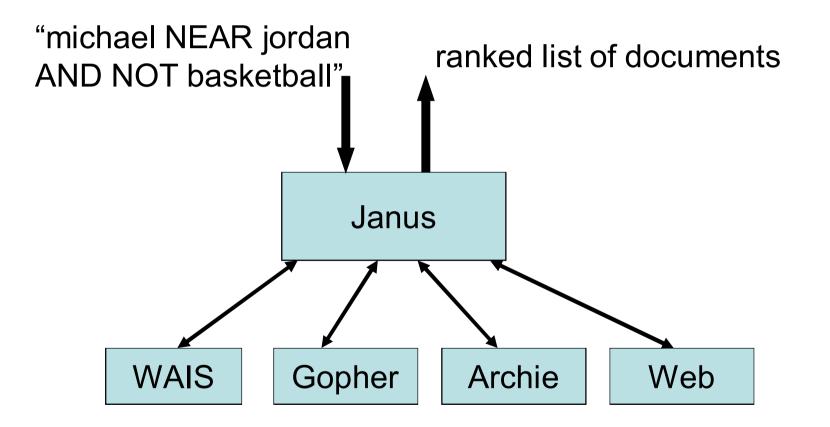
Harnessing the Deep Web

Anand Rajaraman

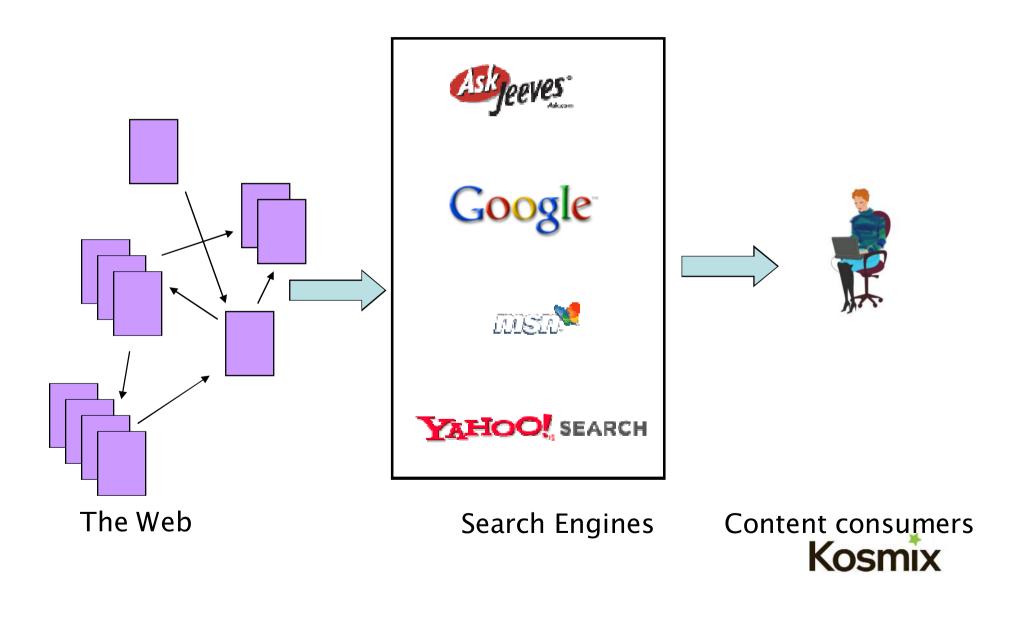


Flashback: PARC, Summer '94

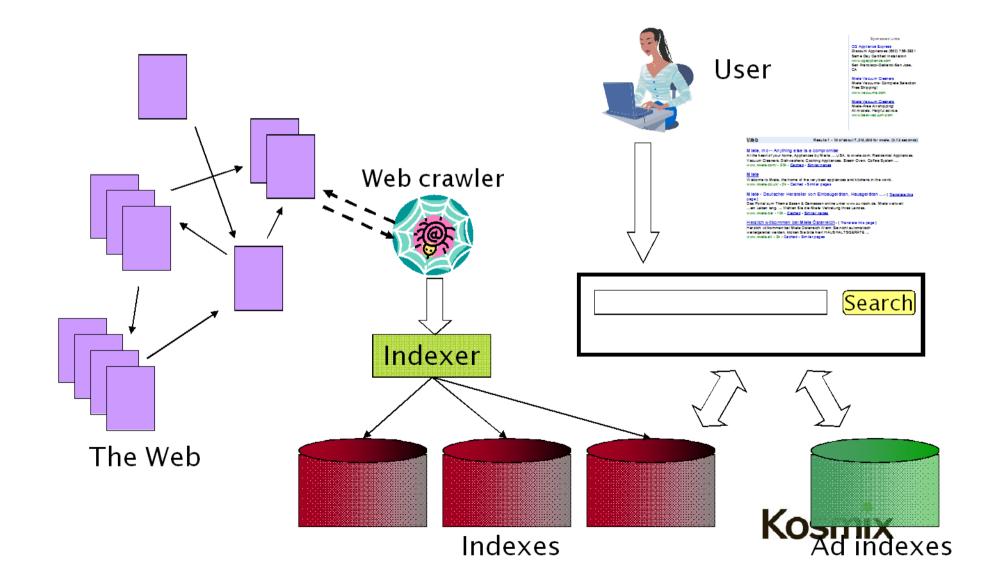




Searching the Web

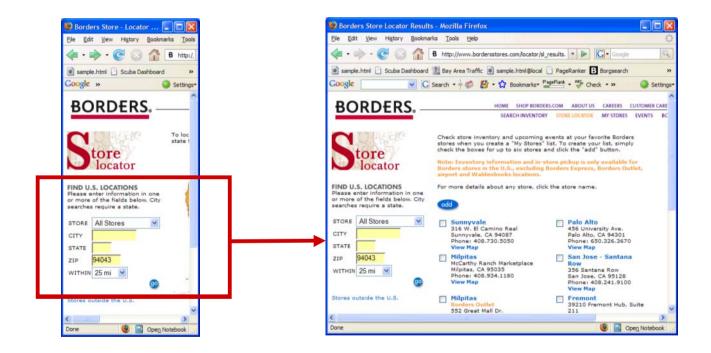


Search Engine 101



What is the Deep Web?

