

Рассказ про участие в RecSys Challenge 2017

Андрей Остапец

Avito.ru

01-07-2017





- Рассказ о конкурсе
- Постановка задачи
- Решения
- Результаты

The ACM Recommender Systems conference



	The ACM Con Recomme	ference Series on nder Systems							
HOME	RECSYS 2017	PAST CONFERENCES	HONORS	WIKI	CONTACT			٩	
Como, I	taly, 27th-31st A	erence on Rec					IT7 (COMO) Conference	No. 1 Jole	
presentati systems. F	ion of new research i Recommendation is a	results, systems and technic a particular form of informa- te a list of information items	ues in the broa tion filtering, tha	d field of rec it exploits pa	ommender ist behaviors	Call for Co	ntributions		
on recom	mender systems, alo	Sys brings together the ma ng with many of the world?	Keynotes						
become the most important annual conference for the presentation and discussion of recommender systems research. RecSys 2017, the eleventh conference in this series, will be held in Como, Italy. It will bring together researchers and practitioners from academia and industry to							is		
compone	ieir latest results and nts in a range of inno 17 program will feat.	RecSys Ch	allenge						
		an industrial track and a de			ne-art in this	Summer 5	ichool		
		gh a rigorous full peer revie a USB drive and via the ACI	Location						
	ed and cited.					Registratio	on		
ACM RecS	ys 2017 will take plac	e in Como, Italy, from Augu	st 27-31, 2017.			Accommo	dations		
						Important	Dates		





Конкурс в 2016 и 2017 годах проводила компания XING. В 2016 году команда Avito участвовала в RecSys Challenge и заняла 7 место. Рассказ о решении

Проблемы прошлого года:

- Хорошо работали алгоритмы, которые выбирали лучшие вакансии среди рекомендованных платформой XING.
- Пользователи часто совершают повторные клики.



В этом году соревнование состояло из 2 этапов:

- 1. Офлайн-этап
- 2. Онлайн-этап

В рамках 2 этапов решалась одна задача:

Работодатель выложил новую вакансию р. Нужно определить тех пользователей, которые (a) могут быть заинтересованы в получении «пуша» с этой вакансией и (б) являются подходящими кандидатами на эту вакансию.



Для этих 2 этапов использовались одни и те же метрики и наборы данных имели одну и ту же структуру (практически). Офлайн-этап являлся отбором команд во вторую стадию конкурса.

- Топ-команды (которые преодолеют baseline от XING) могут участвовать во второй стадии конкурса.
- Победитель второй стадии победитель всего конкурса. Вторая стадия длилась 5 недель. Финальный результат сумма за 2 лучшие недели.

Использование рекомендаций



С офлайн-стадией все понятно (не более 100 пользователей для каждой вакансии)...

В онлайн-стадии фокус на «пуши»:

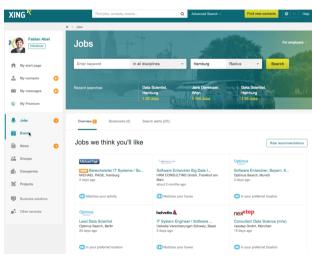
- Раз в 24 часа нужно отправить рекомендации организаторам.
- ullet Каждый пользователь может получить ≤ 1 «пуша» за сутки.
- Для каждой вакансии нужно предоставить список кандидатов из ≤ 250 пользователей.

Различные способы доставки рекомендации:

- Лист рекомендаций.
- Домашняя страница.
- Электронная почта.
- Лист рекомендаций для рекрутера.

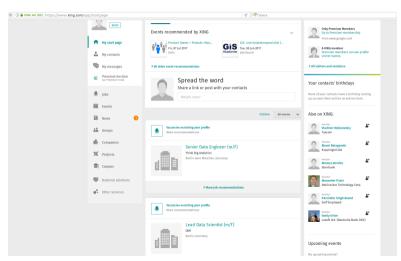
Лист рекомендаций





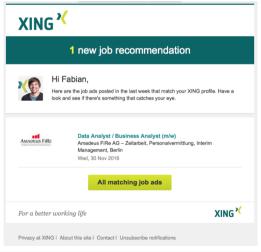
Домашняя страница





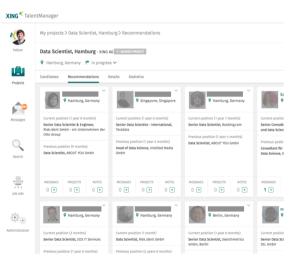
Электронная почта





Лист рекомендаций для рекрутера





Данные



5 файлов:

- Данные о пользователях
- Данные о вакансиях
- Данные о предыдущих взаимодействиях пользователей и вакансий
- Целевые пользователи
- Целевые вакансии



Пользователи:

- ID
- Ключевые слова из текущей работы ([1698419,46526,2394936,219933,1469577])
- Текущий уровень карьеры (от 1 до 6)
- Дисциплина (HR, IT, ...)
- Индустрия (Интернет, Финансы, ...)
- География (страна, регион)
- Опыт работы (стаж, число мест работы)
- Образование
- Премиум
- (Только офлайн-стадия) Интерес к смене работы
- ...

Данные



Вакансии:

- ID
- Ключевые слова из заголовка
- Ключевые слова из описания
- Дисциплина
- Индустрия
- География (страна, регион)
- Занятость (полная, удаленная, ...)
- Время создания
- Премиум
- •



Взаимодействия:

- ID пользователя
- ID вакансии
- Время
- Тип взаимодействия
 - 0 = показ
 - 1 = клик
 - 2 = добавление в закладки
 - 3 = сообщение
 - 4 = удаление
 - 5 = интерес рекрутера

Тренировочная выборка для онлайн-стадии: 1M пользователей, 853K вакансий, 88.7M показов, 4.2M взаимодействий (84.1% clicked, 8.5% deleted, 4.3% bookmarked, 2.3% replied, 0.8% recruiter interest).



