

Query Categorization at Scale

e FOCUS

MAGNETIC™

About Magnetic

MAGNE+IC™



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®

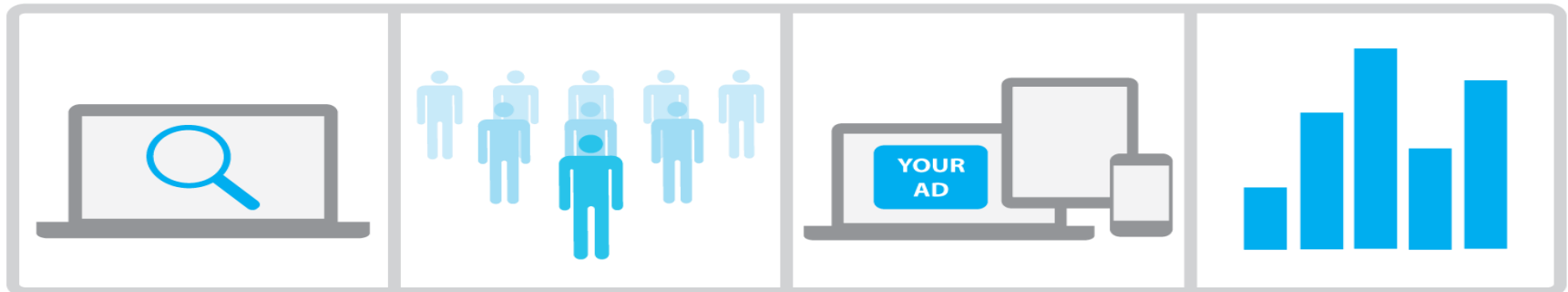
General
Mills

ZenithOptimedia

Allstate.
You're in good hands.

