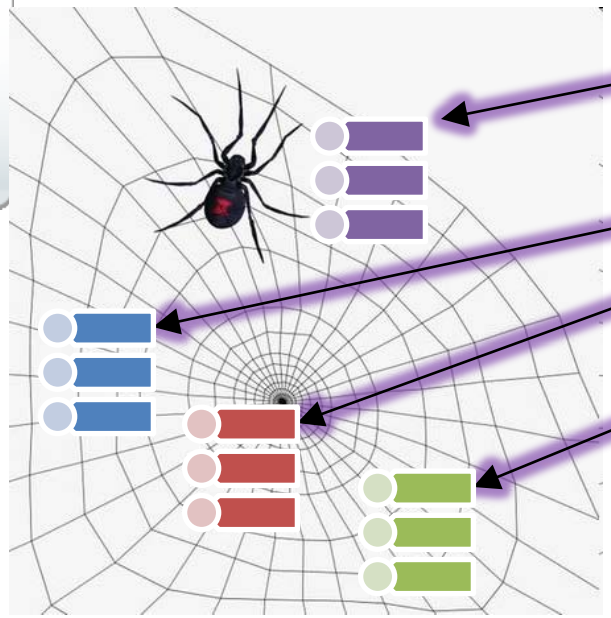
Baltimore Restaurant Week
Eggplant Parmigiana. Fresh sliced eggplant fried to a crisp tenderness, layered with ...
Fresh sliced eggplant fried to a crisp tenderness, layered with mozzarella ...

Done

WoC and User Intent

rich, aggregated data

concept



madonna

mumbai

restaurant

san jose

• • •

Aggregated KB

INDEX

SERP

The "index" is keyed by concept instance, and organizes all relevant information (data describing the concept instance and its relationship to other instances), wherever it is drawn from, in semantically meaningful ways

Understanding Web Content

Concepts

Things of interest to web users:

- Entities
  - E.g., Restaurants
  - Instances with “object” ids, attributes
- Events
  - E.g., NFL games, Concerts, Michael Jackson passes away
  - Instances, some with unique ids
- Topics
  - E.g., High-blood pressure
  - No instances, but many aspects of interest, e.g., drugs, each of which has related concepts (e.g., makers, side-effects, cost) with instances

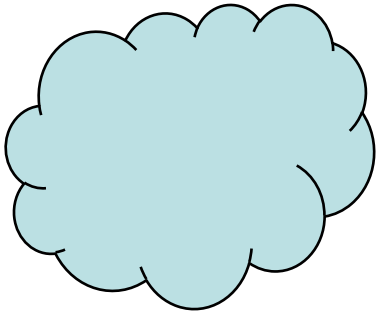
Really, just ER modelling revisited

End-to-End Extraction

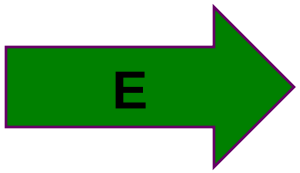
- Let's summarize our extraction goals as a query over a *mythical* true-schools-in-world table:

Q_E=

```
select name, phone, principal-name, api-score, start-
        date,.. (many more)
from true-schools-in-world
where type = "high school" and state = "CA"
```



The Web



Extraction System

V_E=

Name	Phone	…
Cupertino High School	(408) 366-7300	
Fremont High School	(408) 522-2400	
…	…	

An estimate of Q_E(true-schools)

