Query Categorization at Scale

MAGNE+IC™

About Magnetic

MAGNETIC™

- First company to focus 100% on applying search intent to display
- Proprietary media platform and targeting algorithm
- One of the largest aggregators of intent data
- Display, mobile, video and social retargeting capabilities
- Strong solution for customer acquisition and retention

AKQA · Chrysler · Staples · UM · Digitas · Starwood Hotels and Resorts · OMD · Starcom MediaVest Group · Mediacom · Initiative · FedEx · General Mills · ZenithOptimedia · Allstate · Carat

FOCUS

2

Search Retargeting

MAGNETIC™

Search retargeting combines the purchase intent from search with the scale from display

1) Magnetic collects search data

2) Magnetic builds audience segments

3) Magnetic serves retargeted ads

4) Magnetic optimizes campaign

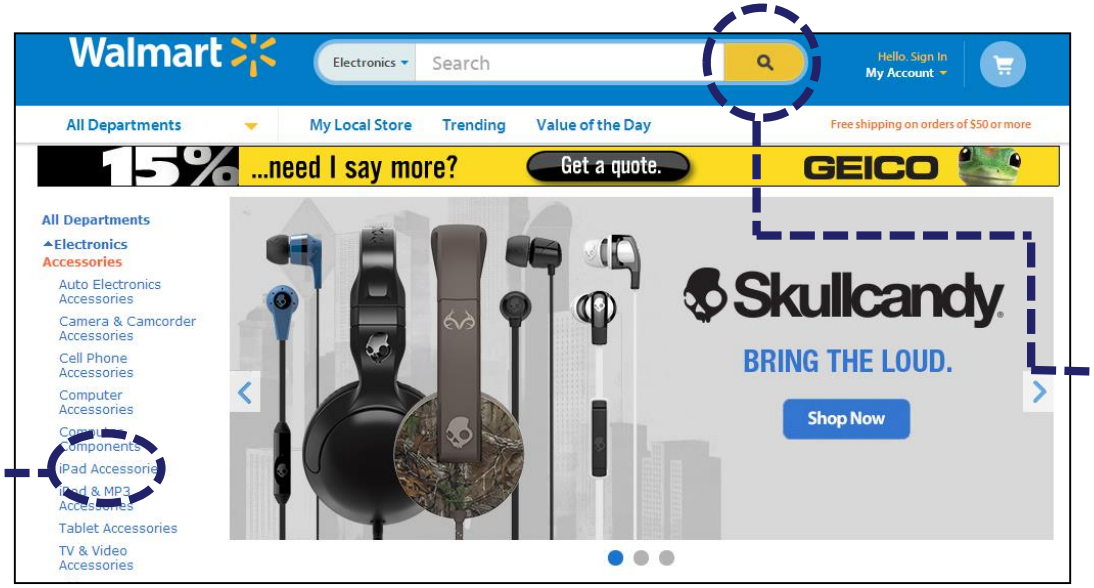
FOCUS

Search Data - Natural and Navigational

MAGNETIC™

Natural Searches and Navigational Searches

Navigational Search: **"iPad Accessories"**

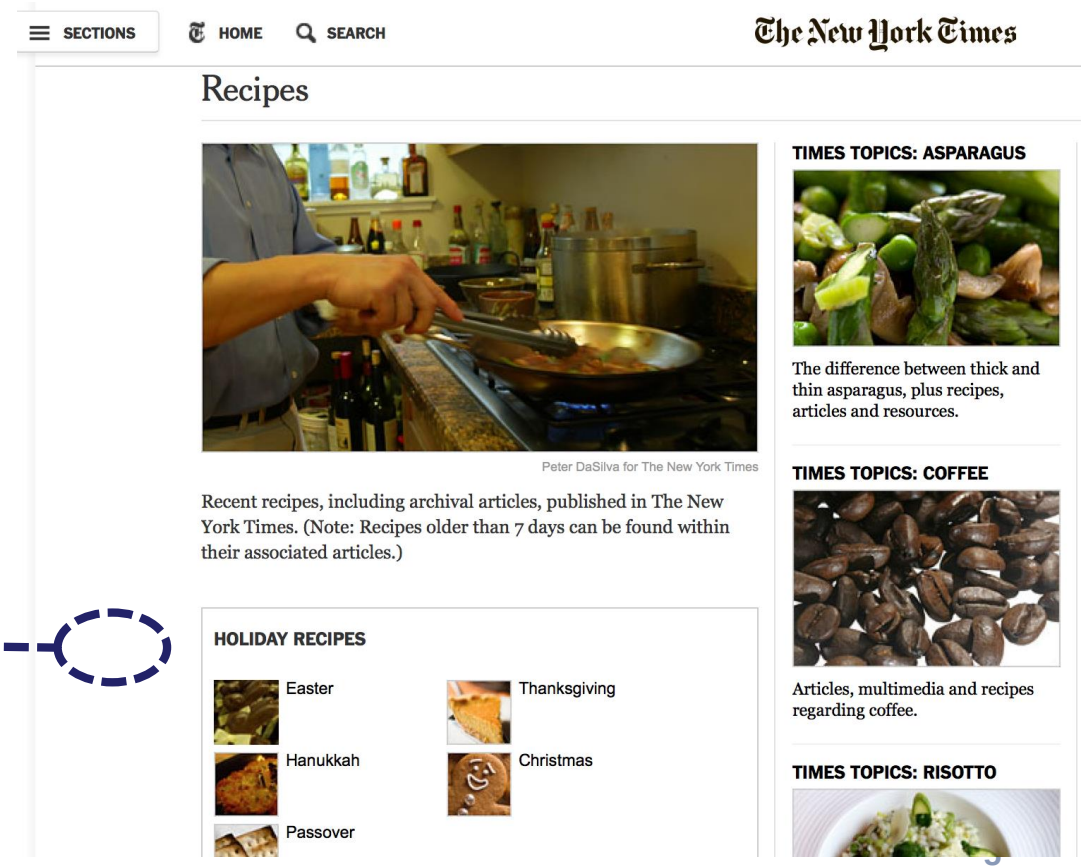


Natural Search: **"iPhone"**

Search Data – Page Keywords

MAGNETIC™

Page keywords from article metadata: **"Recipes, Cooking, Holiday Recipes"**



FOCUS

Search Data – Page Keywords

MAGNETIC™

Article Titles:
“Microsoft is said to be in talks to acquire Minecraft”

Search Data – Why Categorize?

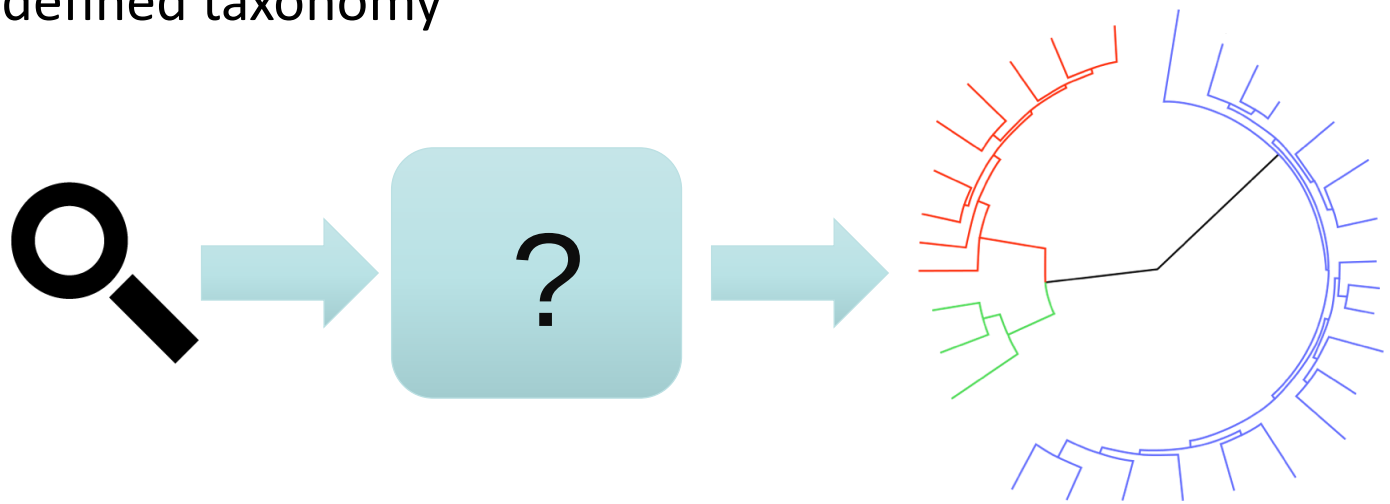
MAGNE+IC™

- Targeting categories instead of keywords = Scale
- Use category name to optimize advertising as an additional feature in predictive models
- Reporting by category is easier to grasp as compared to reporting by keyword

