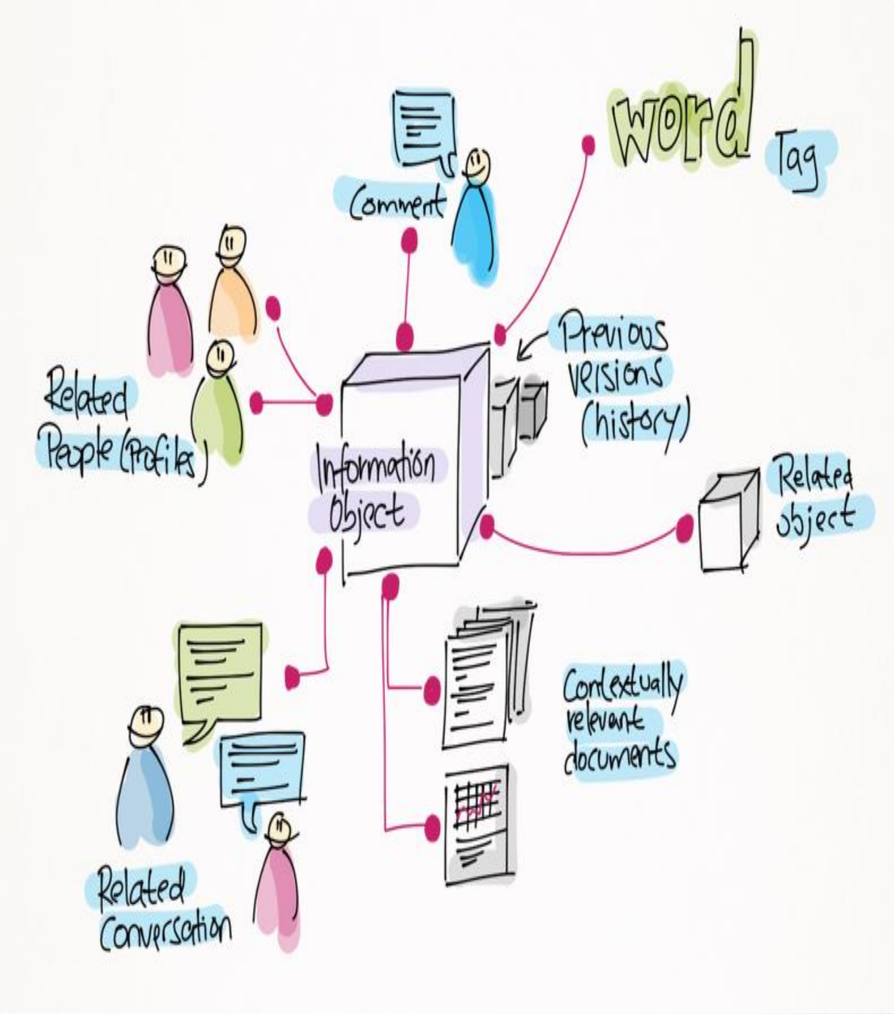
Re-ranking Web Documents Based on Personal Preferences

Satarupa Guha, Priyanshu Agrawal, Shubham Sangal
IIIT Hyderabad

Personalisation… Why ??

- Each Individual is unique.
- Search Engines should re-rank its results based on his profile so that his specific requirements can be met

- Personalisation works using **Individualisation and Contextualisation**
- Individualisation means creating each User's Individual Profile using his **Long-Term History**
- Contextualisation means Identifying relation between sessions using **Short-Term History**



Dataset… Fully Anonymized

- Dataset was provided by Yandex (Almost 16 GB of train data and 400 MB of test Data)
- Search Data for 30 days from a large city was collected.
- 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.
- Training was done on data of first 27 days and testing on last 3 days.