• Content hidden behind HTML forms



Deep = not accessible through search engine crawlers



The Deep Web

• Major gap in search engine coverage



Estimate: The Deep Web contains 500x the data found in the Surface Web

were strate and

all the state of the section of the section of the

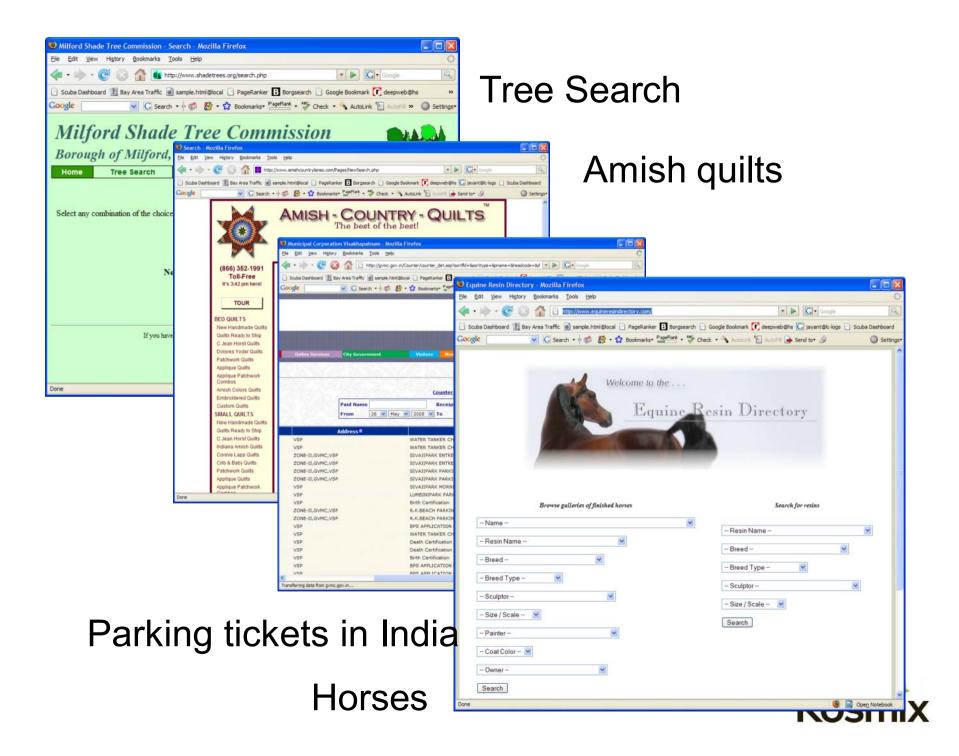
Much of this data is highly structured



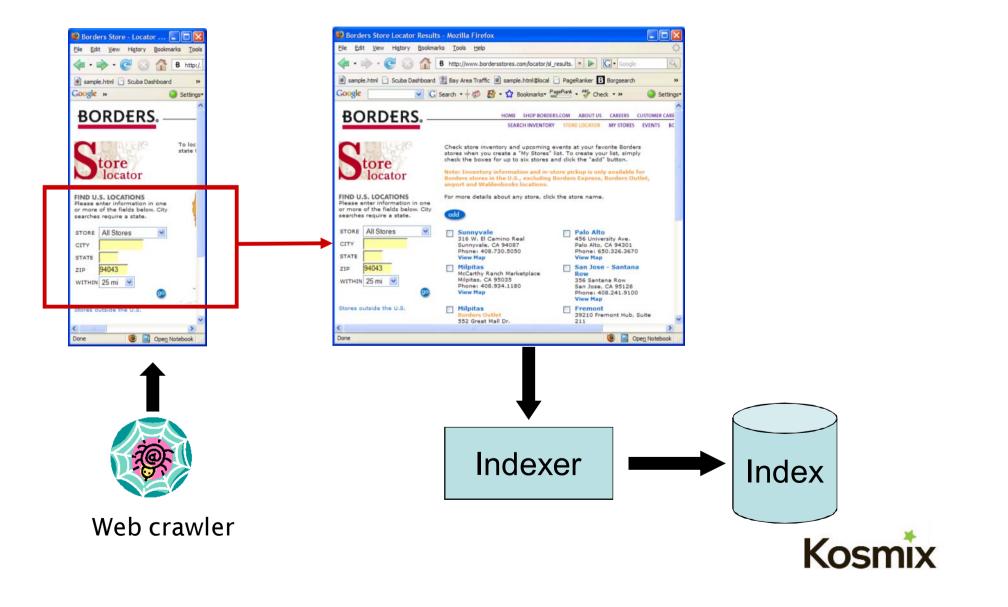
Harnessing the Deep Web

- Two different approaches to deep web:
 - Surfacing content: Google
 - Federated Approach: Kosmix
- Comparisons and future challenges





Surfacing the Deep Web



Surfacing

- At Crawl time
 - Precompute "interesting" form submissions
 - Insert resultimg pages into Google index
- At Query time
 - Nothing!
 - Pages are already in the index



Surfacing Challenges [Madhavan et al., VLDB 2008]