Carat

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



1) Magnetic collects search data

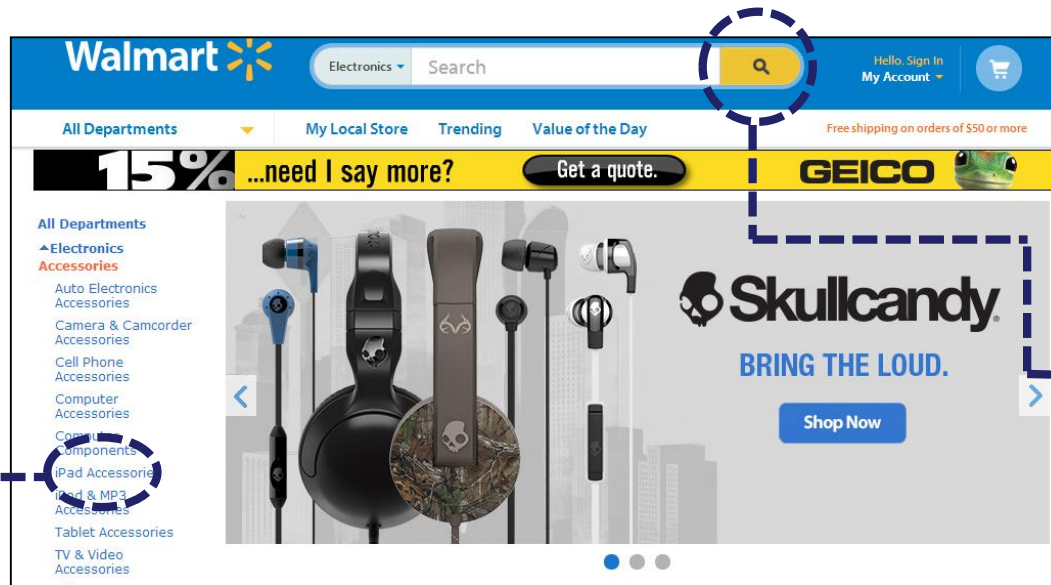
2) Magnetic builds audience segments

3) Magnetic serves retargeted ads

4) Magnetic optimizes campaign

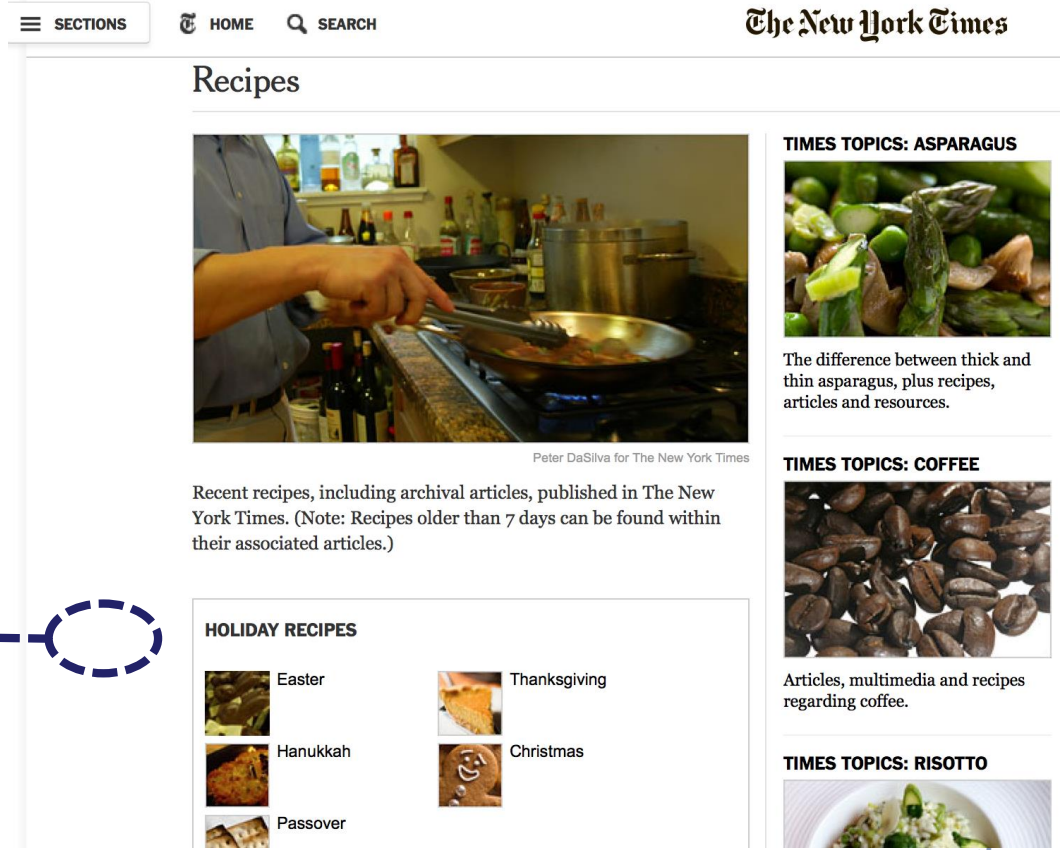
Natural Searches and Navigational Searches

Navigational Search:
"iPad Accessories"



Natural Search:
"iPhone"

Page keywords from article metadata: “Recipes, Cooking, Holiday Recipes”



The screenshot shows the 'Recipes' section of The New York Times website. At the top, there are navigation links for 'SECTIONS', 'HOME', and 'SEARCH'. The main heading is 'Recipes'. Below this, there is a large image of a chef cooking in a kitchen, with a caption 'Peter DaSilva for The New York Times'. To the right of this image is a 'TIMES TOPICS: ASPARAGUS' section with a photo of asparagus and text: 'The difference between thick and thin asparagus, plus recipes, articles and resources.' Below that is a 'TIMES TOPICS: COFFEE' section with a photo of coffee beans and text: 'Articles, multimedia and recipes regarding coffee.' At the bottom left, there is a 'HOLIDAY RECIPES' section with a grid of small images and labels for 'Easter', 'Thanksgiving', 'Hanukkah', 'Christmas', and 'Passover'. At the bottom right, there is a 'TIMES TOPICS: RISOTTO' section with a photo of risotto.