We call this problem End-to-End Extraction

Generative Model of the Web

The true world database

Site queries

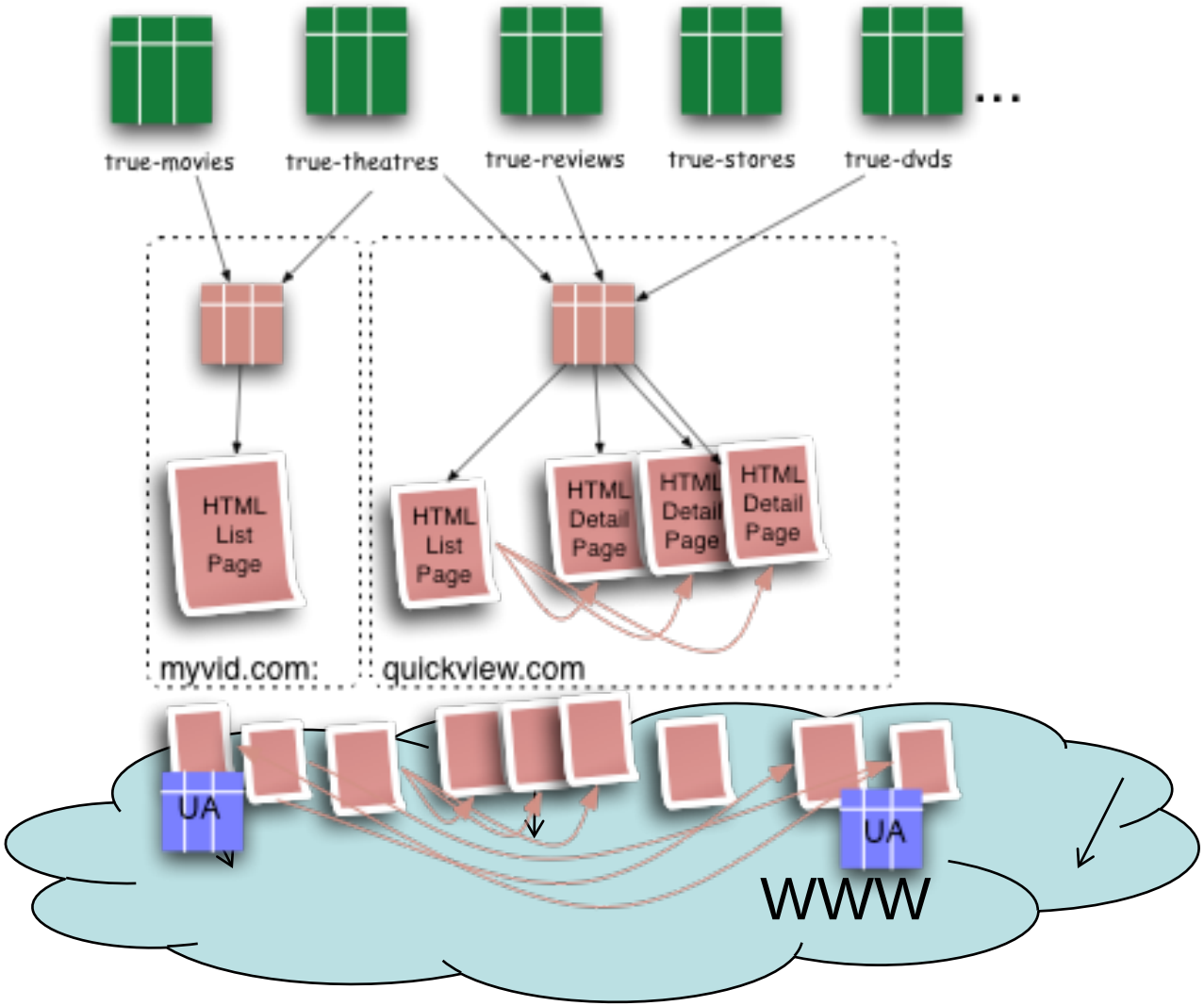
Information loss

Noise addition

Site layout

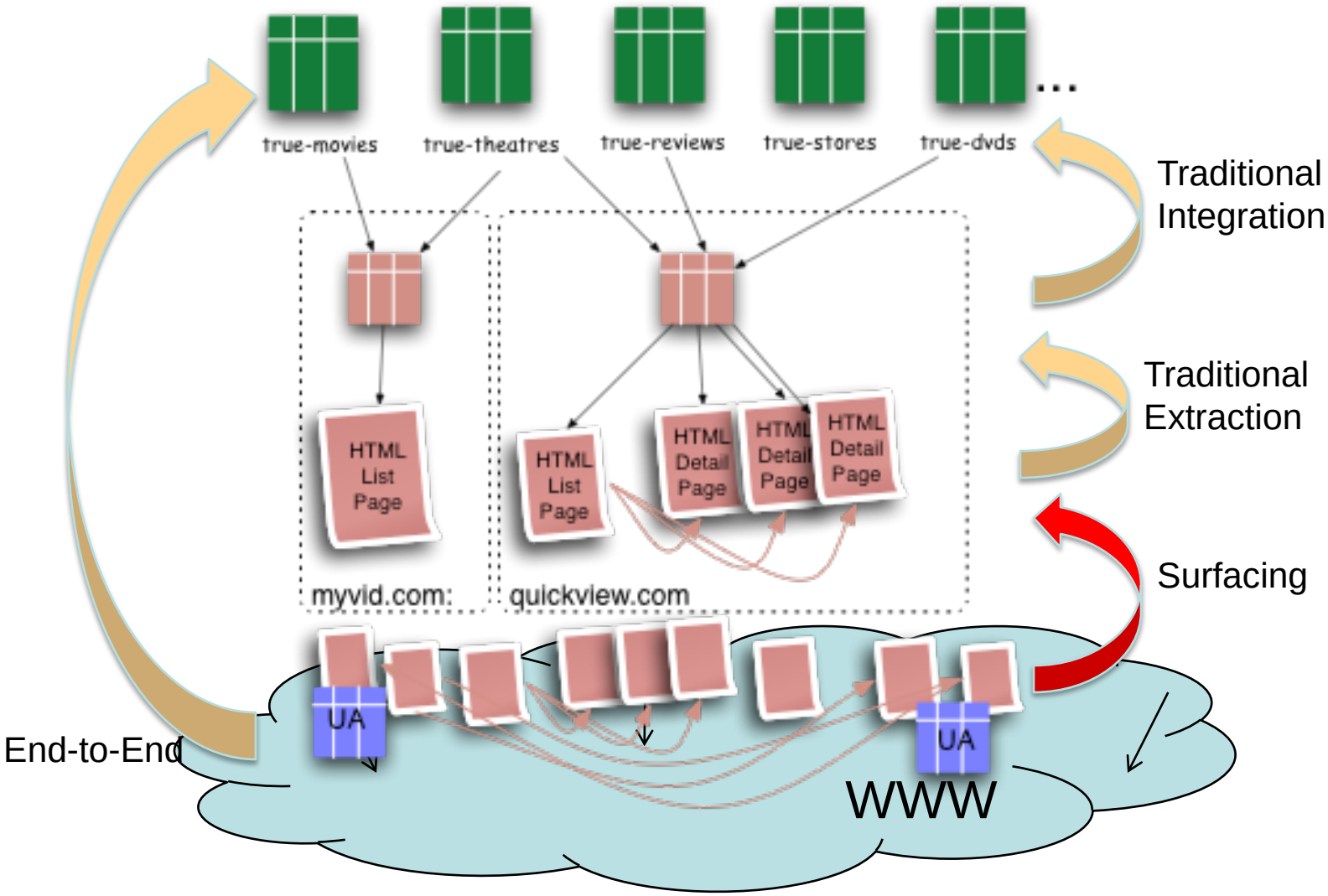
Surround generation

Document Corpus



13

Surfacing, Extraction, Integration



14

Domain-Centric IE

- Our approach is to develop a toolkit for information extraction parameterized by a *domain.* In other words, toolkit input consists of:
  - Q_E, a sample of V_E, and, say, one or two wrappable sites
- How can we hope to fill gaps with low supervision cost?

1. Cheap Supervision
2. Content Redundancy
3. Ability to Rely on Priors
4. Use of surrounding context on web pages/sites

15

Example: Accepted Papers

- Every conference comes with a slightly different format for accepted papers
  - We want to extract accepted papers directly (before they make their way into DBLP etc.)
- Assume
  - Lots of background knowledge (e.g., DBLP from last year)
  - **No** supervision on the target page
- What can you do?



