Information Extraction, Content Curation, and Machine Learning

Raghu Ramakrishnan

Yahoo! Fellow
Chief Scientist, Audience and Cloud Computing

(Many slides courtesy of others at Yahoo!)
Web of Pages

urls

token

julia roberts

mumbai

restaurant

chinese

san jose

WEB

INDEX

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

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

Mentions 1 - 10 out of 137

PC member of SIGMOD 2007
...Jeff Ullman Stanford University Patrick Valduriez INRIA Wei Wang University of North Carolina Gerhard Weikum Max-Planck Institute of Computer Sc. Till Westmann BEA Kyu Young Whang KAIST Jun Yang Yang Duke University...
http://sigmod07.riit.tsinghua.edu.cn/program.shtml

Monday
Oct 30, 2006

"PUBLICATIONS LIST:" page, Patrick O'Neil's homepage
...Invited paper to the IEEE Bulletin on Data Engineering, March 1994,
Data Access Method (LHAM)". Presented at the Fifth International...
http://www.cs.umb.edu/~poneil/publist.html

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

Saturday
Oct 21, 2006

PC member of SMDB 2007
...Canada) Ken Salem (University of Waterloo, Canada) Dennis Shasha
(New York University, USA) Gerhard Weikum (Max-Planck-Institut fur
...
Shower Head

Shower Head stores near Santa Clara, CA

1 Coniiff Plumbing Supply ★★★★★ (5)
coniiff.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)
kbsshowplace.com
(408) 249-9800 - 1200 Campbell Ave, San Jose, CA
Overview | 1 Photo | Directions

25 More Local Results...
Eggplant Parmigiana Restaurants near Baltimore, MD

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. Amici's 🌟🌟🌟🌟🌟 (20)
   aniccis.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
Eggplant Parmigiana Restaurants near Baltimore, MD

1. Pasticcio
   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...
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 ...
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
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:

\[
Q_E = \text{select name, phone, principal-name, api-score, start-date,.. (many more) from true-schools-in-world where type = “high school” and state = “CA”}
\]

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

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

• Approximate Joins for Data Centric XML.
  Nikolas Augustin (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 Abdolaziz (University of Waterloo), Daniel Zilio (IBM Toronto Lab), Fei Chiang (University of Toronto), Andrey Balmin (IBM Almaden Research Center), Kevin Beyers (IBM Almaden Research Center), Calisto Zuzarte (IBM Toronto Lab, Canada)

• XML Profiling as a String Matching Problem
  Christoph Koch (Saarland University), Stefanie Schierzinger (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, Anon 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
  Lin 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
  Kostas G*, (Purdue University), Michael Atallah, (Purdue University), Ashish Vari (Purdue University)
### Extracted Records

<table>
<thead>
<tr>
<th>Record</th>
<th>Title</th>
<th>Authors and Institutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>100765</td>
<td>Improving Information Access for a Community of Practice Using Business Process as Context</td>
<td>Yu Deng (IBM Research), Martha Devanakonda (IBM Research), Nithya Rajanani (IBM Research), Wladek Zadrozny (IBM Research)</td>
</tr>
<tr>
<td>100766</td>
<td>NAGA: Searching and Ranking Knowledge</td>
<td>Genni Kargel (Max Planck Institute), Fabian Suchanek (Max Planck Institute), Georgiana Itu (Max Planck Institute), Maya Ramanathan (Max Planck Institute)</td>
</tr>
<tr>
<td>100767</td>
<td>XOSDB: XQuery Scripting Extension in the AQUAlogic Data Services Platform</td>
<td>Daniel Engelnov (BEA Systems), Vianak Borkar (BEA Systems), Michael Carey (BEA Systems), Dmitry Lyashkin (BEA Systems)</td>
</tr>
<tr>
<td>100768</td>
<td>Efficient Data Interpretation and Compression over RFID Streams</td>
<td>Richard Cecchi (University of Massachusetts), Thanh Tran, Yanlei Xiao (University of Massachusetts Amherst), Prashant Shenoy</td>
</tr>
<tr>
<td>100769</td>
<td>Efficient Merging and Filtering Algorithms for Approximate String Searches</td>
<td>Chen Li (University of California at Irvine), Fabian Lu (Calit2, UC Irvine), Yiming Lu (UC Irvine)</td>
</tr>
<tr>
<td>100770</td>
<td>Automatic Restart Verification for the Functional Testing of a Query Language</td>
<td>Carsten Binnig (University Heidelberg), Donald Kossmann (ETH Zurich), Eric Lo (ETH Zurich), Angel S. Badillo (Microsoft Corporation)</td>
</tr>
<tr>
<td>100771</td>
<td>LOCUST: An Online Analytical Processing Framework for High Dimensional Classification of Data Streams</td>
<td>Chuan Anwaral (IBM T. J. Watson Research Center), Philip Yu (IBM T. J. Watson Research Center)</td>
</tr>
<tr>
<td>100772</td>
<td>On Anti - Corruption Privacy Preserving Publication</td>
<td>Yuefei Tao (Chinese University of Hong Kong), Xiaokai Xiao (Chinese University of Hong Kong), Jie Xing (Chinese University of Hong Kong), Donghui Zhang (Northeastern University)</td>
</tr>
<tr>
<td>100773</td>
<td>Faster Join Enumeration for Complex Queries</td>
<td>Guido Moerkotte (University of Mannheim), Thomas Neumann (Max Planck Institute for Informatics)</td>
</tr>
</tbody>
</table>
Problem: Insufficient information for each of these tasks in a new domain