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

DEALBOOK | Microsoft Is Said to Be in Talks to Acquire Maker of Minecraft

DealB%k WITH FOUNDER
ANDREW ROSS SORKIN

MERGERS & ACQUISITIONS | INVESTMENT BANKING | PRIVATE EQUITY | HEDGE FUNDS | I.P.O./OF

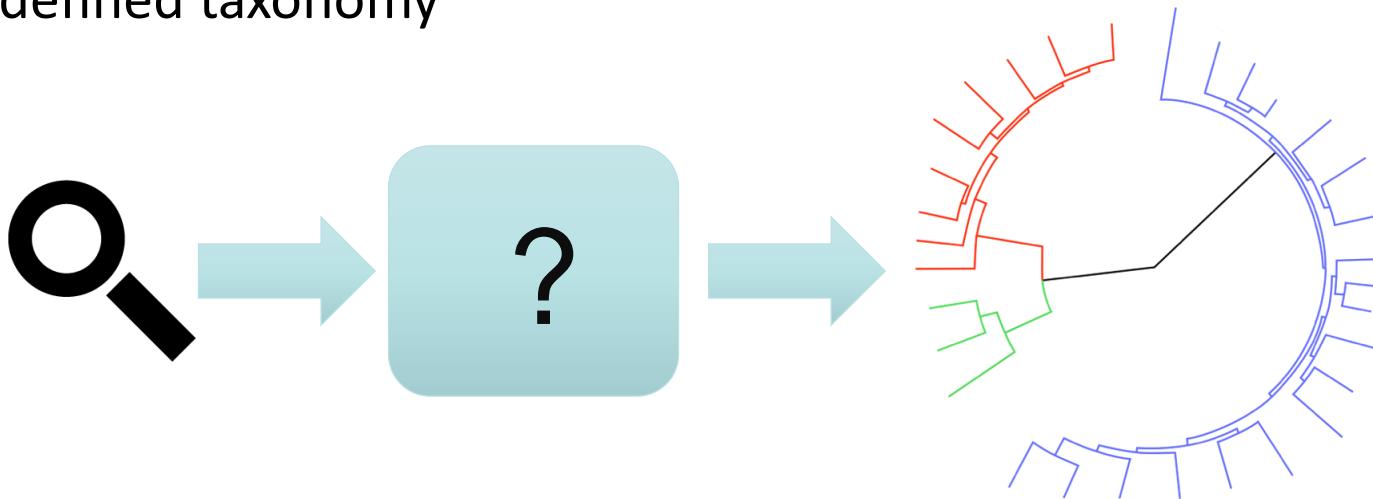
Microsoft Is Said to Be in Talks to Acquire Maker of Minecraft

By NICK WINGFIELD and MICHAEL J. DE LA MERCED SEPTEMBER 9, 2014 8:36 PM [Comment](#)

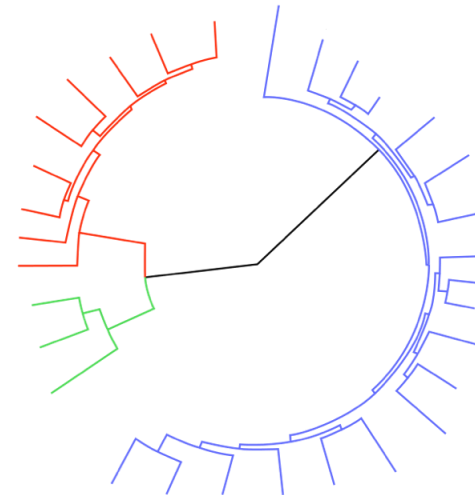
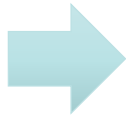


- Targeting categories instead of keywords = Scale
- Use category name to optimize advertising as an additional feature in predictive models
- Reporting by category is easier to grasp as compared to reporting by keyword

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



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



Query Categorization

- 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


fifa 2006

Web Images Maps Shopping Applications More ▾ Search tools

About 95,700,000 results (0.30 seconds)

[2006 FIFA World Cup - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/2006_FIFA_World_Cup ▾
The **2006 FIFA World Cup** was the 18th FIFA World Cup, the quadrennial international football world championship tournament. It was held from 9 June to 9 July ...
2006 FIFA World Cup Final - 2006 FIFA World Cup squads - Knockout stage - Petit

[Images for fifa 2006](#) - Report images



[FIFA 06 - Download](#)
fifa-06.en.softonic.com/ ▾
★★★★★ Rating: 4 - 520 votes - Free - Windows - Game
2006 FIFA World Cup - NBA Live 2001. EA SPORTS FIFA 06 brings the TOTAL FOOTBALL experience to your fingertips. It delivers a combination of attacking ...
Download - Clear filters - Related - See screenshots (7)

[FIFA 2006 - Barcelona vs Real Madrid \(El Clásico \) - YouTube](#)
www.youtube.com/watch?v=QFT2Z1cFeZ4 ▾
Jan 30, 2012 - Uploaded by LightningBS
GAME: FIFA 2006 DIFFICULTY: Professional HALF LENGTH: 6 Minutes STADIUM: Camp Nou PARTICIPANTS ...