- 1. Predicting the correct input combinations
 - Generating all possible URLs is wasteful and unnecessary
 - Cars.com has ~500K listings, but 250M possible queries
- 2. Predicting the appropriate values for text inputs
 - Valid input values are required for retrieving data
 - Ingredients in recipes.com and zipcodes in borderstores.com
- 3. Don't do anything bad!



Informative Query Templates



Result pages different 🗲 informative

http://jobs.shrm.org/search?state=All&kw=&type=All
http://jobs.shrm.org/search?state=AL&kw=&type=All
http://jobs.shrm.org/search?state=AK&kw=&type=All

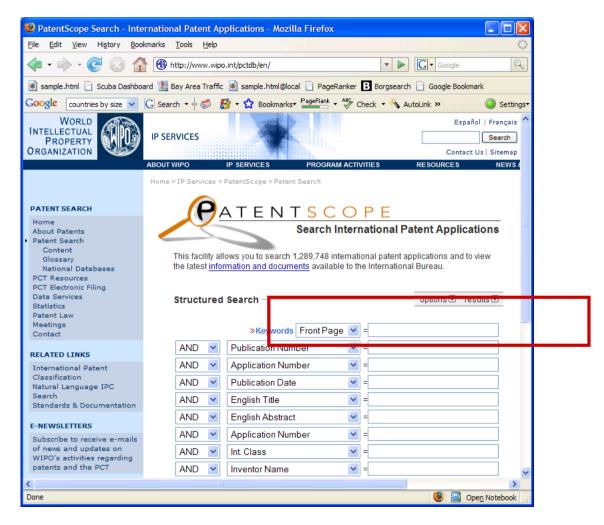
http://jobs.shrm.org/search?state=WV&kw=&type=All

Result pages similar → un-informative

http://jobs.shrm.org/search?state=All&kw=&type=ALL http://jobs.shrm.org/search?state=All&kw=&type=ANY http://jobs.shrm.org/search?state=All&kw=&type=EXACT