Query Categorization Problem

MAGNETIC™

- Input: Query
- Output: Classification into a predefined taxonomy



FOCUS

Query Categorization – Academic Approach

MAGNE+IC™

- Usual approach (academic publications):
  - Get documents from a web search
  - Classify based on the retrieved documents

FOCUS

Query Categorization

MAGNETIC™

- Usual approach:
  - Get results for query
  - Categorize returned documents
- Best algorithms work with the entire web (search API)

Query	Categories
apple	Computers \ Hardware Living \ Food & Cooking
FIFA 2006	Sports \ Soccer Sports \ Schedules & Tickets Entertainment \ Games & Toys
cheesecake recipes	Living \ Food & Cooking Information \ Arts & Humanities
friendships poem	Information \ Arts & Humanities Living \ Dating & Relationships

Long Time Ago …

MAGNE+IC™

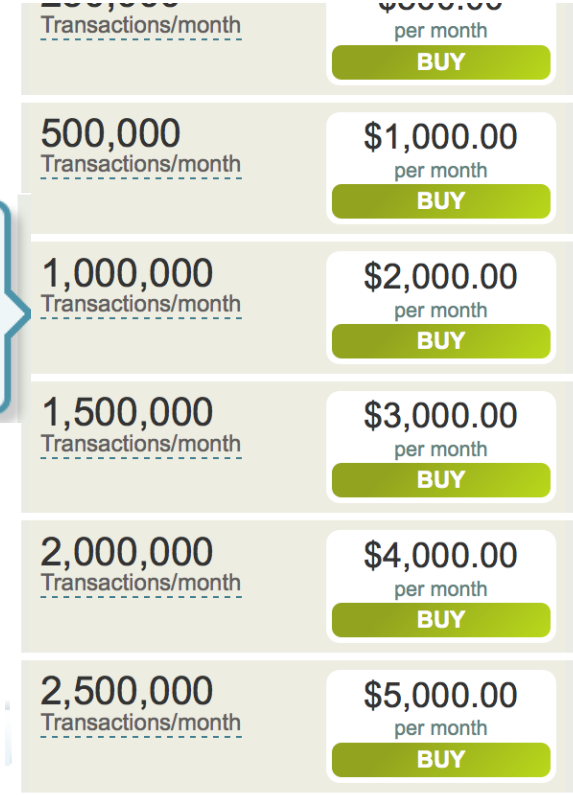
- Relying on Bing Search API:
  - Get search results using the query we want to categorize
  - See if some category-specific “characteristic“ keywords appear in the results
  - Combine scores
  - Not too bad....

Long Time Ago …

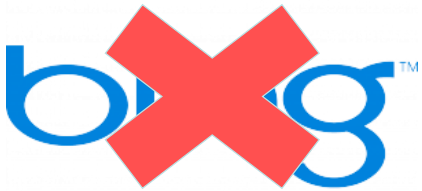
MAGNETIC™

- … But ….

A transaction is one request that returns one page of results. Retrieving multiple pages will result in multiple transactions executed.



- …. We have ~8Bn queries per month to categorize ….
- $2,000 * 8,000 = Oh My!



Our Query Categorization Approach

MAGNETIC™

- Use web replacement – Wikipedia
  - 4.5 million articles
  - 9 million of unique titles
  - 40 GB text
- DBPedia
  - Good categories for articles
  - Additional structured data
- Freebase
  - 170 GB triples
  - 40 million topics

