The Future of Information Discovery: Content Optimization, Social Networks, Interactivity---It's All Very Cloudy

Raghu Ramakrishnan
Chief Scientist, Search & Cloud Computing
Yahoo! Fellow
Main Points

• The medium is being massaged
  – i.e., the canvas is no longer passive
  – It is also very personal, and `always on’
  – *(With apologies to Marshall McLuhan)*

• Search and browse converging
  – Semantics is the new frontier
  – Social networks the new distribution channel

• Learning from Big Data is a key
  – Value of data is now recognized
  – Compute platforms will be increasingly data-centric
The Medium is the Message
McLuhan, 1964: Understanding Media: The Extensions of Man

Wikipedia:

- “People tend to focus on the obvious, which is the content, to provide us information, but in the process, we largely miss the structural changes in our affairs that are introduced subtly, or over long periods of time.”
- McLuhan frequently punned on the word "message" changing it to "mass age", "mess age", and "massage"; a later book, *The Medium is the Massage* was originally to be titled *The Medium is the Message*. 
Yahoo! Data Scale

• **Massive user base and engagement**
  – 650M+ unique users, 11B page visits/month
  – Hundreds of petabytes of storage
  – Hundreds of billions of objects
  – Hundreds of thousands of requests/sec, 200B events/day, 200 PB/day

• **Global**
  – Tens of globally distributed data centers
  – Serving each region at low latencies

• **Challenging Users**
  – Rapidly extracting value from voluminous data
  – Downtime is not an option (outages cost $millions)
  – Variable usage patterns
Content Optimization

Key Features

Package Ranker (CORE)
Ranks packages by expected CTR based on data collected every 5 minutes

Dashboard (CORE)
Provides real-time insights into performance by package, segment, and property

Mix Management (Property)
Ensures editorial voice is maintained and user gets a variety of content

Package rotation (Property)
Tracks which stories a user has seen and rotates them after user has seen them for a certain period of time

Key Performance Indicators

Lifts in quantitative metrics
Editorial Voice Preserved

+79% clicks vs. randomly selected
+250% clicks vs. one size fits all
+43% clicks vs. editor selected
Cloud Structured Storage @ Yahoo!

- Search
- User Login
- Location
- Social Data
- User Generated Content
- Targeted Display Ads
- Stock Portfolio
- Personalized News
- IM Buddies
- Personalized Content
- Reviews and Ratings
- Smart Ads
- Listings
- Saved Locations
By the Numbers

1.5+M read rps

4 Key Platforms

10,000+ servers

10 Geo Zones

102B emails/month

13+B Ad serves/day

11B visits/month

~2B User Ids

~700M Uniq Users*

175+M Users in US

285+M Mail users

40+ Countries

All Data from July/Aug. Worldwide unless indicated to the contrary
* Yahoo!-branded sites
Hadoop: Stability at Scale

Hadoop powers the Yahoo! Network: must be rock-solid

We fix bugs before you see them
• We run very large clusters
• We have a large QA effort
• We run a huge variety of workloads
• Recent spinoff (HortonWorks)
What is PNUTS/Sherpa?

CREATE TABLE Parts (ID VARCHAR, StockNumber INT, Status VARCHAR, ...)

Structured, flexible schema

Parallel database

Hosted, managed infrastructure

Geographic replication
Sherpa Adoption

QoQ 60% growth in projects using Sherpa

In Production

- Abuse Platform
- Answers
- TW Auctions
- COKE
- ConnectedTV
- Groups
- Global Blog Search
- LocDrop
- Mobile Platform
- Mobile Advertising
- Consumer Platform
- Digu
- Search History

- Pipes
- Pulse
- Sports
- Shopping
- UGC Cloud
- YOS Social Directory
- YAP
- YQL
- Aqua (Ad Quality)
- Image Search
- Hosted Search
- And others

No of Customers

Jul '09 Oct '09 Jan '10 Apr '10 Jul '10 Oct '10 Jan '11 Apr '11 Jul '11 Oct '11
10s of data centers, 1000s of machines, PBs of data stored, 100Bs of records, 100Ks of requests/sec

Selective geo-replication, secondary indexes, hashed and ordered tables, flex schemas, eventual and timeline consistency
CORE:
Content Optimization
Recommender Approaches

1. Estimate Most Popular (EMP)
   “What’s most engaging overall?”

2. Behavioral Affinities
   “People who did X, did Y”

3. Attribute Similarities
   “Related items with similar metadata”

4. Social Recommendations
   “What are my trusted connections into?”

5. Business Optimization
   “What generates most business value?”

6. Personalized Recommendations
   “What’s most relevant to me based on my interests, attributes and relationships?”
• What is the best strategy for new articles?
  – If we show it and it’s bad: lose clicks
  – If we delay and it’s good: lose clicks