Example: www.wipo.int



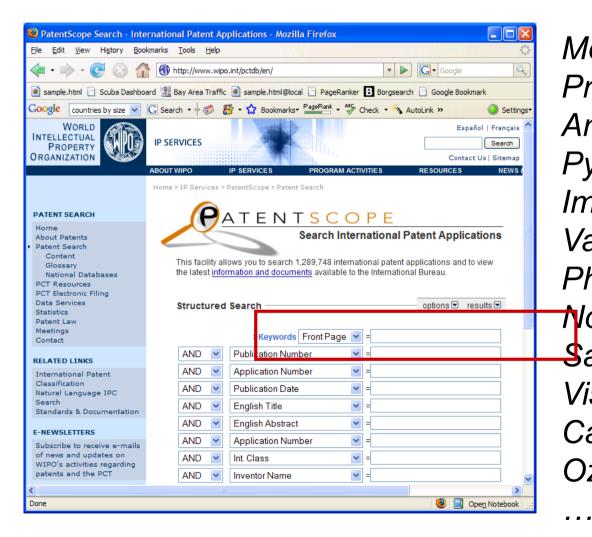


Input values for Generic Search

- Iterative Probing for search boxes
 - Select an initial list of candidate keywords
 - Download pages based on current set of keywords
 - Extract more candidate keywords from result pages
 - Refine the current set of keywords
 - Repeat until no more new candidate keywords
 - Prune list of candidate keywords
- Related Work:
 - Classifying Deep-Web sources [Ipeirotis+]
 - Extracting text documents [Ntoulas+, Barbosa+]



Example: www.wipo.int



Metalworking Protein Antibody Pyrazole Immobilizer Vasoconstriction **Phosphinates** Nosepiece *Sandbridge* Viscosity Carboxydiphenylsulphide Ozonizer



Impact on Query Stream [Madhavan et al, VLDB 2008]

- Crawled ~3M sites
- 50 languages, hundreds of domains
- 1000 queries per-second get results from the deep web!
- Impact mostly on the long and heavy tail of queries



A Closer Look





Structured Data!

| File Edd | Name Internet I | Socimatia Tools telo | | | and the second se | 20 |
|------------------|--|---|--|--------------------|---|---------|
| - | | and the second se | | - | - | |
| 4.19 | .60 | 1 Shate://www.wipe.int/pe | titb/en/ | | GP back | 194 |
| a sample.ht | tel 🗋 Soube Deal | rboard 🔠 Bay Area Traffic 🔳 s | anpie.html@local | PageRanker | Borgsearth | |
| Google [| 0 | C Search + + 1 1 1 1 | 🗘 Bookmarker 1 | · | Deck + # @ Se | minger. |
| a | | A Designed | | | Baselui (Poerçais | 1 |
| 80) × | SERVICES | 1994 C | | | Tearsh | |
| | | and the second | | | Context (in) Sileman | |
| ADC | AUT HERE'S | # SERVICES PROGRAM | ACTIVITES | RESOURCES | | |
| - | ine p. 17 Januara p. Pat | te-Marger p. Palat-1 Saainti | | | | |
| | 0 | | | | | |
| | P | ATENTSC | OPE | | | |
| | | | | | | |
| | | | | | | |
| | | 503 | ch Internatio | onal Patent A | opplications | |
| | <u> </u> | | | | | |
| | | Drive you to search 1,314,348 mem | utional patent app | | | |
| | | | utional patent app | | | |
| | enformation, an | ows you're search 1,314,348 men d <u>Societiants</u> av aliacle to the bilants | utional patent app | pications and to v | ev: the latest | |
| | | ows you're search 1,314,348 men d <u>Societiants</u> av aliacle to the bilants | utional patent app | | ev: the latest | |
| | enformation, an | ows you're search 1,314,348 men d <u>Societiants</u> av aliacle to the bilants | utional patent app | pications and to v | ev: the latest | |
| | enformation, an | own you in search 1,214,348 men d socurrents available to the bland Search | utional patent app | pications and to v | ev: the latest | |
| | Education an | ows you'r search 1,314,348 mer d <u>documenta</u> awliadae to the blant Search » Keywenta Fort Page | utional patent app | pications and to v | ev: the latest | |
| | Structured 1 | even you in search 1,314,346 inter d accuments events to the blank Search 3 Keyvennis Front Page Rabication Namber | utional patent app | pications and to v | ev: the latest | |
| | Structured 1 | ever yets to search 1,314,348 inter d accounts events to be been search skeywards Fort Page / Publication Number | utional patent app | pications and to v | ev: the latest | |
| | Structured 1 | even you be search 1.314.348 intern d documents evaluatie to the bient Search Publication Number Publication Number Publication Date | utional patent app | pications and to v | ev: the latest | |
| | Structured Structured MND & | erres you to asserch 1.314,348 intern discusses and available to the bismo Search Publication Number Application Number Faddication Date 8 English Tate 2 | utional patent app | pications and to v | ev: the latest | |
| land Marit | NID 4 NID 4 NID 4 NID 4 NID 4 NID 4 NID 4 NID 4 | ama pisi in selacini 1.314.340 mini di <u>discutenti si vivalisi in the bienni</u> bearch <u>Publication Number</u> <u>Rabication Sumber</u> <u>Rabication Case a</u> <u>English Restaut</u> | utional patent app | pications and to v | ev: the latest | |
| | NID 4 NID 4 NID 4 NID 4 NID 4 NID 4 NID 4 NID 4 | ever pici in statuti 1.314.348 inter d. Boucesta evaluate is the inter- sector Palacian Nanter Apolication Nanter English Konas English Konas English Konas | utional patent app | pications and to v | ev: the latest | |
| tion nails of | AND W AND W AND W AND W AND W AND W AND W AND W | Inter pis in BARCIN 1.314.348 men d Reconstitute volacies is the Baren Bearch Subjection Number Application Number English Network English Network English Network English Network English Network English Network | vational patient spo atomal Duriess | pications and to v | ev: the latest | |