[2006 FIFA World Cup - Download](#)
2006-fifa-world-cup-germany.en.softonic.com/ ▾
★★★★★ Rating: 4 - 257 votes - Free - Windows - Game
2006 FIFA World Cup, free download. 2006 FIFA World Cup Demo: Celebrate the passion of the World Cup. EA proves it just can't get enough with 2006 FIFA ...

[2006 FIFA World Cup Germany™ - FIFA.com](#)
www.fifa.com/worldcup/archive/germany2006/index.html ▾
Jun 9, 2006 - A look back at the 2006 FIFA World Cup Germany™

- 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....



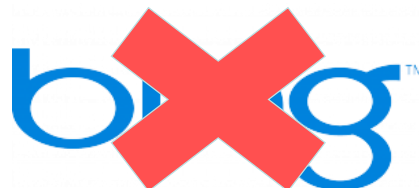
Long Time Ago ...

- ... 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!

250,000 Transactions/month	\$500.00 per month BUY
500,000 Transactions/month	\$1,000.00 per month BUY
1,000,000 Transactions/month	\$2,000.00 per month BUY
1,500,000 Transactions/month	\$3,000.00 per month BUY
2,000,000 Transactions/month	\$4,000.00 per month BUY
2,500,000 Transactions/month	\$5,000.00 per month BUY

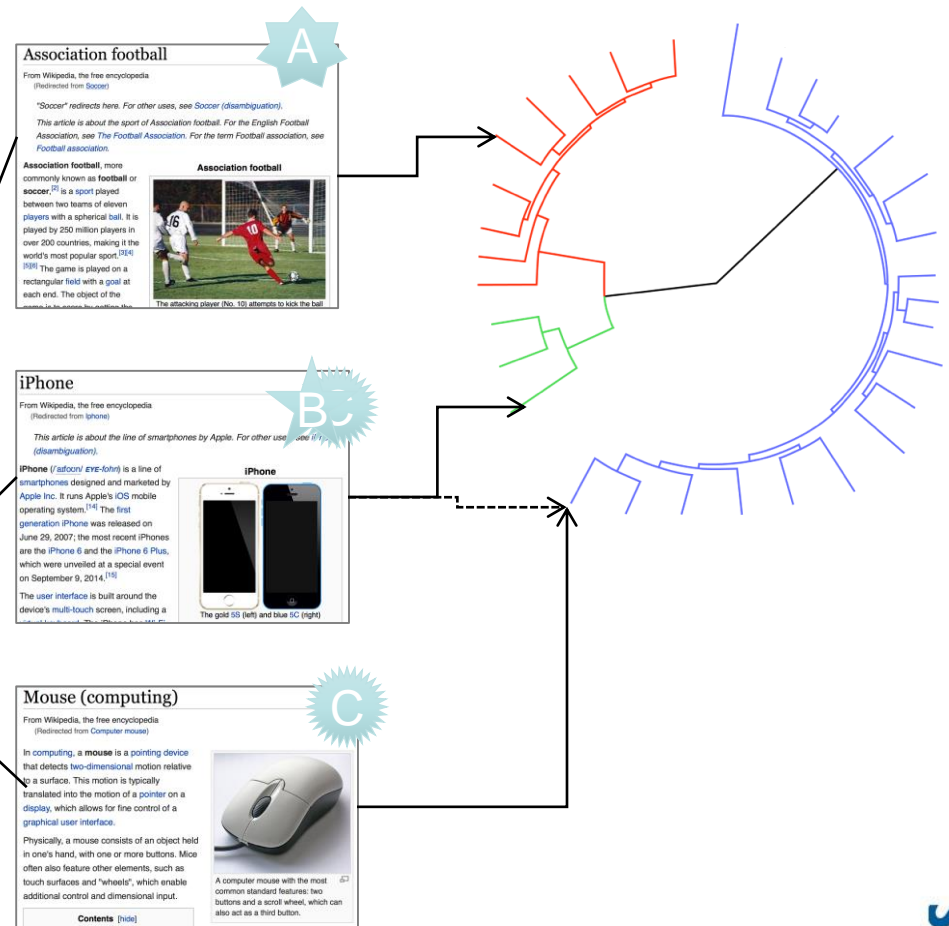
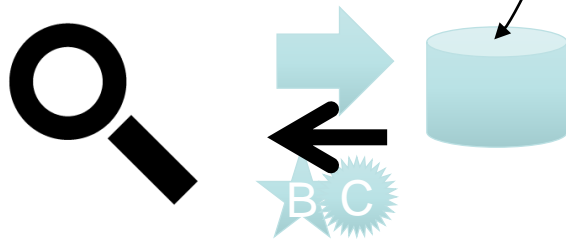


