Combination of Content-Based User Profiling and Local Collective Embeddings for Job Recommendation

Mikhail Kamenshchikov
Avito.ru
18.09.2017

ACM RecSys Challenge 2017

XING

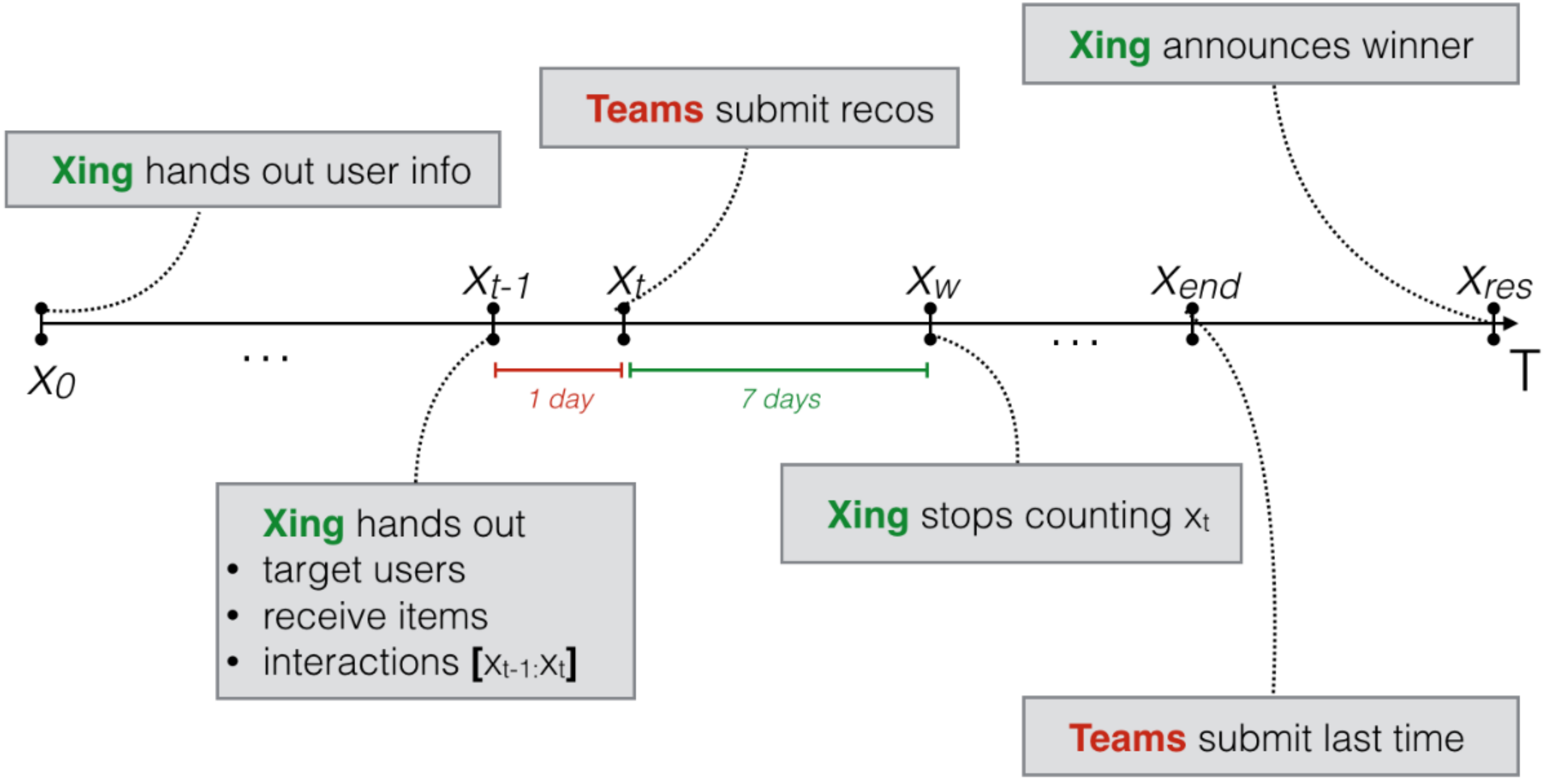
For a better working life

ACM RecSys Challenge

Problem Statement

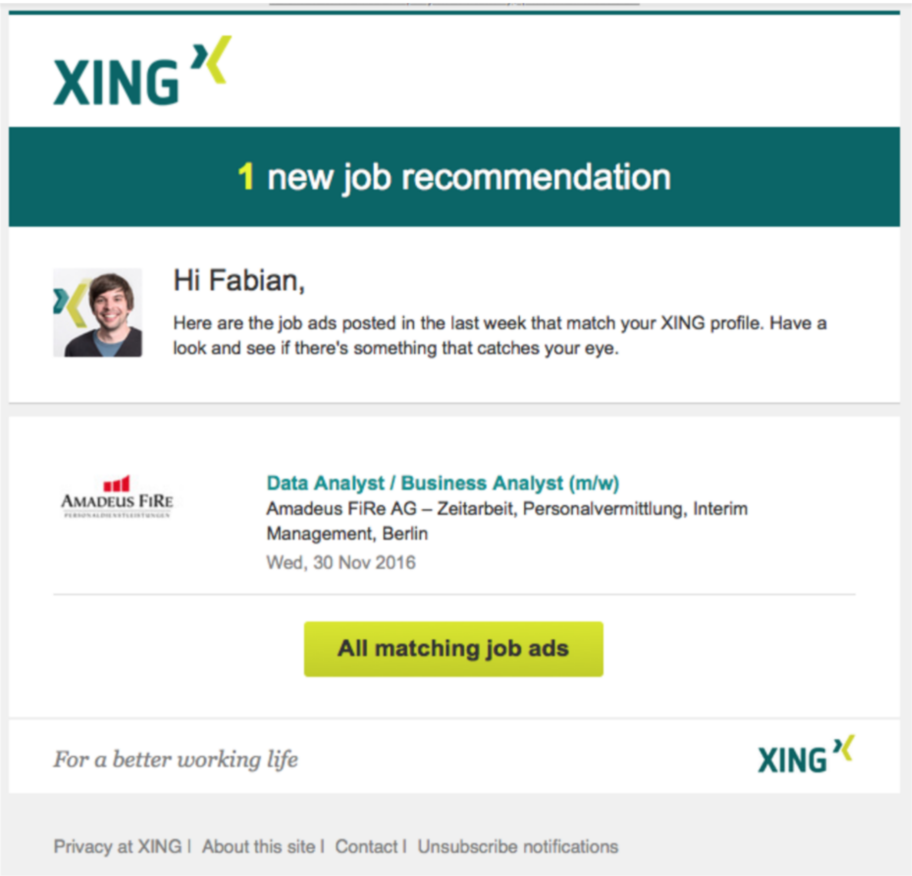
- Given a new job posting p, the goal is to identify those users that (a) may be interested in receiving the job posting as a push recommendation and (b) that are also appropriate candidates for the given job
- Challenge was focused on cold-start recommendation problem
- Challenge consisted of two phases: offline evaluation and online evaluation, where recommendations were shown to real users

Online Phase



Recommendations Delivery Channels

- Activity Stream
- Jobs Marketplace
- E-Mails
- Recruiter Tools
- Push Notifications (main channel for online phase)



Provided Data

- User Data
- Item Data
- User-Item interactions
- Target Users
- Target Items

User Data

- ID
- Keywords from current job title (obfuscated)
- Career level
- Discipline
- Industry
- Geography (country, region)
- Working experience
- Premium

Item Data

- ID
- Keywords from item title
- Keywords from item description
- Career level
- Discipline
- Industry
- Geography
- Employment
- Working experience
- Premium

Interactions

- User ID
- Item ID
- Timestamp
- Interaction type (click, delete, bookmark, etc..)