| 🕑 Borders Store - Locator - Mozilla Fire | efex 🖉 🖾 🖾 |
|---|--|
| Die gatt giew Higtory Bookmarks Too | a tap O |
| 🐗 = 🕪 - 💽 🕢 🟠 🗷 http: | (hown bordersstares.com/locator/locator + 🕼 🚱 - Sampe |
| annoie.html | wea Traffic 🗃 sample html@local 📋 PageRanker 🖪 Borgsearch 🔹 |
| | B - D Sockmeiter Papeline - T Check + = O Settings |
| BORDERS. | HOWE SHOP BORDERLOOM ABOUT US CAREERS CUITOMER |
| BORDERS. | SLARCH INVENTIORY STORE LOCATOR ANY STORES EVENTS |
| Diocator | A state of the sta |
| FIND U.S. LOCATIONS | |
| Please enter information in one or more of the fields below. City searches require a state. | |
| STORE All Stores 💌 | |
| CITY | |
| STATE | and the second se |
| ZIP | |
| WITHIN 25 mi 🕑 | |
| Dares suitaide the U.S. | |
| BUT CARDS | |
| SPECIAL OFFERS | |
| GROUP SAVINGS & SERVICES | |
| BORDERS VISA | |
| Done | Spen hutebook |

| Popular Ba | by Names | 1880 to 2008 | Social Se |
|-------------------------|--------------------------|---------------------|-----------|
| 2 | , | | |
| File View Edit V | isualize Merge | | |
| Current view: All - Sho | w options (filter/aggreg | ate/choose columns) | |
| rank 🗸 | Male name 🔻 | Female name - | year |
| 1 | John | Mary | 1880 |
| 2 | William | Anna | 1880 |
| 3 | James | Emma | 1880 |
| 4 | Charles | Elizabeth | 1880 |
| 5 | George | Minnie | 1880 |

A Different Approach

Then

• Explosion of documents



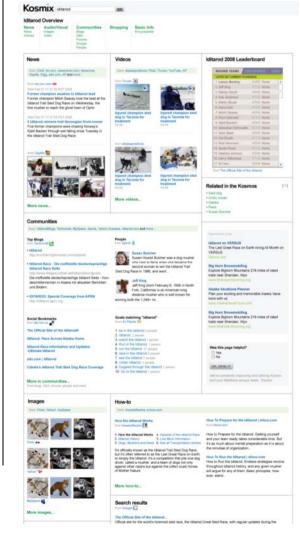
- But the documents were primarily of the same information type
- Search engines surfaced the top 10 relevant documents for the query with a snippet for each document

Now

• Explosion of structured information



Solution Kosmix Topic Pages



Kosmix: Introducing Topic Pages

Search Results = Find needle in haystack

Caltrain schedule Bi Rite Creamery San Francisco

Topic Page = Get 360-degree overview of topic

Kate Winslet Citizen Kane Cholesterol



Kosmix Demo



Consumer Validation -- RightHealth

- Second most visited health site on the Web per Hitwise
- 6.4 million monthly unique users per comScore