- Use web replacement – Wikipedia
 - 4.5 million articles
 - 9 million of unique titles
 - 40 GB text
- DBPedia
 - Good categories for articles
 - Additional structured data
- Freebase
 - 170 GB triples
 - 40 million topics

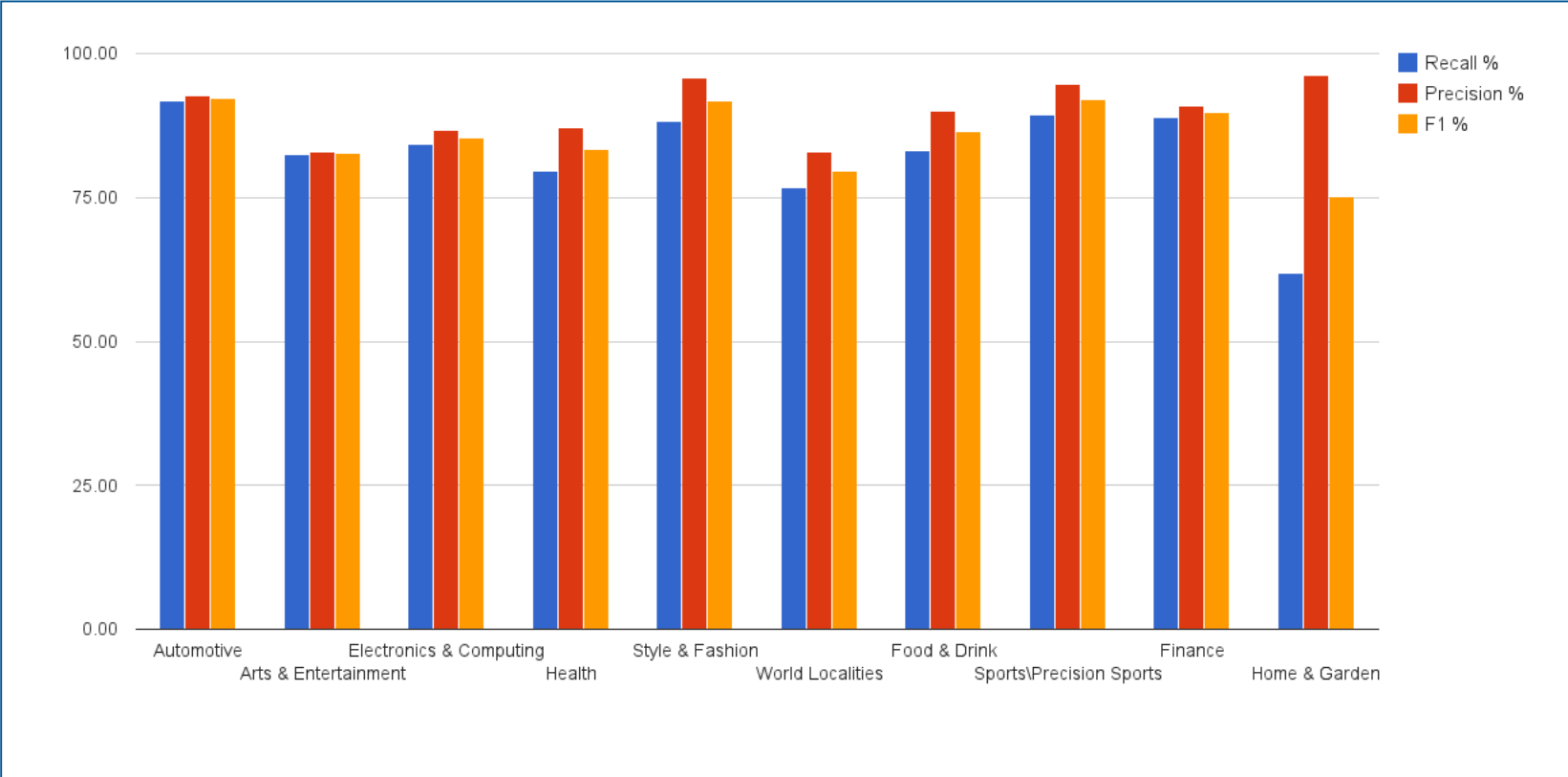