• Solution: Show it while we don’t have much data if it looks promising
  – Classical multi-armed bandit type problem
  – Our setup is different than the ones studied in the literature; new ML problem

CTR Curves for Two Days

Each curve is the CTR of an item in the Today Module over time

Traffic obtained from a controlled randomized experiment
Things to note:
(a) Short lifetimes, (b) temporal effects, (c) often breaking news stories
Degrees of Personalization

1. Most Popular
Most engaging overall based on objective metrics

2. Most Popular + Per User History
Engaging overall, and aware of what I’ve already seen

3. Light Personalization
More relevant to me based on my age, gender and property usage

4. Deep Personalization
Most relevant to me based on my deep interests and relationships

Voice and Business Rules
Real-time Dashboard
Business Optimization
CORE Modeling Overview

Offline Modeling
- Exploratory data analysis
- Regression, feature selection, collaborative filtering (factorization)
- Seed online models & explore/exploit methods at good initial points
- Reduce the set of candidate items

Online Learning
- Online regression models, time-series models
- Model the temporal dynamics
- Provide fast learning for per-item models

Explore/Exploit
- Multi-armed bandits
- Find the best way of collecting real-time user feedback (for new items)

Large amount of historical data (user event streams)
CORE: Scoring + Modeling Framework

User

ULT (user, article, context, click)

Property
(instrumentation )

Request

CORE: Scoring: onstage

Ranked List

CORE: Ingestion Collection Normalization

LES: Logical Event Stream

Models, Index

Core Modeling Grid: async

Every 5 mins
  • Create/Update item models
  • Build index
  • Also build Core profiles

Core User Profiles

CMS: Content Agility

Article Content

Content Agility

CMS:

LES: Logical Event Stream

Core User Profiles

Sherpa: onstage

BT User Profiles

Ad Behavior Targeting

Ingestion Collection Normalization

Ingestion Collection Normalization
CORE Dashboard: Overall CTR

Compare performance of models and historical benchmarks

Compare buckets and models over time

See which content was promoted most across time

Compare bucket metrics
CORE Dashboard: Segment Heat Map
Examples

- **ACQUISITION**: A “Star Trek” package was #3 with 18-20 demo, #2 with 21-24 demo, but #9 overall. We can acquire younger audiences with targeted content like this.

- **ENGAGEMENT**: “Kobe’s astonishing shot” was #25 with women, but #5 with men. We can better engage men (or sports fans) by showing more like this, women by showing less.

- **REACH**: A package about a hair-pulling soccer player was just plain interesting to everyone (#1-3). We can maintain reach by programming content for the mass audience.
SEARCH TRENDS
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. Amicii ★★★★☆ (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
Julia Roberts - Wikipedia

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...
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.

ANSWERS, NOT LINKS
Web of Concepts

• The web is a source of information about concepts people care about
• The “index” is keyed by concept instance, and organizes all relevant information, wherever it is drawn from, in semantically meaningful ways
  – E.g., Information about Mumbai, or the Dish-Dash restaurant, is organized in terms of meaningful attributes (population, type of cuisine) and can be retrieved and presented to address whatever the user wants to know about Mumbai or Dish-Dash

Ever since the opening of the Suez Canal in the 1860s, the principal gateway to the Indian subcontinent has been MUMBAI (Bombay), the city Aldous Huxley famously described as "the most appalling... of either hemisphere". Travellers tend to regard time spent here as a rite of passage to be survived rather than savoured. But as the powerhouse of Indian business, industry and trade, and the source of its most seductive media images, the Maharashtrian capital can be a compelling place to kill time. Whether or not you find the experience enjoyable, however, will depend largely on how well you handle the... more
Understanding Web Content

with a little help from your friends
Generative Model of the Web

- The true world database
- Site queries
- Information loss
- Noise addition
- Site layout
- Surround generation
- Document Corpus

![Diagram showing the true world database and its interactions with site queries, information loss, noise addition, site layout, surround generation, and document corpus.](image)
Surfacing, Extraction, Integration

End-to-End

Traditional Extraction

Traditional Integration

Surfacing

WWW
DBLife: Community Information Mgmt

- Integrated information about a (focused) real-world community
- Collaboratively built and maintained by the community
  - Semantic web, "bottom-up"
- Joint work with AnHai Doan, Pedro Domingos, Warren Shen and others at Wisconsin
Summary of Gaps

We have observed three key gaps in the state-of-the-art

Gap 1: Finding the pages to extract from may be a significant portion of the supervision cost

Gap 2: Must extract across diverse content types to gather all attributes

Gap 3: Supervision costs for information integration may exceed costs for extraction
Integration: Entity Extraction/Deduping

co-authors = A. Doan, Divesh Srivastava, ...