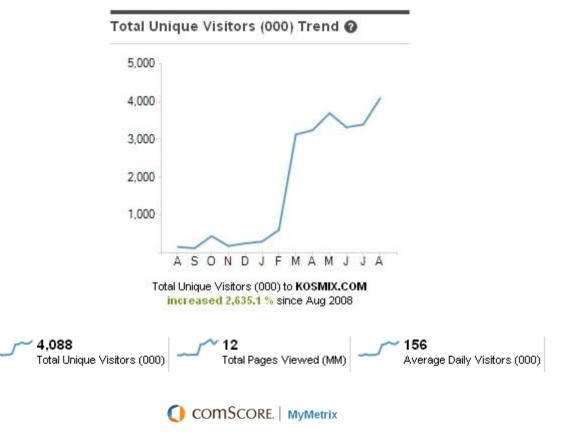
| Health and Medical - Information | | | |
|----------------------------------|---------------------------|----------|--|
| | Websites (2,737 returned) | Visits 🔻 | |
| 1 | WebMD | 10.20% 💻 | |
| 2 | Righthealth.com | 5.67% | |
| ≤ 3 | MedicineNet | 3.48% 💻 | |
| v 4 | MayoClinic.com | 3.48% 💻 | |
| ≙ 5 | Drugs.com | 3.33% 💻 | |
| ₹6 | MSN Health | 3.23% 💻 | |
| <u>∽</u> 7 | AOL Health | 2.58% 💻 | |
| 78 | Yahoo! Health | 2.05% 💻 | |
| <u>~</u> 9 | QualityHealth.com | 1,66% 💻 | |
| 7 10 | MedlinePlus | 1.66% 💻 | |

hitwise



Consumer Validation -- Kosmix

Kosmix.com was launched in Dec 2008, has already grown to 4.1M uniques*



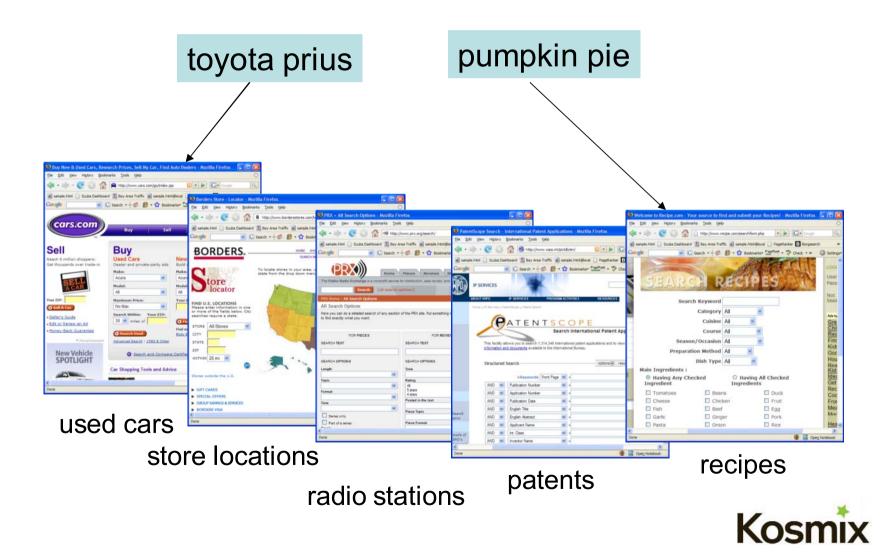


Kosmix [#]Health.

- 10 million unique visitors/month
- 35 million searches/month
- Rapid growth over the past few months



Federated Approach



Federated approach to the Deep Web

- It is hard to separate structured data from its applications
 - 15 years of data integration research & experience
- Let domain specialists create and maintain structured data

– And expose it as web services

Route queries to specialist services



Trends Favor the Federated Approach

- Social Media
- Real-Time Information
- Specialized search engines
- Innovative visualizations
- Business Model issues
- Algorithmic Content
- Availability of APIs

flickr

en





twitter



Deep Web Driver: Social Media



- Content volume grows rapidly
- Access controls can prevent indexing
- Opportunities for personalization



Deep Web Driver: Real-Time





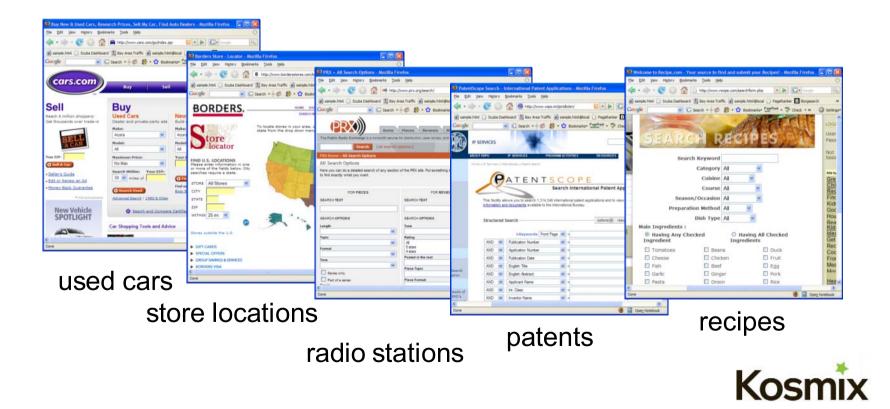


- Earthquake in China (2008)
- US Airways 1549 Hudson landing (2009)
- Iran elections (2009)



Specialized Search Engines

It's a shame to take all this richness and compress it into 10 results links!