Data Format

The log represents a stream of user actions, with each line representing a session metadata, a query action, or a click action. Each line contains tab-separated values according to the following format:

Session metadata (TypeOfRecord = M):

SessionID TypeOfRecord Day USERID

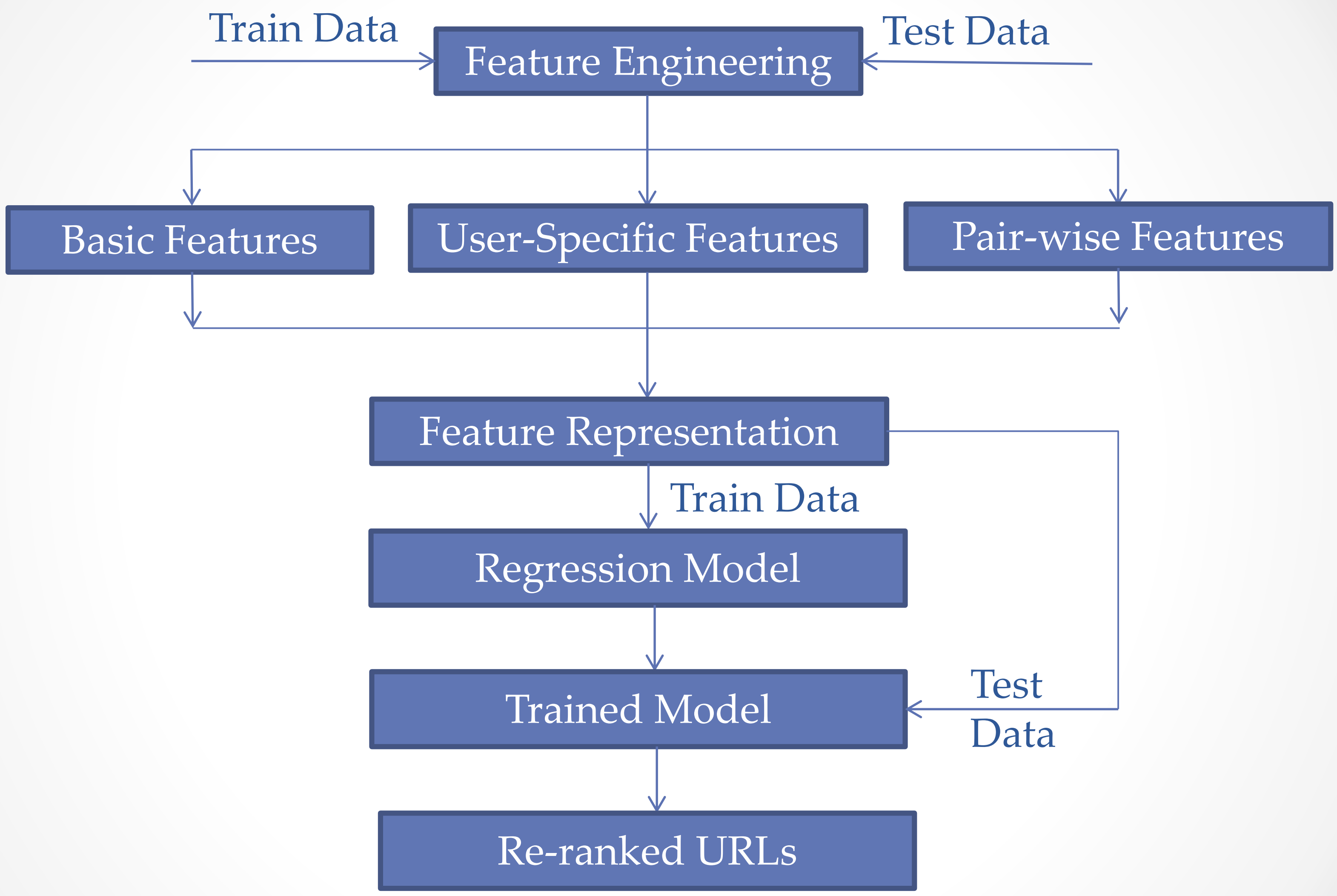
Query action (TypeOfRecord = Q or T):