DBpedia

Freebase

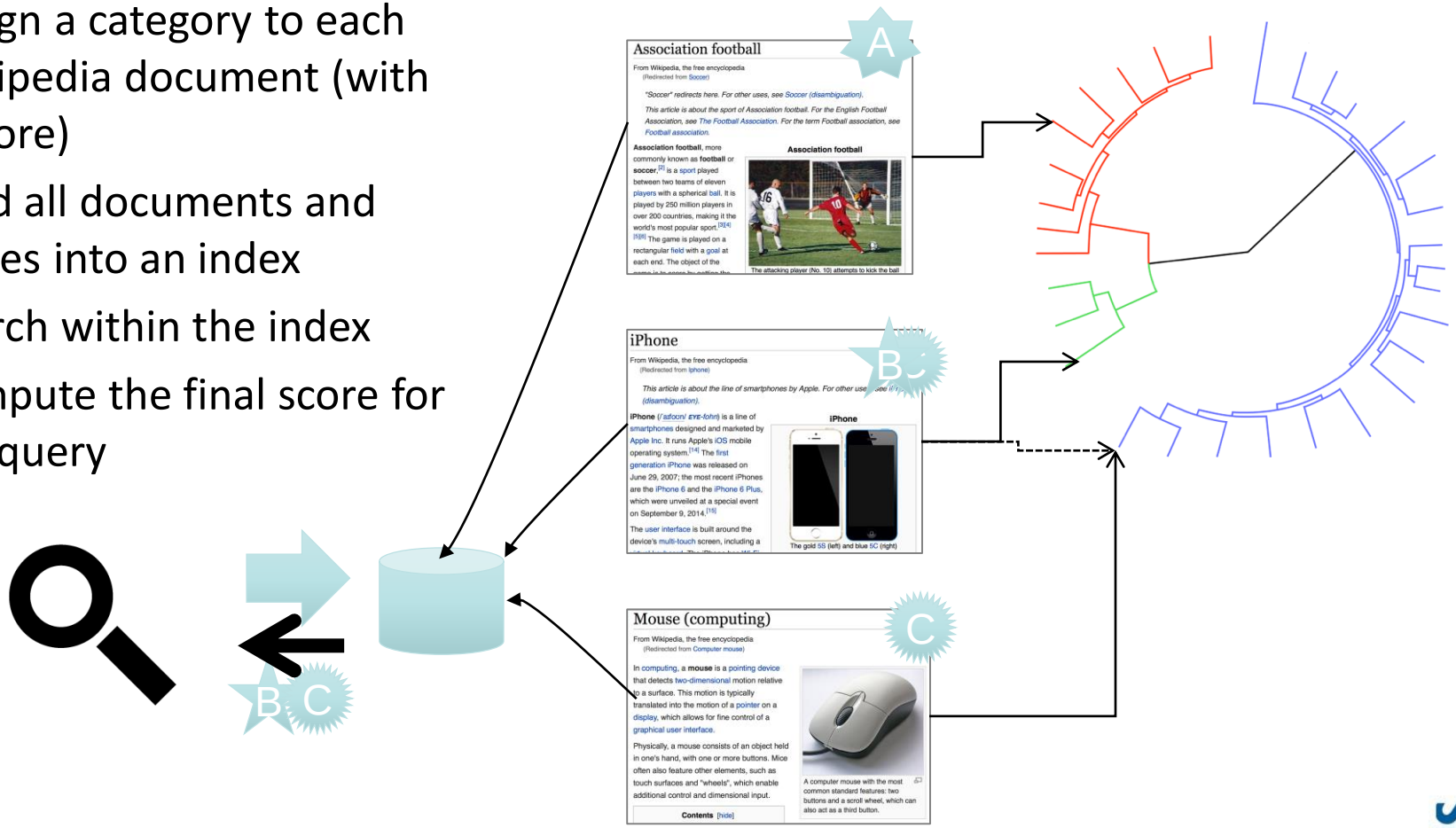
All available for download!