- Obtain Domain Knowledge
- Find sources relevant to the domain
- Understand the sources & Extract
- Reconcile conflicting and duplicate information

Can we bootstrap with a few examples in each?
Solution:
Solve the Problems Jointly

1. Obtain Domain Knowledge
2. Find sources relevant to the domain
3. Understand the sources & Extract
4. Reconcile conflicting and duplicate information
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
<table>
<thead>
<tr>
<th>Hadoop Core (Core, Pig, Oozie, Hive, Howl)</th>
<th>Ad BT and Inventory prediction, Content Agility, UDA, COKE, Mail Spam, Search, APG, Labs, Insights, Analytics</th>
<th>1+ million jobs per month</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>3.7 PB processed daily</td>
</tr>
<tr>
<td></td>
<td></td>
<td>90B events and 120 TB daily</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70+ PB of Data</td>
</tr>
</tbody>
</table>

Map-Reduce and more …

**HADOOP: SCALABLE ANALYTICS**
<table>
<thead>
<tr>
<th></th>
<th>2008</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Webmap</strong></td>
<td>~70 hours runtime</td>
<td>~73 hours runtime</td>
</tr>
<tr>
<td></td>
<td>~300 TB shuffling</td>
<td>~490 TB shuffling</td>
</tr>
<tr>
<td></td>
<td>~200 TB output</td>
<td>~280 TB output</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+55% hardware</td>
</tr>
<tr>
<td><strong>Terasort</strong></td>
<td>209 seconds</td>
<td>62 seconds</td>
</tr>
<tr>
<td></td>
<td>1 Terabyte sorted</td>
<td>1 Terabyte, 1500 nodes</td>
</tr>
<tr>
<td></td>
<td>900 nodes</td>
<td>16.25 hours</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 Petabyte, 3700 nodes</td>
</tr>
<tr>
<td><strong>Largest cluster</strong></td>
<td>2000 nodes</td>
<td>4000 nodes</td>
</tr>
<tr>
<td></td>
<td>• 6PB raw disk</td>
<td>• 16PB raw disk</td>
</tr>
<tr>
<td></td>
<td>• 16TB of RAM</td>
<td>• 64TB of RAM</td>
</tr>
<tr>
<td></td>
<td>• 16K CPUs</td>
<td>• 32K CPUs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• (40% faster CPUs too)</td>
</tr>
</tbody>
</table>
Example: User Activity Modeling

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

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Possible Values</th>
<th>Typical values per user</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pages</td>
<td>~ MM</td>
<td>10 – 100</td>
</tr>
<tr>
<td>Queries</td>
<td>~ 100s of MM</td>
<td>Few</td>
</tr>
<tr>
<td>Ads</td>
<td>~ 100s of thousands</td>
<td>10s</td>
</tr>
</tbody>
</table>

Hadoop pipeline to model user interests from activities
## User Modeling Pipeline

<table>
<thead>
<tr>
<th>Component</th>
<th>Data Processed</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Acquisition</td>
<td>~ 1 Tb per time period</td>
<td>2 – 3 hours</td>
</tr>
<tr>
<td>Feature and Target Generation</td>
<td>~ 1 Tb * Size of feature window</td>
<td>4 - 6 hours</td>
</tr>
<tr>
<td>Model Training</td>
<td>~ 50 - 100 Gb</td>
<td>1 – 2 hours for 100’s of models</td>
</tr>
<tr>
<td>Scoring</td>
<td>~ 500 Gb</td>
<td>1 hour</td>
</tr>
</tbody>
</table>
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?