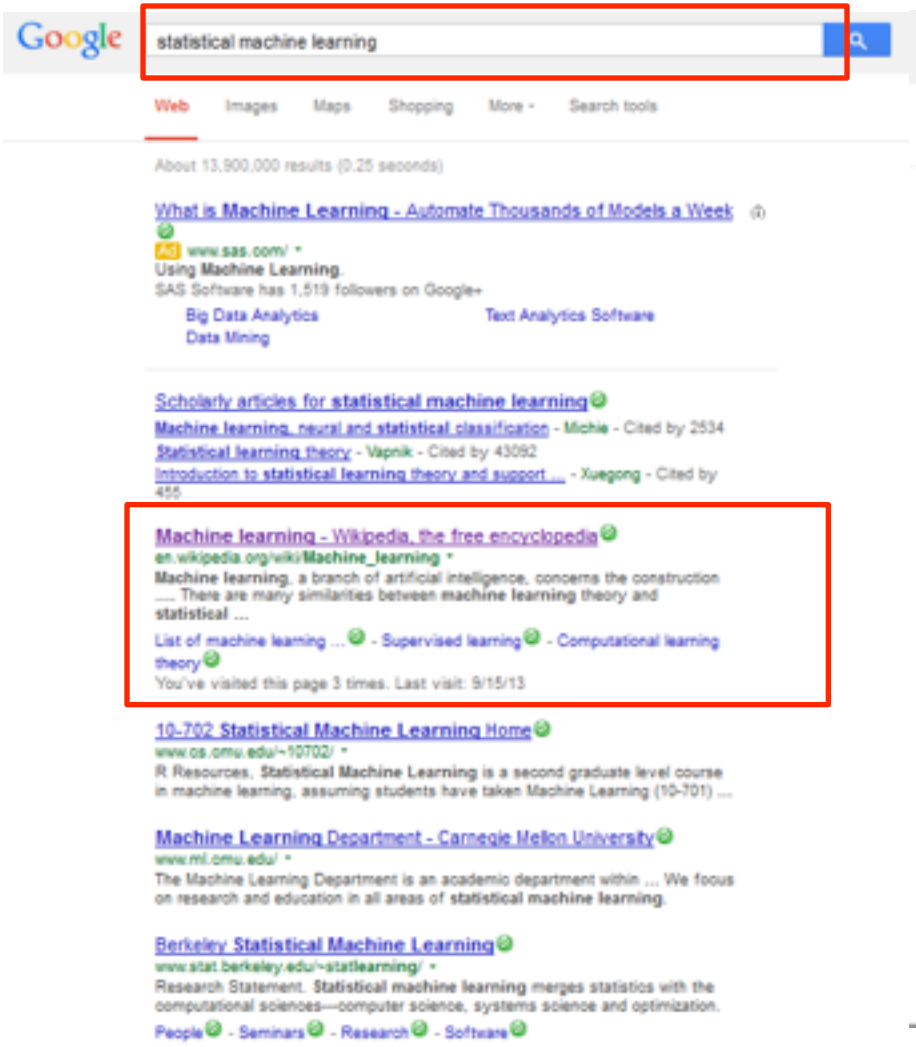
Personalized Web Search

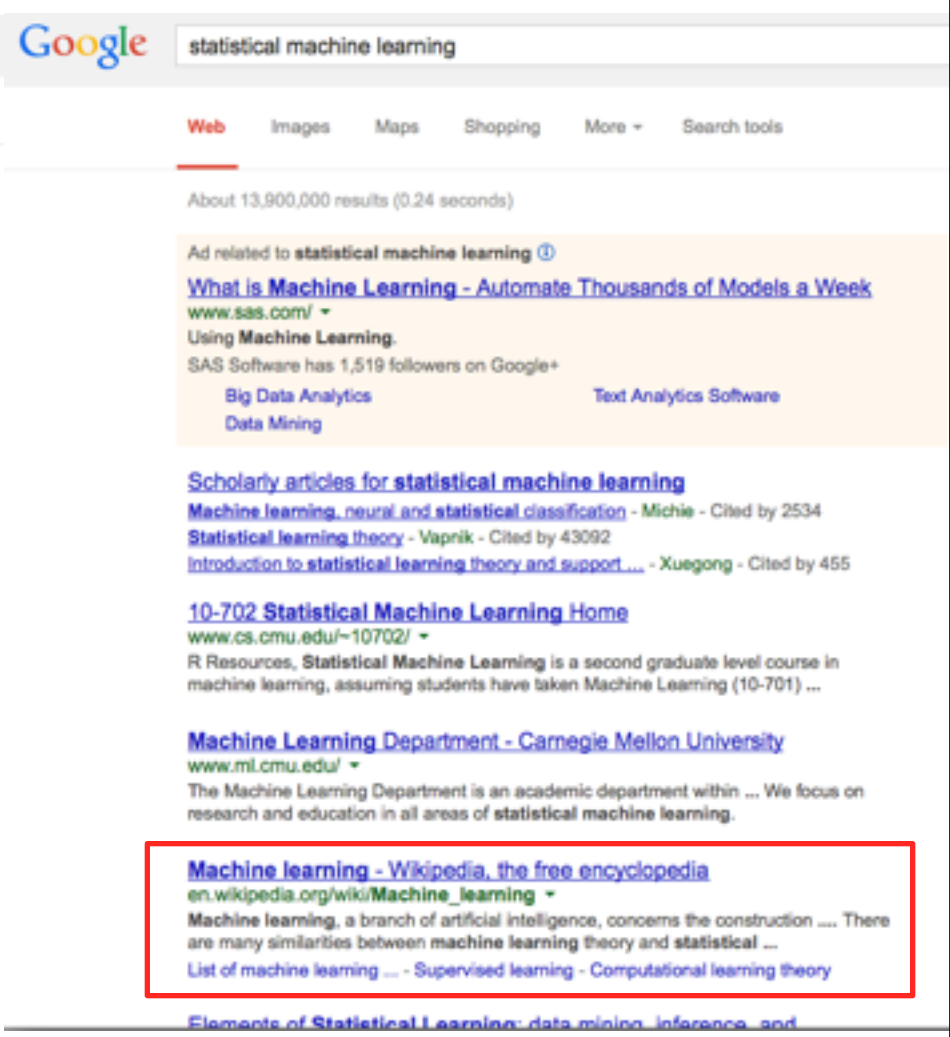
Anantha Raghuraman
Asish Ghoshal

Overview

- Introduction
- Problem Statement and Data format
- Evaluation Criteria
- Algorithm
- Results
- Conclusion

What is personalized web-search?





Problem Statement

- Re-rank the 10 URLs of each *Search Engine Result Page* according to the personal preferences of the users.
- The training period corresponds to 27 days of search activity. The next 3 days correspond to the test period.
- The problem statement was taken from kaggle.com.

Raw Data

```
34573630 M 28 15
34573630 0 Q 0 10507991 3139706,2771252,3808573 34169548,3278460 34165793,3278348
35438447,3339074 15367590,1582976 31337693,3075260 43622876,3822427
26061675,2596986 29897513,2901859 39010230,3548763 62850010,4824984
34573630 6 C 0 34169548
34573630 250 T 1 2338342 1255686,3591321,1687414,3416146,4342041 56906042,4503913
21293423,2183949 3580938,482441 21291242,2183806 14221334,1461559 43622870,3822427
58185226,4577130 6936569,855329 5736329,747654 52480003,4295034
```

- Extracted from Yandex logs.
- Data is hashed for anonymity.

Data Characteristics

- Noteworthy characteristics of the dataset:
- Unique queries: 21,073,569
- Unique urls: 703,484,26
- Unique users: 5,736,333
- Training sessions: 34,573,630
- Test sessions: 797,867
- Clicks in the training data: 64,693,054
- Total records in the log: 167,413,039
- Training data size : **45 GB**!

Sample Training Data (post-extraction)

User-ID	Session-ID	SERP-ID	QUERY ID	QUERY-TIME	QUERY TERMS	URLS	DOMAINS	CLICK_TIME	Dwell time
58794	8975	98579	23597	500	9804, 67805, 68743	76940 9865 789976 59 785575 498598 589598 0 98454 78504	5680 78948 69 598349 855 3974 5985 540127 984 9347	1500 700 0 0 0 0 900 0 0 0	420 38 0 0 0 0 502 0 0 0

Evaluation Criteria I
Relevance

Rank results according to relevance

- **2 (Highly relevant):** dwell time > 400 time units or last click in a session
- **1 (Relevant):** dwell time between [50 , 400) time units
- **0 (Irrelevant)**: dwell time < 50 time units

Evaluation Criteria II
Normalized Discounted Cumulative Gain (NDCG)

- Popular measure for evaluating web search and related tasks
- Two assumptions:
  - Highly relevant documents are more useful than marginally relevant document
  - the lower the ranked position of a relevant document, the less useful it is for the user, since it is less likely to be examined