Our Query Categorization Approach

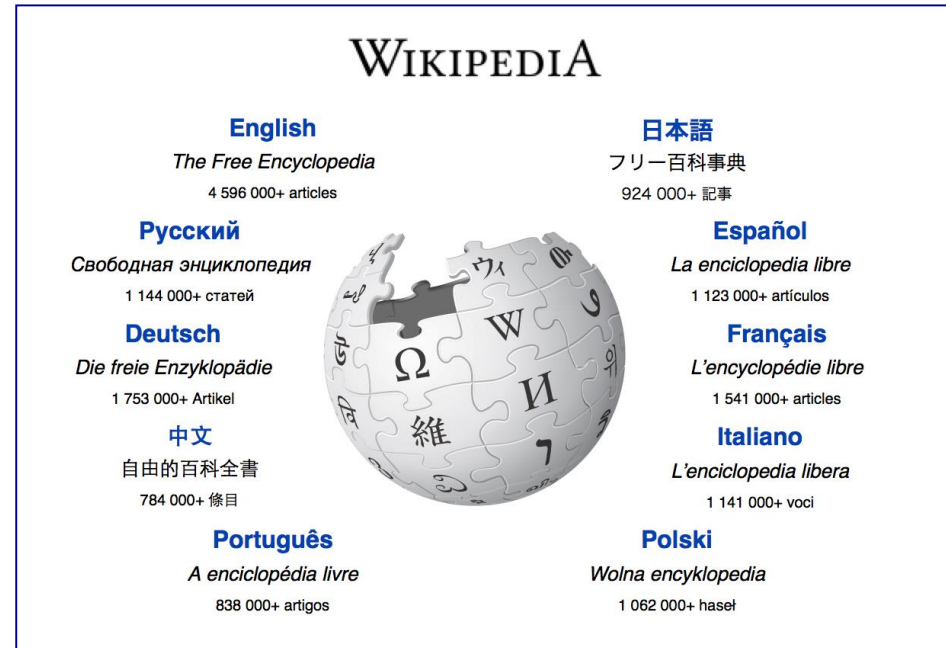
- Assign a category to each Wikipedia document (with a score)
- Load all documents and scores into an index
- Search within the index
- Compute the final score for the query