FOCUS

Our Query Categorization Approach

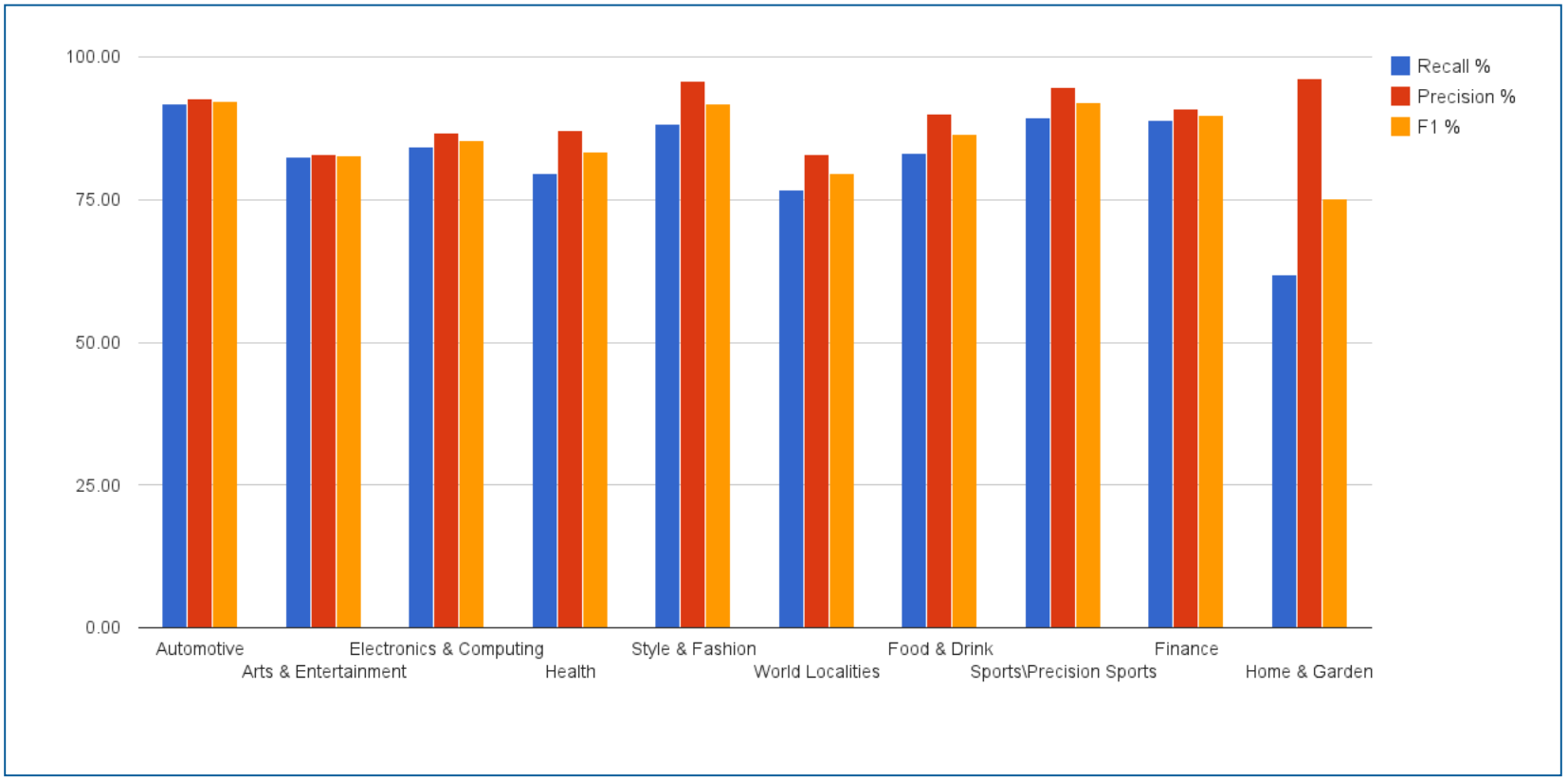
MAGNETIC™

- Assign a category to each Wikipedia document (with a score)
- Load all documents and scores into an index
- Search within the index
- Compute the final score for the query



