KNOW THY USERS: COMPILING AND SERVING THE CONTENT PREFERENCES OF OUTBRAIN’S USER BASE

Ronny Lempel
VP Recommendations, Outbrain
What Is Outbrain?
The Lighthouse

Help people discover content they can trust to be interesting, relevant and timely for them.
Over
25 BILLION PAGE VIEWS PER MONTH
Over 200 Billion Recommendations Served Per Month
AGENDA

• WHAT IS OUTBRAIN
• PERSONALIZATION AND OUTBRAIN’S USER PROFILE
• IMPLEMENTATION
• ADVANCED IMPLEMENTATION
• SOME NUMBERS
PERSONALIZATION
TWO TRACKS OF PERSONALIZATION

• Content-based
  Understand the affinity of each user to each piece of content, based on that user’s historical content consumption

• Collaborative filtering
  Understand the affinity of each user to each piece of content, based on co-consumption patterns of the entire user population

Outbrain’s platform is a hybrid that taps both sources
CONTENT BASED RECOMMENDATIONS

Standard three step recipe:

1. **Understand** what content is about
   - Represent content in a rich feature space, consisting of site sections, categories, topics, entities

2. **Aggregate** the features of the content consumed by each user to model that user
   - Represent the user in a similar feature space

3. **Match** content to a user

WSDM'2016
CONTENT CONSUMPTION EVENTS TURNED INTO PROFILES

Page Views, Clicks, Shares, Video play/pause, ...
USER PROFILE: FAVORITE CATEGORIES

YOUR CATEGORIES

<table>
<thead>
<tr>
<th>Category</th>
<th>Computed</th>
<th>Server Side</th>
<th>Cold Start</th>
<th>Cookie</th>
</tr>
</thead>
<tbody>
<tr>
<td>Football</td>
<td>0.19639465</td>
<td>0.19639464</td>
<td>0.029425267</td>
<td>0.125</td>
</tr>
<tr>
<td>Basketball</td>
<td>0.10662436</td>
<td>0.10662436</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Software</td>
<td>0.061638538</td>
<td>0.061638538</td>
<td>--</td>
<td>0.125</td>
</tr>
<tr>
<td>Real Estate</td>
<td>0.060585096</td>
<td>0.060585093</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>
USER PROFILE: FAVORITE SITES

YOUR SOURCES

<table>
<thead>
<tr>
<th>Source</th>
<th>Computed</th>
<th>Server Side</th>
<th>Cold Start</th>
<th>Cookie</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://mmqb.si.com/">http://mmqb.si.com/</a></td>
<td>0.11572199</td>
<td>0.115721986</td>
<td>--</td>
<td>0.083333336</td>
</tr>
<tr>
<td><a href="http://www.si.com/river">http://www.si.com/river</a></td>
<td>0.090729184</td>
<td>0.09072918</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td><a href="http://www.ynet.co.il/home/0,7340,L-3,00.html">http://www.ynet.co.il/home/0,7340,L-3,00.html</a></td>
<td>0.07776445</td>
<td>0.07776445</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td><a href="http://www.ynet.co.il/home/0,7340,L-2,00.html">http://www.ynet.co.il/home/0,7340,L-2,00.html</a></td>
<td>0.07048833</td>
<td>0.07048833</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td><a href="http://recode.net/category/media/">http://recode.net/category/media/</a></td>
<td>0.06857925</td>
<td>0.06857924</td>
<td>--</td>
<td>0.16666667</td>
</tr>
<tr>
<td><a href="http://www.nbcphiladelphia.com/">http://www.nbcphiladelphia.com/</a></td>
<td>0.058901973</td>
<td>0.058901973</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td><a href="http://www.tricities.com/news/local/">http://www.tricities.com/news/local/</a></td>
<td>0.056125287</td>
<td>0.056125283</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>
## User Profile: Site-Category Combinations