Deep Web Driver: Specialized Search Engines

- Vertical search engines exist for several domains
 - Travel, real estate, autos, ...
- They use specialized ranking functions
 Very different from standard web search
- And innovative visualizations of results
 - Going beyond 10 links per page

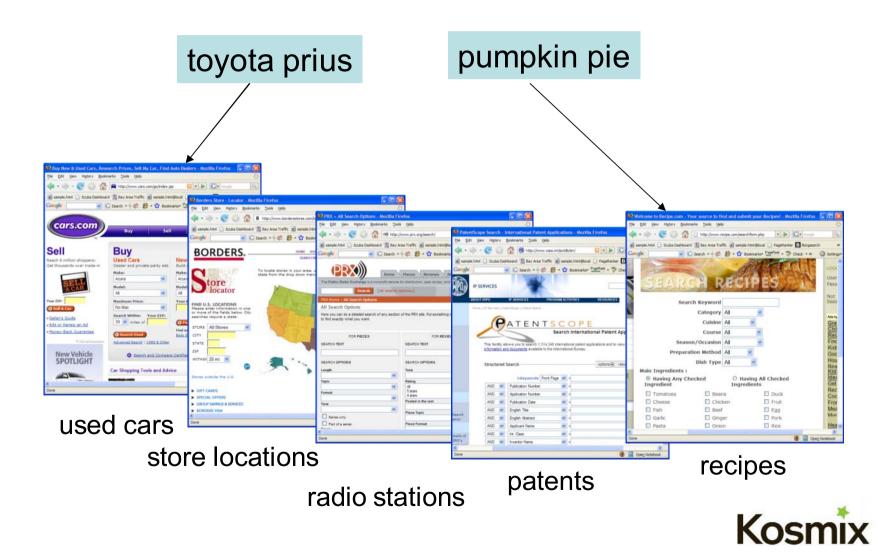


Challenges for the federated approach

- Data source integration
- Data source selection
- Query transformation
- Results Layout
- Performance



Data Source Selection

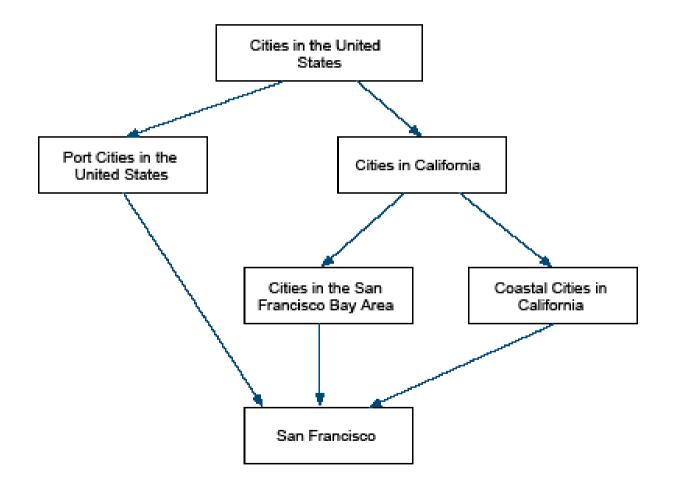


Data Source Selection

- The system knows about 1000s of data sources
- Cannot access each source for every query
 - Unacceptable load on service providers
- Need to select a small set of sources for each query
 - Without access to the source data



The Taxonomy





The Taxonomy

- Millions of nodes arranged in a directed acyclic graph (DAG)
 - Captures ISA relationships
 - And many other domain-specific relationships
- Created over 3 years using a combination of algorithms and human curation
 - Mined from dozens of sources, including Wikipedia, DMOZ, and deep vertical sources
 - Refreshed on a daily basis to capture new entities and relationships



Tagging Data Sources

- Data sources are "tagged" using taxonomy nodes
- A data source can be tagged to several nodes in taxonomy, at different levels