SessionID TimePassed TypeOfRecord SERPID QueryID ListOfTerms ListOfURLsAndDomains

Click action (TypeOfRecord = C):

SessionID TimePassed TypeOfRecord SERPID URLID

Approach


Train Data
Feature Engineering
Test Data
Basic Features
User-Specific Features
Pair-wise Features
Feature Representation
Train Data
Regression Model
Trained Model
Test Data
Re-ranked URLs

Feature Engineering

- Initially each URL is labelled with a relevance label of:
  - 0 (irrelevant) : no clicks or click less than 50 time units
  - 1 (relevant): clicks with 50 to 399 time units
  - 2 (highly relevant): dwell time grater than 400 units

- **Basic Features** used include URL-query instance such as original position of URL for query , URLId , DomainId of URLs , TermIds of query and various joint features such as URLId X position , URLId X QueryId , DomainId X TermId etc.

Feature Engineering

- **User Specific Features** include for each user , dividing each URL in 5 categories based on relevance labels and whether URL was previously displayed or skipped to the User.

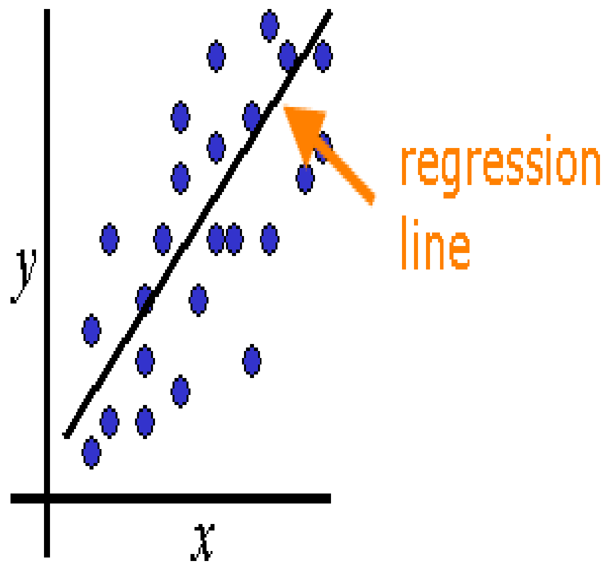
- **Pairwise Features** include for each query checking the relative position of a pair of URL