Results Quality - Precision/Recall

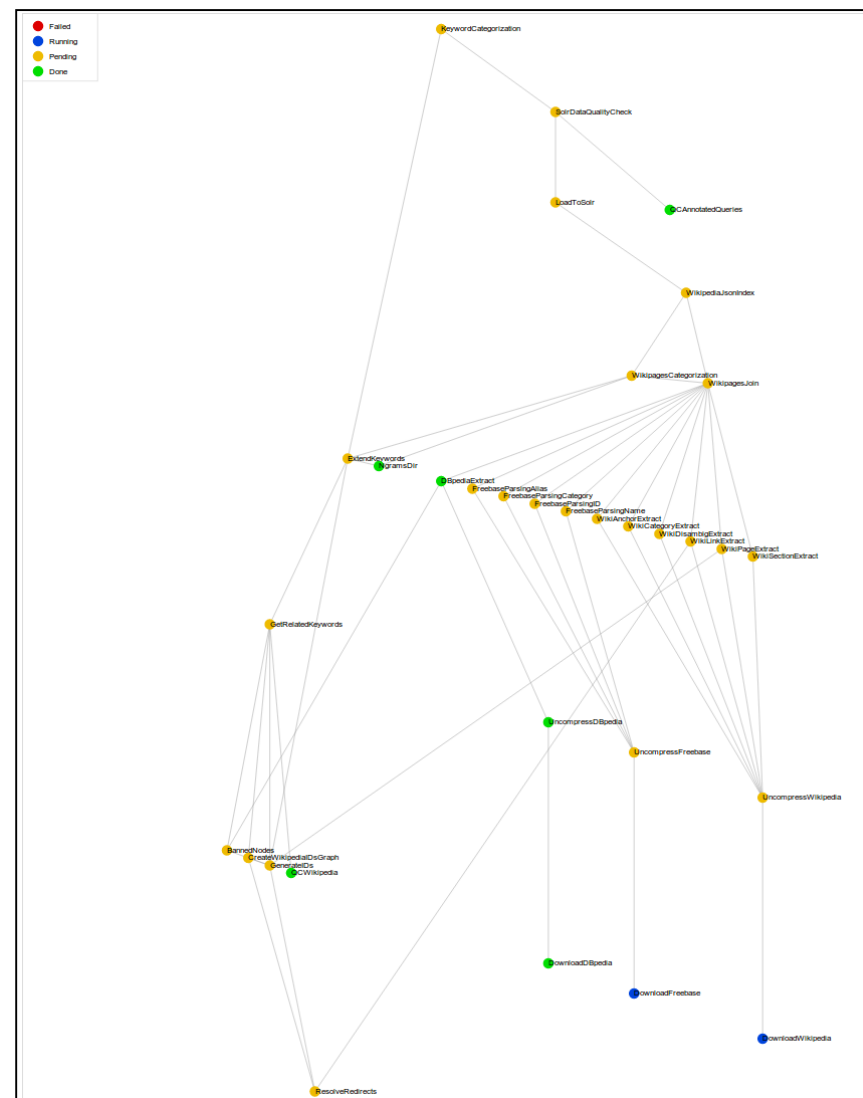


- In development
- Combining indexes for multiple languages into one common index
- Focus:
 - Spanish
 - French
 - German
 - Portuguese
 - Dutch

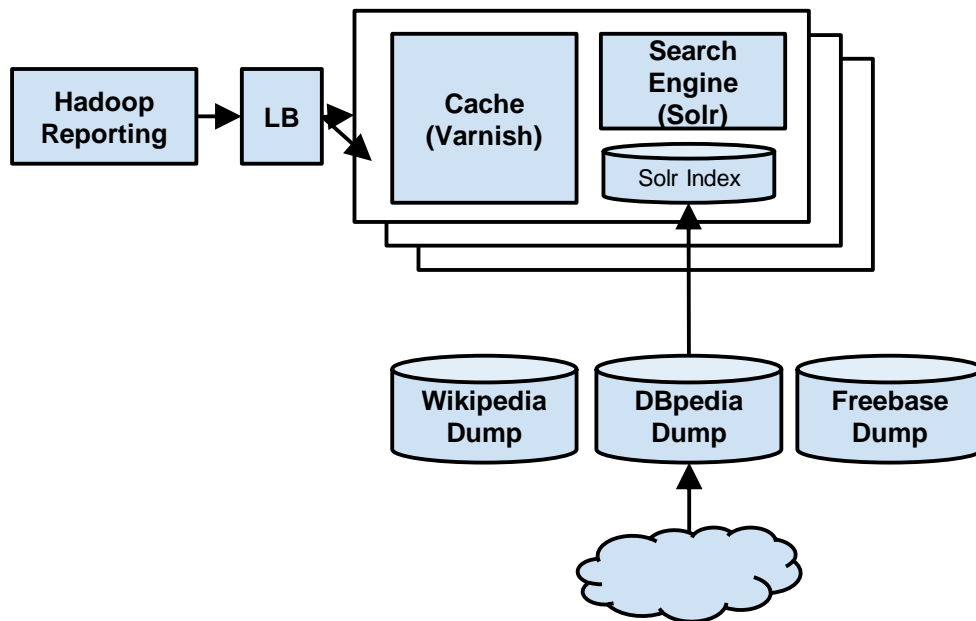


Preprocessing Workflow

- 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



- 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



- 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'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 pozornost

MAGNETIC™

e FOCUS