List of Accepted Papers

The list of accepted papers is already available.

- Research Full Presentations (75)
- Research Short Presentations (44)
- Research Poster Presentations (72)
- Industrial Full Presentations (10)
- Industrial Short Presentations (3)
- Industrial Poster Presentations (4)
- Demos (23)

Research Full Presentations (75)

- **Muse: Mapping Understanding and deSign by Example**
  Bogdan Alexe (UC Santa Cruz), Laura Chiticariu (UC Santa Cruz), Renee Miller (University of Toronto), Wang-Chiew Tan (UC Santa Cruz)
- **Transformation-based Framework for Record Matching**
  Arvind Arasu, Surajit Chaudhuri (Microsoft Research, USA), Raghav Kaushik (Microsoft Research)
- **Usage-based schema matching**
  Hazem Elmeleegy (Purdue University), Mourad Ouzzani (Purdue University),

Down the Page a Bit

-:: ICDE 2008 - Cancun, Mexico ::- - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address C:\IEProject\Presentations\RaghuKDD\list\www.icde2008.org.html

-:: ICDE 2008 - Cancun, Mexico ::-

- **NAGA: Searching and Ranking Knowledge**
  Gjergji Kasneci (Max-Planck Institute), Fabian Suchanek (Max-Planck Institute), Georgiana Ifrim (Max-Planck Institute), Maya Ramanath (Max-Planck Institute), Gerhard Weikum (Max-Planck Institute for Informatics)
- **Approximate Joins for Data Centric XML**
  Nikolaus Augsten (Free University of Bozen), Michael Boehlen (Free University of Bolzano-Bozen), Curtis Dyreson (Washington State University), Johann Gamper (Free University of Bozen-Bolzano)
- **The space complexity of processing XML twig queries**
  Ziv Bar-Yossef (Technion, Israel), Mirit Shalem (Technion, Israel)
- **XML Index Recommendation with Tight Optimizer Coupling**
  Iman Elghandour (University of Waterloo), Ashraf Aboulnaga (University of Waterloo), Daniel Zilio (IBM Toronto Lab), Fei Chiang (University of Toronto), Andrey Balmin (IBM Almaden Research Center), Kevin Beyer (IBM Almaden Research Center), Calisto Zuzarte (IBM Toronto Lab, Canada)
- **XML Prefiltering as a String Matching Problem**
  Christoph Koch (Saarland University), Stefanie Scherzinger (Saarland University), Michael Schmidt (Freiburg University)
- **Efficient Processing of XML Update Streams**
  Leonidas Fegaras (University of Texas at Arlington)

Research Short Presentations (44)

- **Validating Multi-column Schema Matchings by Type**
  Bing Tian Dai (National University of Singapore), Nick Koudas (University of Toronto), Divesh Srivastava (AT&T Labs Inc.), Anthony Tung (NUS), Suresh Venkatasubramanian (University of Utah)
- **LOCUST: An Online Analytical Processing Framework for High Dimensional Classification of Data Streams**
  Charu Aggarwal (IBM T. J. Watson Research Center), Philip Yu (IBM T. J. Watson Research Center)
- **k-Anonymization Revisited**
  Tamir Tassa (The Open University of Israel), Aristides Gionis, Arnon Mazza
- **RiTE: Providing On-Demand Data for Right-Time Data Warehousing**
  Christian Thomsen (Aalborg University), Torben Bach Pedersen (Aalborg University), Wolfgang Lehner (Dresden University of Technology)
- **Summarizing Graph Patterns**
  Liu Yong (HIT)
- **Compact Similarity Joins**
  Brent Bryan (Carnegie Mellon University), Frederick Eberhardt (Carnegie Mellon University), Christos Faloutsos (Carnegie Mellon University)
- **Standing Out in a Crowd: Selecting Attributes for Maximum Visibility**
  Muhammed Miah (University of Texas Arlington), Gautam Das, Vagelis Hristidis (Florida International University), Heikki Mannila (University of Helsinki)
- **Injector: Mining Background Knowledge for Data Anonymization**
  Tiancheng Li (Purdue University), Ninghui Li (Purdue University)
- **Efficient Data Authentication in an Environment of Untrusted Third-Party Distributors**