  f(URLi, URLj) = 1 if URLi occurs at better position than URLj

  f(URLi, URLj) = -1 if URLj occurs at better position than URLi

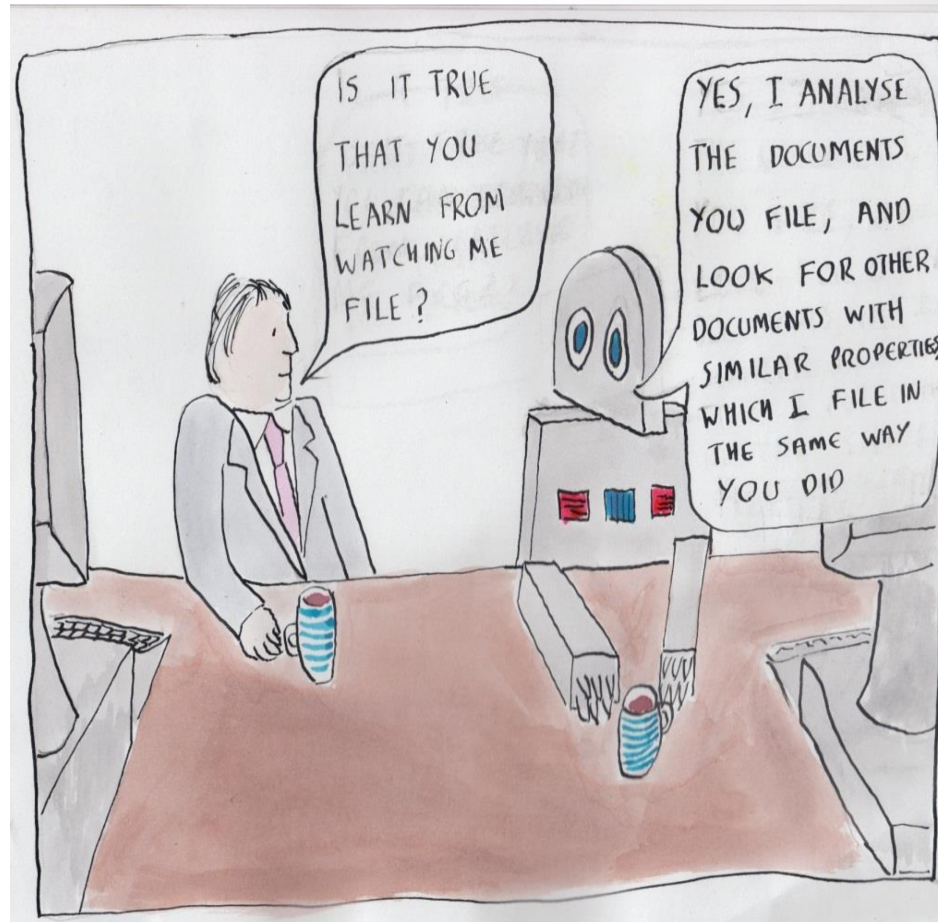
Logistic Regression

- Simple Logistic Regression Model was used.
- URL with positive relevance was labelled 1. Rest labelled -1.
- Relevance Labels assigned to each URL w.r.t to each User were used as Weight.
- **Out of Core** learning Algorithms are used as they can work with huge Data in hours.



Vowpal Wabbit for Logistic Regression

- Fast Machine Learning tool.
- Works on huge data in a parallel manner.
- RAM efficient.
- Fast Convergence to a Good Predictor.
- Handles TBs of Data in Matter of Hours.