Last.fm \rightarrow Music Epicurious \rightarrow Recipes (under Food & Wine) MetroLyrics \rightarrow Lyrics (under Music) TripAdvisor \rightarrow Hotels (under Travel)



Query Categorization

- Kosmix Categorization Service (KCS)
- Given a query, KCS identifies the taxonomy nodes related to the query

Taylor Swift → Music
Pinot Noir → Wine, Viticulture (+ many others)
Eiffel Tower → Monuments, Tourist Attractions,
History of France (+ many

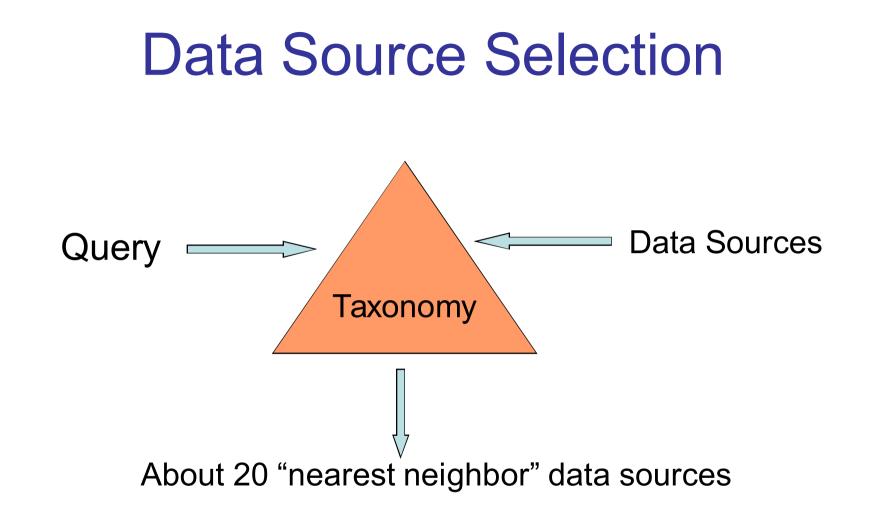
others)



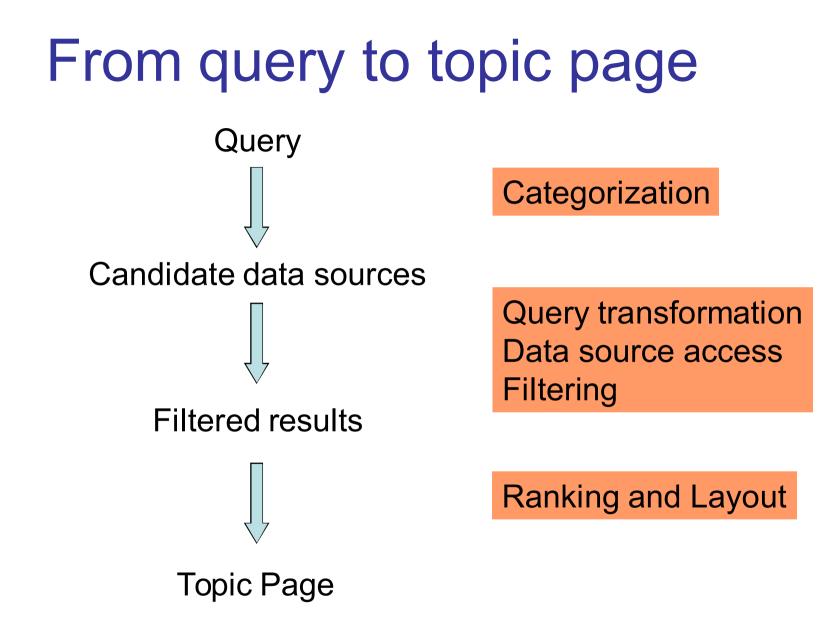
KCS output for Pinot Noir













Read VLDB 2009 paper for:

- Hybrid approach
 - Combines Indexing and Federation
- Taxonomy
- Categorization
- Query transformation
- Disambiguation
- Filtering
- Performance



Comparing the approaches

- Surfacing is great for:
 - Simplicity
 - Coverage
 - Performance
- Federated approach better for:
 - Real-time information
 - Aggregating specialized search engines
 - Social media



Challenges

- Can we combine the advantages of the surfacing and federated approaches?
- Is there a way to scale the federated approach rapidly?
- How can we surface data with more structure?



References

- Jayant Madhavan et al. Google's Deep-Web Crawl. VLDB 2008.
- Anand Rajaraman. *Kosmix: High Performance Topic Exploration using the Deep Web.* VLDB 2009.