Extracted Records

C:\IEProject\Presentations\RaghuKDD\bibRecords.out.html - Microsoft Internet Explorer

Record 100765

Improving Information Access for a Community of Practice Using Business Process as Context Yu Deng (IBM Research) , Murthy Devarakonda (IBM Research) , Nithya Rajamani (IBM Research) , Wlodek Zadrozny (IBM Research)

Record 100766

NAGA : Searching and Ranking Knowledge Gjergji Kasneci (Max - Planck Institute) , Fabian Suchanek (Max - Planck Institute) , Georgiana Ifrim (Max - Planck Institute) , Maya Ramanath (Max - Planck Institute) , Gerhard Weikum (Max - Planck Institute for Informatics)

Record 100767

XQSE : XQuery Scripting Extension in the AquaLogic Data Services Platform Daniel Engovatov (BEA Systems) , Vinayak Borkar (BEA Systems) , Michael Carey (BEA Systems) , Dmitry Lychagin (BEA Systems) , Till Westmann (BEA Systems) , Warren Wong (BEA Systems)

Record 100768

Efficient Data Interpretation and Compression over RFID Streams Richard Cocci (University of Massachusetts) , Thanh Tran , Yanlei Diao (University of Massachusetts Amherst) , Prashant Shenoy

Record 100769

Efficient Merging and Filtering Algorithms for Approximate String Searches Chen Li (University of California at Irvine) , Jiaheng Lu (CallT2 , UC Irvine) , Yiming Lu (UC Irvine)

Record 100770

Automatic Result Verification for the Functional Testing of a Query Language Carsten Binnig (University Heidelberg) , Donald Kossmann (ETH Zurich) , Eric Lo (ETH Zurich) , Angel Saenz - Badillos (Microsoft Corporation)

Record 100771

LOCUST : An Online Analytical Processing Framework for High Dimensional Classification of Data Streams Charu Aggarwal (IBM T . J . Watson Research Center) , Philip Yu (IBM T . J . Watson Research Center)

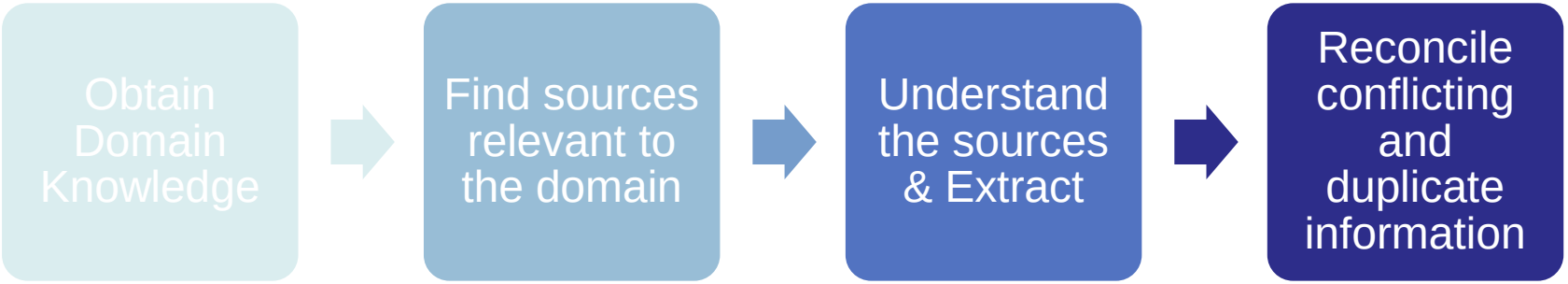
Record 100772

On Anti - Corruption Privacy Preserving Publication Yufei Tao (Chinese University of Hong Kong) , Xiaokui Xiao (Chinese University of Hong Kong) , Jiexing Li (Chinese University of Hong Kong) , Donghui Zhang (Northeastern University)

