Competition

CIKM Cup 2016 Track 2: Personalized E-Commerce Search Challenge

Organized by spirinus - Current server time: Sept. 9, 2016, 7:05 a.m. UTC

▶ **Current**	Next
Phase 1: Validation Leaderboard	**Phase 2: Test Leaderboard**
Aug. 5, 2016, midnight UTC	Oct. 2, 2016, midnight UTC

Learn the Details Phases Participate Results

Forums (/forums/7901/)

Get Data

Submit / View Results

Data Files

The dataset includes user sessions extracted from an e-commerce search engine logs, with anonymized user ids, hashed queries, hashed query terms, hashed product descriptions and meta-data, log-scaled prices, clicks, and purchases. There are eight different files described below. The files can be downloaded **here** (https://drive.google.com/open?id=0B7XZSACQf0KdXzZFS21DblRxQ3c).

train-queries.csv and test-queries.csv (~869.5MB)

- queryId (serial)
- sessionId (serial)
- userId (serial)
- tmeframe (time since the first query in a session, in milliseconds)
- duration (page dwell time, in milliseconds)
- eventdate (calendar date)
- searchstring.tokens (comma separated hashed query tokens; empty if it is a query-less case)
- categoryId (product category ID; empty if it is a query-full session)
- **items (productIDs returned by the default ranking algorithm on the SERP; this IDs must be re-ranked).**
- is.test (TRUE/FALSE; TRUE if it is a test query)

About (https://github.com/codalab/codalab/wiki/Project_About_CodaLab) Forum (https://groups.google.com/forum/#!forum/codalabdev)

- regionId (geographical region of a query; serial).

An example Query object looks as follows:*1;;16327074;311;2016-05-09;16655,244087,51531,529597,58153;0;62220,33969,30311,32902,8252,13682,*

products.csv (~7.3MB)

- productID (serial)
- priceLog2 (log-transformed product price)
- product.name.tokens (comma separated hashed product name tokens)
- imageName (name of the corresponding product image)

An example Product object looks as follows: *1;1;10;4875,776,56689,18212,18212,4896*

product_images.zip (might be released after 50% of the competition time, currently under consideration, ~14GB)

To find an image for a product, one should use *imageName* attribute from the **products.csv**.

product-categories.csv (~2MB)

- productCategoryID (serial)
- productID
- categoryID

An example ProductCategory object looks as follows: *1;139578;1096*

train-purchases.csv (749KB)

- sessionId (serial)
- timeframe (time since the first query in a session, in milliseconds)
- eventdate (calendar date)
- ordernumber (serial product orderID; groups all products purchased together ~ shopping cart; if a user bought several products, there are several records sharing the same ordernumber)
- itemId (purchased product)

An example Purcahse object looks as follows: *100030;1861906;2016-04-20;2963942;377191*

train-item-views.csv (42.7MB)

- sessionId
- userId
- itemId
- timeframe (time since the first query in a session, in milliseconds)
- eventdate (calendar date)

An example ItemView object looks as follows: *1;;81766;526309;2016-05-09*

Data Pre-processing

To allay privacy concerns the user data is fully anonymized. Only meaningless numeric IDs of users, queries, query terms, sessions, URLs and their domains are released. The queries are grouped by sessions. Specifically, we applied the following pre-processing before the release of the dataset for the Personalized E-commerce Search Challenge:

1. Take the most recent six months of logs of an e-commerce search engine.
2. Remove queries without clicks.
3. Detect sessions using a 1-hour of inactivity heuristic (for web search the session segmentation heuristic is ~20 min).
4. Find the first query in each session and replace timestamps for all events in the session relative to the first query, i.e. the timestamp for the first query in the session is 0 and all other events have non-negative delta-timestamps.
5. To hash textual data, we: (1) build the vocabulary by concatenating all available textual data such as queries, product titles, product descriptions; (2) for each unique word assign a hash-code using an MD5-based hash function; (3) replace each word with the corresponding hash-code.
6. Using the same transformation as in step 4, we hash the names for product images.
7. Prices are subject to log transformation and subsequent rounding to the nearest smallest integer, i.e. if a product costs 3.89, then the obfuscated price will be 1; if a product costs 4.89, then the obfuscated price will be 2.
8. For training, we take all sessions before a certain timestamp.
9. For testing, we take the last session for each user, find the first query in this session, and hide all actions after the query action (when a SERP with the results is presented). The goal is to re-rank the products on that SERP.

Dataset Statistics

- The number of sessions: 573,935
- The number of products: 134,319,529
- The number of products viewed from search (including browsing after SERP): 2,451,565
- The average number of products viewed per search session (including browsing after SERP): 4.271
- The number of SERP clicks on products: 1,877,542
- The average number of SERP click per search session: 3.271
- The number of products purchased from search: 68,818
- The average number of products purchased from search session: 0.119

We use 50% of test queries for the Validation Leaderboard (phase 1) and 50% for the Test Leaderboard (phase 2). We don't disclose which test queries are used for the public leaderboard and which test queries are used for the private leaderboard. Every submission must contain re-ranked productIDs for all test queries/SERPs.