Scoring

```
score(targetItems) = targetItems.map(item => score(item, recommendations(item))).sum
```

```
score(item, users) =
  users.map(u => userSuccess(item,u)).sum + itemSuccess(item, users)

  userSucess(item, user) =
    (
        if (clicked) 1 else 0
      + if (bookmarked || replied) 5 else 0
      + if (recruiter interest) 20 else 0
      - if (delete only) 10 else 0
    ) * premiumBoost(user)

  premiumBoost(user) = if (user.isPremium) 2 else 1

  itemSuccess(item, users) =
    if (users.filter(u => userSuccess(item, u) > 0).size >= 1) {
      if (item.isPaid) 50
      else 25
    } else 0
```

Avito

Baseline solution

- Extract features from user-item interactions
- Target - positive interaction
- XGBoost
- Features:

  - number of matches in ids (int)

  - discipline match (binary)

  - career level match (binary)

  - industry match (binary)

  - country match (binary)

  - region match (binary)
- Score - 10004

Content-Based User Profiling: Title Match

- Score pairs with non-empty title intersection
- For each token *t* calculate 3 IDF-like measures on User Title, Item Title and Item Tags:

$$F_t = \log\left(\frac{\#\text{unique tokens}}{\#\text{token occurrences}}\right)$$

- Total token score is calculated as following:

$$\text{score}_t = \frac{20 * UF_t * IF_t * TF_t}{\sqrt{|u|}}$$

- Pair score is a sum of token scores in title intersection:

$$\text{score}(u, i) = \sum_t score_t$$

Content-Based User Profiling: User Interest Title Match

- Quite similar to previous, but use user interactions history
- Calculate similarity between titles / tags of clicked items

Content-Based User Profiling: Rankers

- Base score is calculated as mentioned above
- Ranker is some multiplicative weight, based on similarity of user/item parameters
- Career Level Ranker
- Discipline Ranker
- Industry Ranker
- User Behaviour Ranker
- Premium Ranker

Content-Based User Profiling: Career Level Ranker

- Career Level Difference: $CLD(u, i) = |U_{CL} - I_{CL}|$

$$w_{CL}(u,i) = \begin{cases} 1.2, & if\ CLD(u,i) \leq 1 \\ 0.7, & if\ CLD(u,i) = 3 \\ 0.5, & if\ CLD(u,i) \geq 4 \end{cases}$$

$$score(u,i) = w_{CL} * score(u,i)$$

Avito

Content-Based User Profiling: Discipline & Industry Rankers

- On exact field match multiply score by $w > 1$
- Otherwise multiply score by $w < 1$

Content-Based User Profiling: User Behaviour Ranker

- Good feature - user clicked on the item with same features
- Positive / Negative actions ratio
- User was recently active (only offline phase)
- User have already clicked on this item (only offline phase)

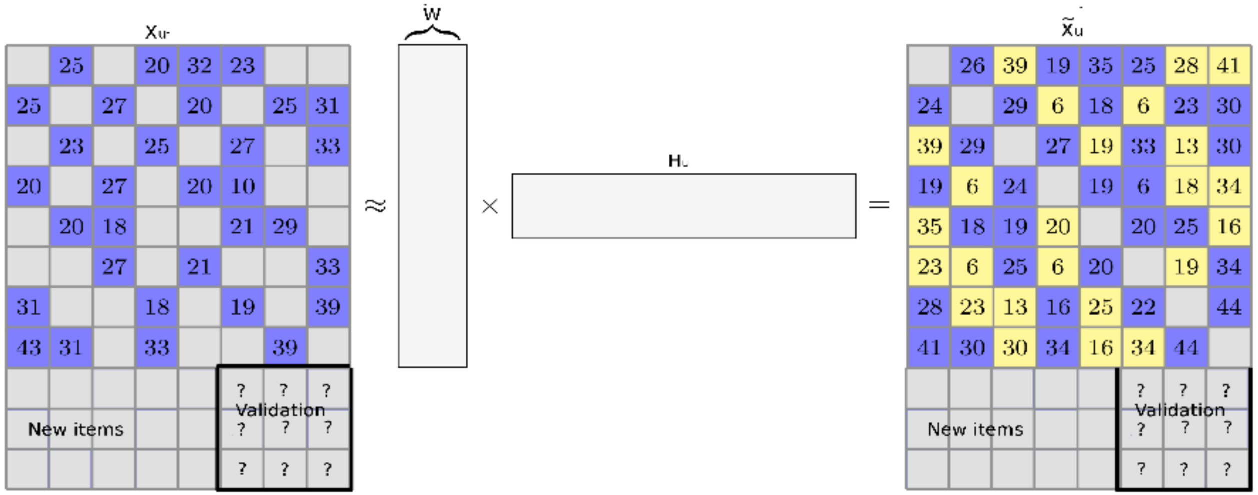
Content-Based User Profiling: Premium Ranker