Record 100773

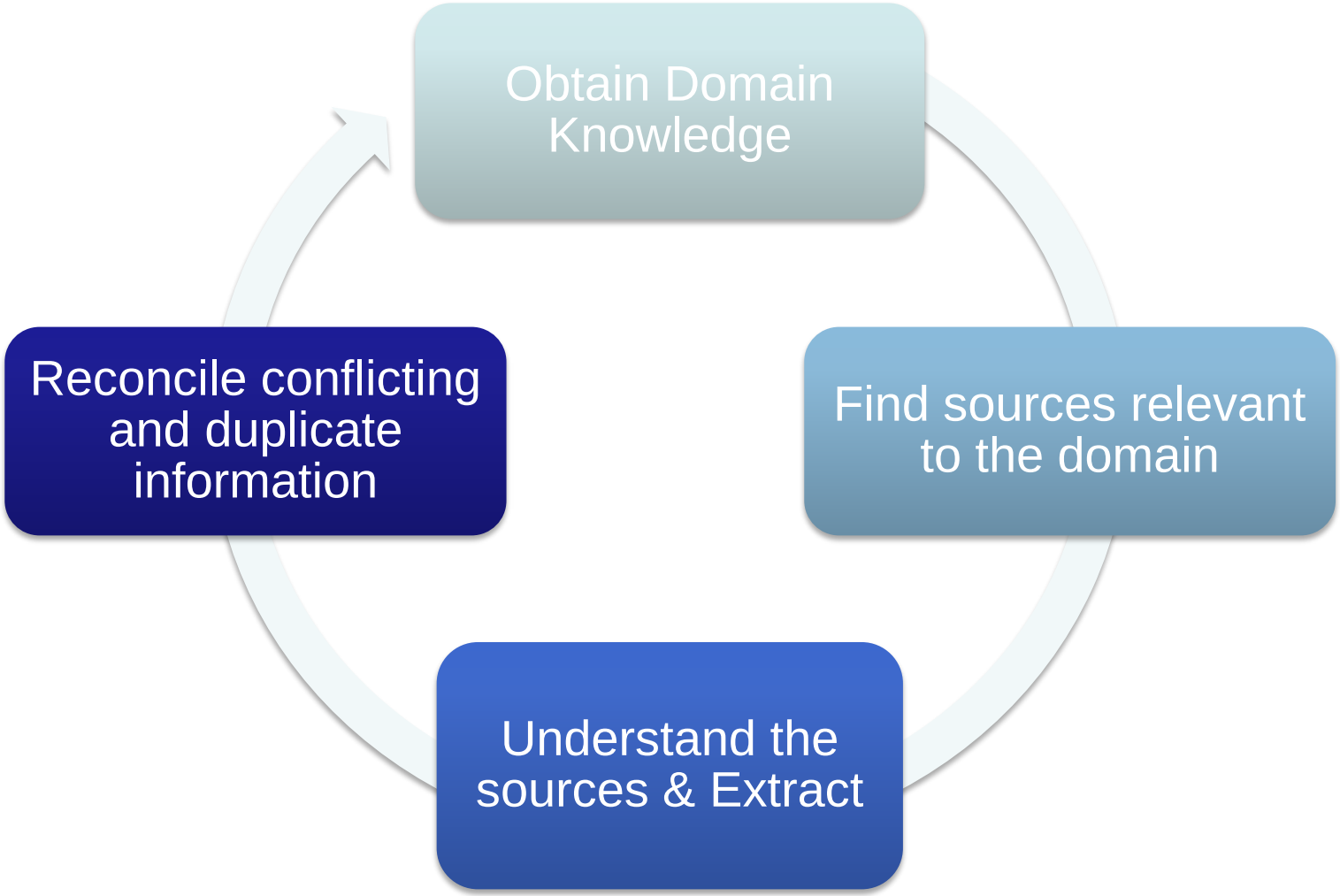
Faster Join Enumeration for Complex Queries Guido Moerkotte (University of Mannheim) , Thomas Neumann (Max - Planck - Institut Informatik)

Problem:
Insufficient information for each of these tasks in a new domain



Can we bootstrap with a few examples in each?