Results Quality - Precision/Recall

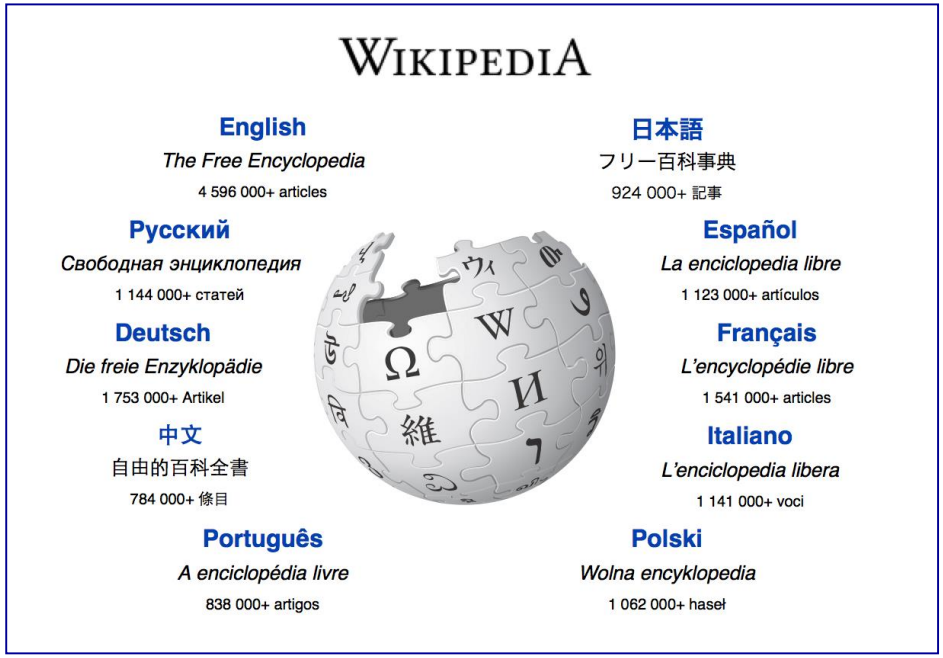
MAGNE+IC™



Categorizing Other Languages

MAGNETIC™

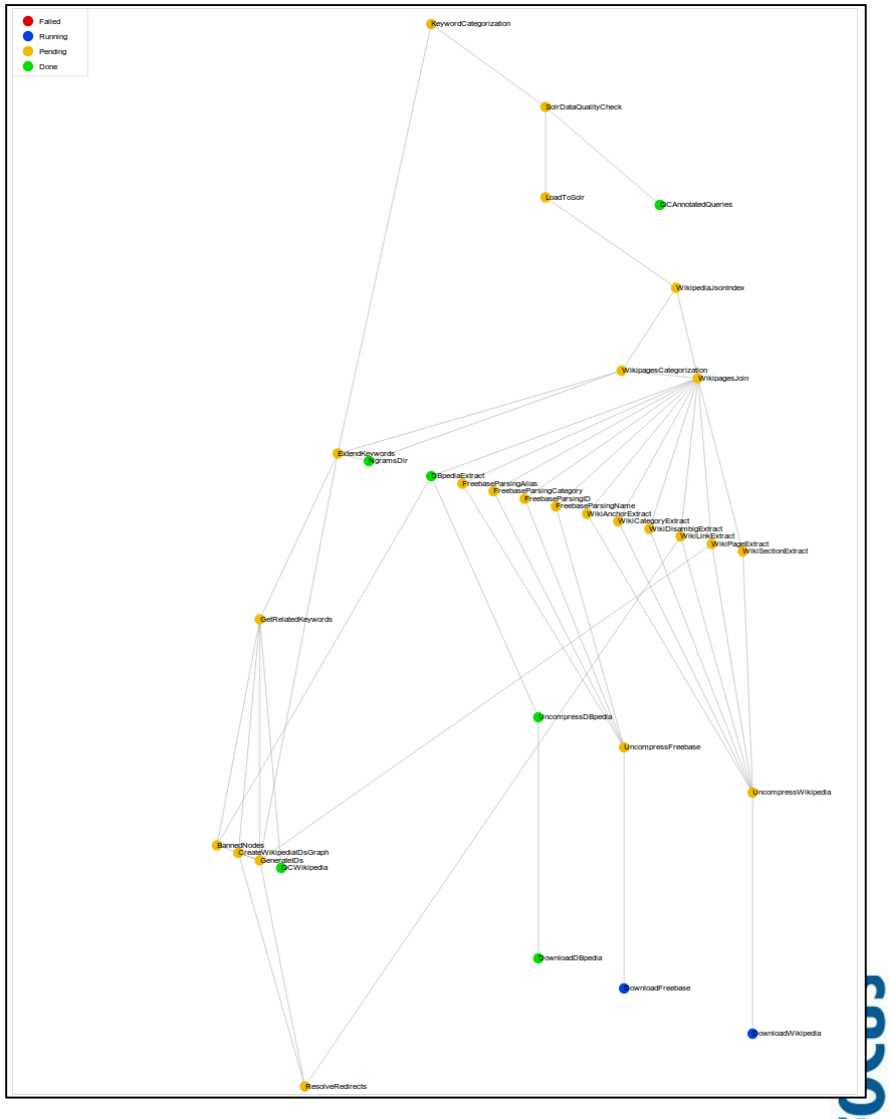
- In development
- Combining indexes for multiple languages into one common index
- Focus:
  - Spanish
  - French
  - German
  - Portuguese
  - Dutch