- Premium users & items contribute more to target metric
- Increase weight for such users / items

Content-Based User Profiling: Final Predictions

- Score all user-item pairs
- Apply rankers
- Threshold scores by some value
- Take top-100 users for each item
- Works pretty fast (~30 min)
- Offline score: 32493

Matrix Factorization & Local Collective Embeddings



MF: $\min : J = ||X_u - WH_u||^2 + \lambda(||W||^2 + ||H_u||^2)$

LCE:

$$\min : J = \frac{1}{2}[\alpha||\mathbf{X_s} - \mathbf{WH_s}||^2 + (1-\alpha)||\mathbf{X_u} - \mathbf{WH_u}||^2 + \\ + \lambda(||\mathbf{W}||^2 + ||\mathbf{H_s}||^2 + ||\mathbf{H_u}||^2)] \quad (1)$$

Local Collective Embeddings

- Let A - nearest neighbour graph with *n* edges (nearest item pairs):

$$
\begin{aligned}
S &= \frac{1}{2}\sum_{i,j=1}^{n} ||w_i - w_j||^2 \mathbf{A}_{ij} \\
&= \sum_{i=1}^{n}(w_i^T w_i)\mathbf{D}_{ii} - \sum_{i,j=1}^{n}(w_i^T w_j)\mathbf{A}_{ij} \\
&= \mathrm{Tr}(\mathbf{W}^{\mathrm{T}}\mathbf{D}\mathbf{W}) - \mathrm{Tr}(\mathbf{W}^{\mathrm{T}}\mathbf{A}\mathbf{W}) = \mathrm{Tr}(\mathbf{W}^{\mathrm{T}}\mathbf{L}\mathbf{W})
\end{aligned}
$$

- Optimization Problem:

$$
\begin{aligned}
\min : J = \frac{1}{2}[\alpha||\mathbf{X_s} - \mathbf{W}\mathbf{H_s}||^2 + (1-\alpha)||\mathbf{X_u} - \mathbf{W}\mathbf{H_u}||^2 + \\
+ \beta\mathrm{Tr}(\mathbf{W}^{\mathrm{T}}\mathbf{L}\mathbf{W}) + \lambda(||\mathbf{W}||^2 + ||\mathbf{H_s}||^2 + ||\mathbf{H_u}||^2)]
\end{aligned}
$$

Final Ensembling

- Weighted sum of two models: $score = 0.8 * score_{LCE} + (score_{CB})^{0.15}$
- +8.1% on local validation
- Scores were much lower on the last two weeks, so final results include only CB-model

Offline Phase Results

Official, April 16th

Rank	Team	Score
1	Lunatic Goats	71002
2	layer6.ai	68072
3	Hushpar	61427
4	rho	59461
5	Get all the data	57043
6	chome	53566
7	Amethyst	50069
8	leavingseason	43183
9	LongLiveSea	41472
10	Druid	39579
11	guang	39344
12	Donau	38014
13	YunOS	36590
14	chiyou	36616
15	Think More	36133
16	better	35137
17	Taoist	35112
18	Avito	32493
19	passionate17	32765
20	RecoPassion	30991

Online Phase Results

Rank	Team	Score
1.	**layer6.ai**	**10963**
2.	**Lunatic Goats @PoliMi**	**9741**
3.	**CTL@Fuji Xerox**	**9648**
4.	rho	9536
5.	leavingseason	9173
6.	Get all the data	8906
7.	Avito	8710
8.	Donau	7062
9.	poem in rain	6563
10.	YunOS	5444
11.	RecoPassion	3780
12.	Endeavour	3338
13.	Degree of Belief	3323
14.	Druid	3291
15.	Taoist	3133
16.	Hushpar	1852
17.	JKU-Alpha	1615
18.	Amethyst	834

Thank You!

vleksin@avito.ru
aostapets@avito.ru
makamenshchikov@avito.ru
dkhodakov@avito.ru
vnrubtsov@avito.ru