Solution:
Solve the Problems Jointly



Joint Extraction & Reconciliation

Underlying reasoning is fully probabilistic.

- Helps us automatically bootstrap from a few example entities or parsed pages.
- Helps us extract unknown entities.
- Helps us understand unseen pages that only talk about unknown/unpopular examples.

Cheap Supervision

- Leverage starting estimate of V_E, plus alternate forms found on initial sites, to allow noisy annotation of pages
  - [Canisius EMNLP 2007] “Bootstrapping Information Extraction with Field Books”
  - [Gupta PVLDB 2009] “Table Augmentation Queries”
- Similar supervision leveraged in [Elmeleegy PVLDB 2008] in a cross-domain fashion

29

Hadoop Core (Core, Pig, Oozie, Hive, Howl)

Ad BT and Inventory prediction, Content Agility, UDA, COKE, Mail Spam, Search, APG, Labs, Insights, Analytics

1+ million jobs per month
3.7 PB processed daily
90B events and 120 TB daily
70+ PB of Data

Map-Reduce and more …

HADOOP: SCALABLE ANALYTICS

Hadoop Applications @ Yahoo!

	2008	2009
Webmap	~70 hours runtime ~300 TB shuffling ~200 TB output	~73 hours runtime ~490 TB shuffling ~280 TB output +55% hardware
Terasort	209 seconds 1 Terabyte sorted 900 nodes	62 seconds 1 Terabyte, 1500 nodes 16.25 hours 1 Petabyte, 3700 nodes
Largest cluster	2000 nodes •6PB raw disk •16TB of RAM •16K CPUs	4000 nodes •16PB raw disk •64TB of RAM •32K CPUs •(40% faster CPUs too)

Example: User Activity Modeling

Large dimensionality vector describing possible user activities
But a typical user has a sparse activity vector

Attribute	Possible Values	Typical values per user
Pages	~ MM	10 – 100
Queries	~ 100s of MM	Few
Ads	~ 100s of thousands	10s

Hadoop pipeline to model user interests from activities

User Modeling Pipeline

Component	Data Processed	Time
Data Acquisition	~ 1 Tb per time period	2 – 3 hours
Feature and Target Generation	~ 1 Tb * Size of feature window	4 - 6 hours
Model Training	~ 50 - 100 Gb	1 – 2 hours for 100's of models
Scoring	~ 500 Gb	1 hour

Model Training

Supervised models trained using a variety of techniques:

- Regressions
  - Different flavors: Linear, Logistic, Poisson etc.
  - Constraints on weights
  - Different regularizations: L_1 and L_2
- Decision trees
  - Used for both regression and ranking problems
  - Boosted trees
- Naïve Bayes
- Support vector machines
  - Commonly used in text classification, query categorization etc.
- Online learning algorithms

Model Training

- Maximum Entropy modeling
- Log-linear link function.
- Classification problems in large dimensional, sparse features
- Constrained Random Fields
- Sequence labeling and named-entity recognition problems

Some of these algorithms are implemented in Mahout

Not all algorithms are easy to implement in MR framework

Hadoop Pipelines

- Pipeline workflows run repeatedly (e.g., daily, hourly)
- Incremental evaluation support needed
  - Semi-naïve style techniques can help
  - NOVA and other projects
- Soft real-time constraints

- Natural point to inject streaming analytics
- Key observation—Hadoop is being used as more than an analytics platform!
  - Data acquisition, warehouse
  - Lots to optimize here—e.g., # copies of shared files

Questions?