Evaluation and Results

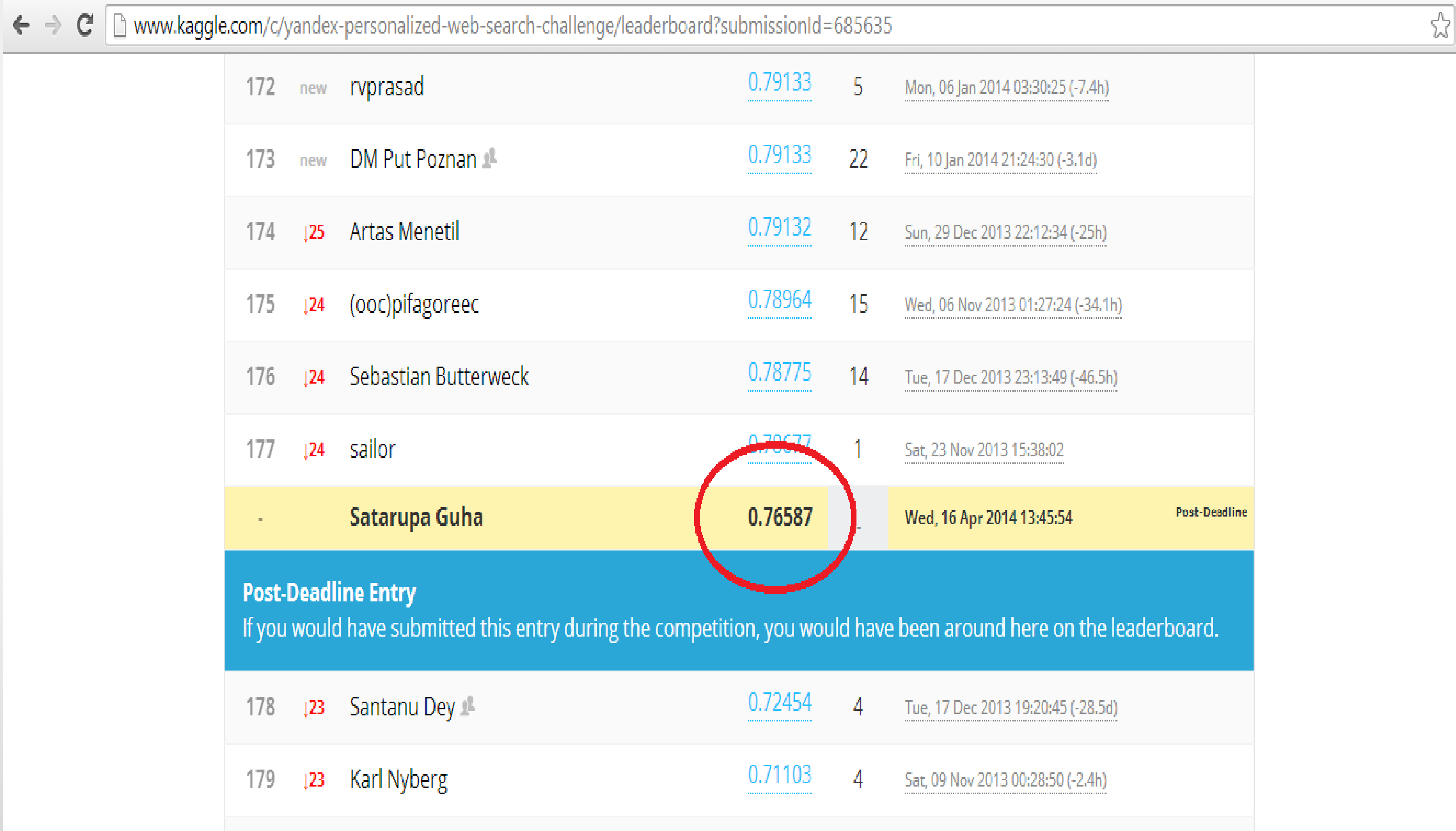
- We used NDCG (Normalized Discounted Cumulative Gain)for evaluation of our results.
- Calculated using the ranking of URLs for each query, and then averaged over queries.

$$DCG@10 = \sum_{i=1}^{10} \frac{2^{rel_i - 1}}{log_2(i+1)}$$

where rel_i, $i = 1,2, \ldots 10$ is the relevance list that contain 10 URL's relevance values. IDCG@10 is the maximum possible (ideal) DCG for a given set of queries, documents, and relevance value. Then, NDCG@10 is given by

$$NDCG@10 = \frac{DCG@10}{IDCG@10}$$

- Uploaded the final Result file online on Kaggle Portal for Evaluation.
- Our Team (Satarupa Guha) Got NDCG Score 0.76857
- Team who secured first Position got 0.80725



172	new	rvprasad	0.79133	5	Mon, 06 Jan 2014 03:30:25 (-7.4h)
173	new	DM Put Poznan	0.79133	22	Fri, 10 Jan 2014 21:24:30 (-3.1d)
174	↓25	Artas Menetil	0.79132	12	Sun, 29 Dec 2013 22:12:34 (-25h)
175	↓24	(ooc)pifagoreec	0.78964	15	Wed, 06 Nov 2013 01:27:24 (-34.1h)
176	↓24	Sebastian Butterweck	0.78775	14	Tue, 17 Dec 2013 23:13:49 (-46.5h)
177	↓24	sailor	0.78677	1	Sat, 23 Nov 2013 15:38:02
-		**Satarupa Guha**	**0.76587**	-	**Wed, 16 Apr 2014 13:45:54** Post-Deadline
178	↓23	Santanu Dey	0.72454	4	Tue, 17 Dec 2013 19:20:45 (-28.5d)
179	↓23	Karl Nyberg	0.71103	4	Sat, 09 Nov 2013 00:28:50 (-2.4h)

Post-Deadline Entry
If you would have submitted this entry during the competition, you would have been around here on the leaderboard.

Thank You !!!