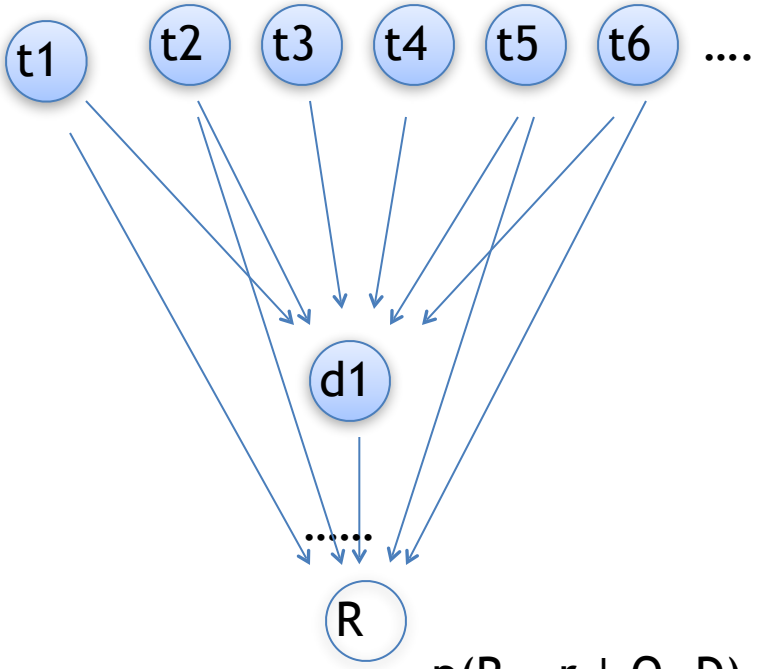
NDCG - Example

	Ground Truth		Ranking Function$_1$		Ranking Function$_2$	
	Document Order	r_i	Document Order	r_i	Document Order	r_i
1	d4	2	d3	2	d3	1
2	d3	1	d4	1	d2	1
3	d2	1	d2	1	d4	2
4	d1	0	d1	0	d1	0
	$NDCG_{GT}$=1.00		$NDCG_{RF1}$=1.00		$NDCG_{RF2}$=0.9203	

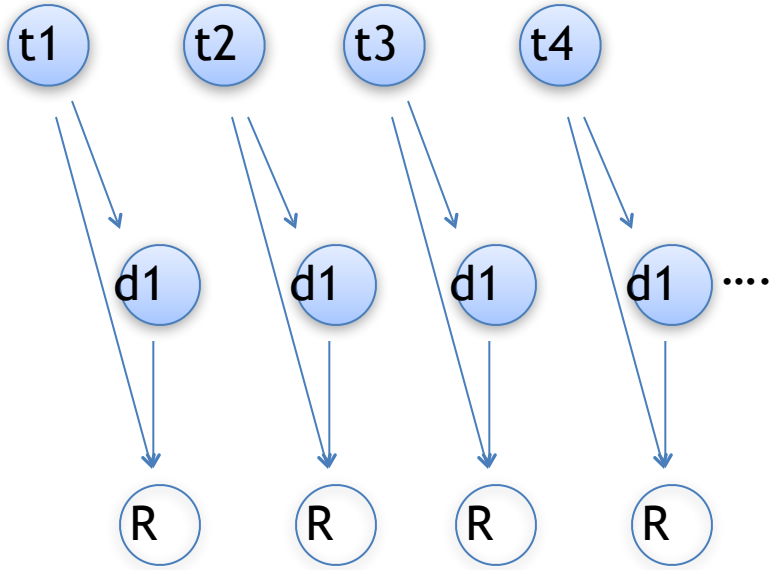
DCG (actual) = 2 + 1/log2 + 1/ log3 + 0/ log4 = 3.63 » NDCG = 1

DCG (obtained) = 1 + 1/log2 + 2/ log3 + 0/ log4 = 3.26 » NDCG = 0.90

Generative model for relevance



max



$$p(R = r \mid Q, D) = \max_{T \in Q} p(R = r \mid D, T)$$

$$p(R = r \mid T, D) = p(R = r, T, D) / p(D, T)$$

$$R = \{0, 1, 2\}$$

Results

- Our Score: 0.76359
- Random baseline: 0.47972
- Top score: 0.80268

Future Work

- Use query click entropy to avoid personalizing queries with low click entropy.
- Exploit pattern in the short-term session.