Preprocessing Workflow

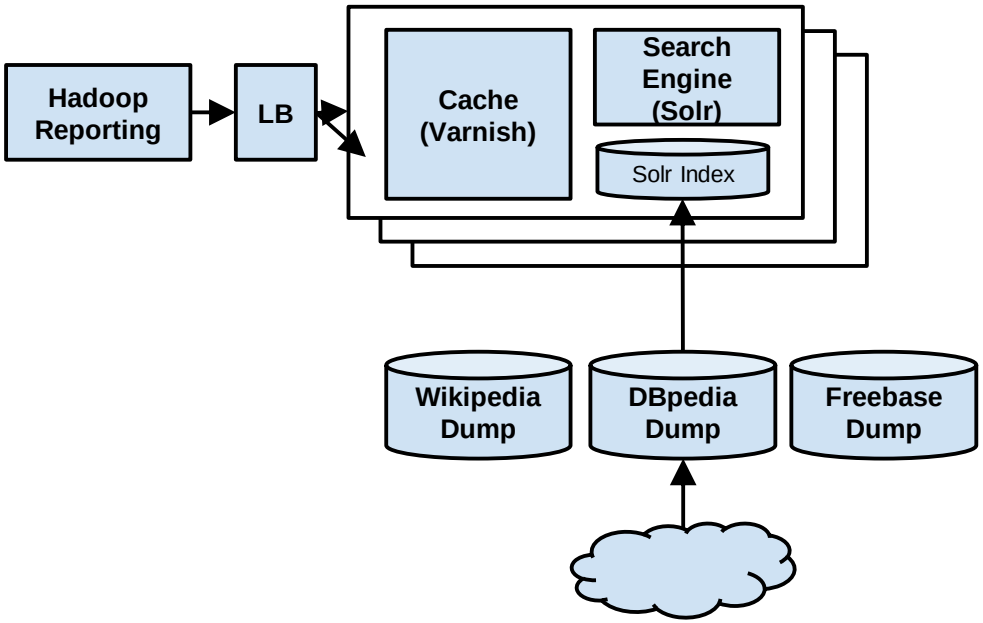
MAGNETIC™

- Automatized Hadoop and local jobs
- Luigi library and scheduler
- Steps:
  - Download
  - Uncompressing
  - Parse Wikipedia/Freebase/DBpedia
  - Generate N-grams
  - JOIN together (Wikipage) – JSON per page
  - Preprocess Wikipage categories
  - Produce JSON for Solr or local index
  - Load to SOLR + Check quality



Query Categorization: Scale

MAGNETIC™



- Scale is achieved by combination of multiple categorization boxes, load balancing, and Varnish (open source) cache layer in front of Solr
- We have 6 servers in production today
- Load Balancer - HAProxy
- Capacity – 1,000 QPS/server
- More servers can be added if needed

VARNISH CACHE

Lucene Apache Solr

Architected for Scale

MAGNETIC™

- Bidders, AdServers developed in Python and use PyPy VM with JIT
- Response time critical - typically under 100ms as measured by exchange
- High volume of auctions – 200,000 QPS at peak
- Hadoop – 25 nodes cluster
- 3 DC – US East, West and London
- Data centers have multiple load balancers – HAProxy
- Overview of servers in production:
    - US East: 6LB, 45 Bidders, 6 AdServers, 4 trackers, 25 Hadoop, 9 Hbase, 8 Kyoto DB
    - US West: 3LB, 17 Bidders, 6 AdServers, 4 trackers, 4 Kyoto DB
    - London: 8 Bidders, 2 AdServers, 2 trackers, 4 Kyoto DB

ERD Challenge

MAGNETIC™

- ERD'14: Entity Recognition and Disambiguation Challenge
- Organized as a workshop at SIGIR 2014 Gold Coast
- Goal: Submit working systems that identify the entities mentioned in text
- We participated in the “Short Text” track
  - Entities (people, locations, organizations …) in queries
- 19 team participated in the challenge
- We took 4th place

Ďakujeme za pozornosť

MAGNETIC™

eFOCUS

© 2010 Digit, s.r.o.