Raghu Ramakrishnan
Mass Collaboration for IE

Not David!

Picture is removed if enough users vote “no”.
Mass Collaboration Meets Spam

Jeffrey F. Naughton swears that this is David J. DeWitt
Some Closely Related Work

- Avatar (IBM Almaden)
- DeepWeb & WebTables (Google)
- DBLife (U. Wisconsin)
- KnowItAll & TextRunner (U. Washington)
- Nell (CMU)
- Rexa (U. Mass)
Social Networking
Social Networks: Distro to Go!
Connect with your friends around the shows you love

IntoNow from Yahoo! makes engaging with your friends around your favorite television shows easy and fun. Just tap the green button when you’re watching, and IntoNow will identify the show, right down to the episode. Once identified it’s easy to share with your friends on Twitter or Facebook.

IntoNow:

- **Candace T** is watching **SpongeBob SquarePants**
  - Season 1, Episode 21
- **John A** is watching **The Walking Dead**
  - Season 2, Episode 9
- **David S** is watching **Parenthood**
  - Season 3, Episode 17

Discover what your friends are into

Use IntoNow to discover new shows, discuss your favorites with friends, and learn more about whatever you’re into! IntoNow is a great source of recommendations for what’s on right now for you, based on what your friends are into. Find out what shows you share in common, get notified.
Social Advertising

Recommend ads based on private shopping histories of “friends” in the social network.

- Nikon
- HP
- Nike

- Armani
- Gucci
- Prada

(Personalized Social Recommendations: Machanavajjhala et al., VLDB 2011)
Privacy in Social Advertising

Items (products/people) liked by Alice’s friends are better recommendations for Alice.

Fact that “Betty” liked “VistaPrint” is leaked to “Alice”
Privacy in Social Advertising

Alice is recommended ‘X’

Can we provide *accurate* recommendations to Alice while ensuring that Alice *cannot* deduce that Betty likes ‘X’?
Takeaway …

• “For majority of the nodes in the network, recommendations must either be inaccurate or violate differential privacy!”
  
  – Maybe this is a “bad idea”

  – Or, Maybe **differential privacy is too strong a privacy definition to shoot for.**
Related Work

• Amazon, Netflix, Y! Music, etc.:
  – Collaborative filtering with large content pool (see KDD Cup, 2011)
  – Achieve lift by eliminating bad articles
  – We have a small number of high quality articles

• Search, Advertising
  – Matching problem with large content pool
  – Match through feature based models
Mobius: Unified Messaging and Data for Mobile Apps

Byung-Gon Chun, Carlo Curino, Russell Sears, Alexander Shraer, Samuel Madden*, Raghu Ramakrishnan

Yahoo! Research, *MIT

mobicloud@yahoo-inc.com
Mobile Apps Connect to Cloud

Access, Create, Share Data/Messages

84% of apps are “cloud-connected”

(source Appcelerator)
Mobile App Developers’ Challenges

1. Scalable Backend
Mobile App Developers’ Challenges

2. Speed: wireless networks
   - Caching
   - Prefetching
Mobile App Developers’ Challenges

3. Notification

Push Alerts
Mobile App Developers’ Challenges

4. Messaging

Share News

Share News

Share News
Mobile App Developers’ Challenges

5. Disconnection

- Read news
- Post on Facebook
Mobius

- MUD API: (continuous) read/write of data
- Protocols to handle disconnection (and svn-like conflict resolution)
- Predicate-based notification (receiver and sender predicates)
- Automatic caching/prefetching (global policy generated by the cloud)
Mobius Architecture

Device

- App 1
  - MUD lib
  - MUD service
  - Notification client
  - Data client
  - Database

- App 2
  - MUD lib
  - MUD service
  - Notification client
  - Data client

Cloud

- Notification server
- Caching policy generator
- MUD gateway/processor
- Storage system

Connections:
- Notifications from Device to Cloud
- Reads/writes between Device and Cloud
Main Points

• The medium is being massaged
  – i.e., the canvas is no longer passive
  – It is also very personal, and `always on’
  – *(With apologies to Marshall McLuhan)*

• Search and browse converging
  – Semantics is the new frontier
  – Social networks the new distribution channel

• Learning from Big Data is a key
  – Value of data is now recognized
  – Data marketplaces will emerge
  – Compute platforms will be increasingly data-centric