### Your Source-Category Pairs

<table>
<thead>
<tr>
<th>Source and Category</th>
<th>Computed</th>
<th>Server Side</th>
<th>Cold Start</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://mmqb.si.com/">http://mmqb.si.com/</a>, Sports/Football</td>
<td>0.30465043</td>
<td>0.30465043</td>
<td>0.004823738</td>
</tr>
<tr>
<td><a href="http://www.nbcbphiladelphia.com/">http://www.nbcbphiladelphia.com/</a>, Technology_and_Internet/Software</td>
<td>0.07679236</td>
<td>0.07679236</td>
<td>--</td>
</tr>
<tr>
<td><a href="http://www.tricities.com/news/local/">http://www.tricities.com/news/local/</a>, Health/Mental_Health</td>
<td>0.049294382</td>
<td>0.049294382</td>
<td>--</td>
</tr>
<tr>
<td><a href="http://finance.townhall.com/">http://finance.townhall.com/</a>, Business_and_Finance/Real_Estate</td>
<td>0.04844852</td>
<td>0.04844852</td>
<td>--</td>
</tr>
<tr>
<td><a href="http://www.si.com/nfl">http://www.si.com/nfl</a>, Sports/Football</td>
<td>0.047520883</td>
<td>0.047520883</td>
<td>0.006322426</td>
</tr>
<tr>
<td><a href="http://www.si.com/mlb">http://www.si.com/mlb</a>, Sports/Baseball</td>
<td>0.042384487</td>
<td>0.042384487</td>
<td>--</td>
</tr>
<tr>
<td><a href="http://recode.net/category/media/">http://recode.net/category/media/</a>, Business_and_Finance/Media</td>
<td>0.039736275</td>
<td>0.039736275</td>
<td>--</td>
</tr>
</tbody>
</table>
## User Profile: Favorite Topics

### Your Topic Models

<table>
<thead>
<tr>
<th>Topic Model</th>
<th>Computed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cinema</td>
<td>0.5205062</td>
</tr>
<tr>
<td>Comics</td>
<td>0.35867336</td>
</tr>
<tr>
<td>Literature</td>
<td>0.23290905</td>
</tr>
<tr>
<td>Career</td>
<td>0.17538102</td>
</tr>
<tr>
<td>Lifestyle</td>
<td>0.14753719</td>
</tr>
<tr>
<td>Names</td>
<td>0.14105599</td>
</tr>
<tr>
<td>Television</td>
<td>0.13184603</td>
</tr>
<tr>
<td>Communication</td>
<td>0.06536578</td>
</tr>
<tr>
<td>Interpersonal relationships</td>
<td>0.05690778</td>
</tr>
<tr>
<td>Lifehacks</td>
<td>0.05634296</td>
</tr>
<tr>
<td>Celebrity Gossip</td>
<td>0.054192513</td>
</tr>
<tr>
<td>Content Marketing</td>
<td>0.043678593</td>
</tr>
<tr>
<td>History</td>
<td>0.03707295</td>
</tr>
</tbody>
</table>
ILLUSTRATIVE USER TOPICS REPRESENTATION

MUSIC

HACKER

ATTACK
IMPLEMENTATION (ONLINE AND INCREMENTAL)
Infrastructure Building Blocks

• Apache Cassandra (née Facebook)
  – Distributed NoSQL database
  – Horizontally scalable through sharding and replication
  – Multi-DC aware via cross-DC replication

• Apache Kafka (née LinkedIn)
  – Scalable publish-subscribe messaging system
  – Messages are consistently partitioned by key
  – Consumers reading a partition consume messages in FIFO order
UPDATING THE PROFILE — CONCEPTUAL FLOW

Click!

User Interaction

DocID
userID
Action

Categories:
• Nature
• Leisure

Topics:
• Hiking

Entities:
• Yosemite

UserID’s Profile

Categories:
• Football
• Investing
• Leisure

Topics:
• Hiking
• Startups

Entities:
• Yosemite

aggregate and decay

enrich

Document Service
Updating the Profile — Architectural Flow

User interaction

Outbrain Gateway

WhoAmI (user service)

Write User Event

Event

Raw Events

Cassandra

Document Service

Get Doc Features

kafka

UPAG

Write Top-K

Aggregate & Decay

Online

Cassandra

Offline

Cassandra

WSDM'2016
• In addition to user interests, Outbrain’s user profile also saves some raw events (non-aggregated data)
• Example: several hundred recent pages views per user
• Motivation: do not recommend content the user has already consumed by filtering candidates against this list
  – Also filter some recently or frequently served recommendations
SERVING FLOW AND UPDATE FLOW

- The offline repository stores the full user profile
- The online repository stores only the data required for serving: top-K values of user preference
- Motivation: minimize data to be read at serving, reduce latency
ADVANCED
KAFKA PARTITIONING AND CACHING

- Updating a user’s profile requires reading it
- Users consume content in bursts – locality of reference
- By partitioning Kafka messages by userID, all interactions of a user are routed to the same consumer, allowing the consumer to cache user profiles
**MULTI-DATACENTER SUPPORT: PROBLEM**

- Within a single DC, the usage of Kafka ensures that per user, interactions are processed sequentially in FIFO order by UPAG.
- Outbrain operates out of several data centers, and every user’s interactions may be arrive to each one.
- Cassandra has cross-DC replication capabilities: a key updated here can (soon) be read there.
- Processing interactions independently in each DC might cause *race conditions* between Cassandra reads and writes across DCs.
MULTI-DATACENTER SUPPORT: SOLUTION

- Maintain offline and online profiles per user per DC
- Update flow: each DC updates its own set of profiles with the interactions it processes
  - Those get replicated to other DCs, where they are read-only
- Serving flow: in each DC, WhoAmI reads, aggregates and top-k’s all online profiles of the user
- Potential loss of precision in final top-k
LEARNING TO AGGREGATE

• Users consume sequences of documents
• Documents are annotated with semantic metadata (categories, topics, entities)
• How should we transform the activity steam into a profile representing users’ current content affinities and interests?
• Profile should be incrementally maintained rather than recomputed from all raw events over and over again
LEARNING TO AGGREGATE — PROPOSED METHODOLOGY

• Difficulties:
  – Explicit consumption data; rejection is implicit
  – Consumption heavily biased by publisher’s editorial staffs
  – Popularity bias is always tricky to overcome in recommender systems

• Methodology:
  – Build user profiles incrementally from training stream of raw events
  – Test procedure: match two users to two stories, based on profiles and story semantics (test cases chosen to have exactly one correct result)

WSDM’2016
In addition to user interests, Outbrain’s user profile also saves some raw events (non-aggregated data).

Example: several hundred recent pages views per user.

Motivation: do not recommend content the user has already consumed by filtering candidates against this list.

Fact: Outbrain’s crawler assigns document ids sequentially.
  - So recent documents have close-by document identifiers.
RAW EVENTS AND BIDIRECTIONAL DGAPS

- Compression opportunity: since users most often consume current content, most recent page-views have close-by docids.
- Solution: d-gap encode the docids of recent page views, as done when compressing postings lists.
- Twist: gaps can be negative.
- Can’t sort the gaps, since they are naturally sorted by recency.
- Solution: add a sign bit, or encode XOR-gaps.

Timestamp | Doc id | d-gap
--- | --- | ---
T-48 | 2096 | 2096
T-26 | 4200 | 2104
T-26 | 4155 | -45
T-25 | 4300 | 145
T-2 | 6000 | 1700
T-1 | 2250 | -3750
SOME NUMBERS

• 3 Data centers
• Accommodates 35,000 recommendation requests per second, served on average at 50 ms
• 30 WhoAmI machines
• 8 UPAG machines
• 2 Cassandra clusters of 80 + 20 machines with 30 TB of compressed data
• 300,000/sec read operations from Cassandra at < 10 ms per operation
Thank you
rlempel at outbrain dot com

We’re hiring!!
www.outbrain.com/about/careers