Для каждой вакансии *item* алгоритм должен отдавать список рекомендаций *recommendations(item)*.

```
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

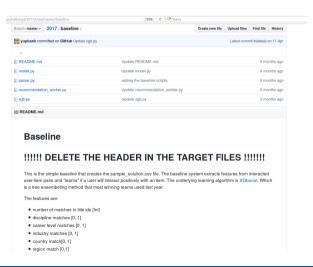
itemSuccess(item, users) =
    if (users.filter(u => userSuccess(item, u) > 0).size >= 1) {
        if (item.isPaid) 50
        else 25
        else 0
```

Вызовы



- Баланс между интересами пользователей и интересами рекрутеров.
- Баланс между релевантностью и заработком.
- Проблема «холодного старта».
- Умная посылка «пушей».





Замечания к baseline для офлайн-стадии



- 322М взаимодействий; 75К целевых пользователей; 47К целевых вакансий.
- Взаимодействия содержат огромное количество полных дубликатов. После удаления получаем 220М взаимодействий (всего 28М, если удаляем повторные взаимодействия).
- Выкидываем просмотры. Получаем обучающую выборку в 6М примеров.
- Делаем предсказание для пары (user, item) только если есть пересечения в заголовках. Около 50М предсказаний.
- На сервере не работал xgboost. Использование H2O!
- Для всех вакансий рекомендуем топ-100 самых активных пользователей: результат 519.
- 6 фич дают результат в 10004 на лидерборде, 25 ?

Что-то идет не так...

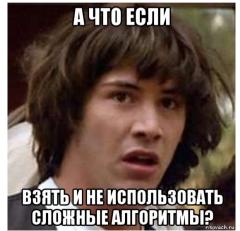


Submission History

	Score	Rank (at time of submission)	Labbel	Time
	-5552.00	16		2017-03-11T18:27:23.000+01:00
-67	1955.00	14		2017-03-10T22:59:52.000+01:00
(4:	3863.00	13		2017-03-10T17:05:04.000+01:00
١	4294.00	13	Fix	2017-03-10T07:54:37.000+01:00
	10218.00	11	lc7653	2017-03-09T20:42:59.000+01:00
	11731.00	4	le37	2017-03-09T20:16:26.000+01:00
	8339.00	13	last_actions	2017-03-09T19:51:00.000+01:00
	10950.00	5	+user_id	2017-03-09T01:58:58.000+01:00
	8214.00	8	Fix: h2o with clicks features	2017-03-08T18:31:41.000+01:00
	11332.00	3	h20 with clicks features	2017-03-08T10:39:01.000+01:00
	10714.00	3	h2o (500 trees)	2017-03-07T09:30:15.000+01:00
	8988.00	6	h2o	2017-03-06T21:08:58.000+01:00
	10004.00	3	baseline	2017-03-05T20:33:46.000+01:00
'	519.00	4	top users (5-100,3-30,1-211 events)	2017-03-05T20:32:41.000+01:00



Вероятная проблема: не смог настроить негативный семплинг.



Интересы пользователя



- Для вычисления оценки для пары (пользователь, вакансия) используются суммы 3 типов:
 - Близость по профилю пользователя
 - Интересы к параметрам, исходя из истории пользователя.
 - Прошлые взаимодействия айтема и пользователя (неактуально для онлайн-части).
- Для каждой пары (пользователь, вакансия) считаем оценку, если есть либо пересечение по заголовкам или пересечение с тегами вакансии, либо пользователь просматривал объявление, которое пересекается по заголовку или тэгам с вакансией для которой сейчас вычисляется оценка. Пример: «Аналитик» совершал клик на «Data Scientist», потенциальные рекомендации это вакансии со словами «Аналитик», «Data», «Scientist».

Близость ключевых слов пользователя к заголовку вакансии



Для каждого токена t вычислялось 3 типа «аналога» IDF по словарям с UF - User Title, IF - Item Title, TF - Item Tags:

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

Для каждого пользователя u близость по заголовку вакансий и профиля пользователя формировалась следующим образом:

$$score_t = \frac{w * UF_t * IF_t * TF_t}{\sqrt{|u|}}$$

где |u| - это кол-во уникальных ключевых слов пользователя.

Финальная оценка для пары (u_k, i_k) вычисляется как сумма по всем токенам t из пересечения ключевых слов заголовка пользователя и заголовка вакансии:

$$score(u_k, i_k) = \sum score_t$$

Близость по тексту



Схожим образом считались:

- Близость ключевых слов пользователя к тегам вакансии
- Близость токенов из заголовка вакансий, на которые кликал пользователь к заголовку вакансии.
- Близость токенов из тегов вакансий, на которые кликал пользователь к тегам вакансии.

Другой вариант подсчета:

- Полное вложение ключевых слов пользователя в ключевых слова из заголовка вакансии
- . . .



Близость по профилю для уровня карьеры была основана на разнице между уровнем пользователя и вакансии: $CLD(u,i) = |u_{CL} - i_{CL}|$

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

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

Индустрия и дисциплина - категориальные признаки: оценка умножалась на w, где w>1 при точном совпадении и w<1 иначе.

Близость по взаимодействиям считалась немного иначе.

Основная идея: модифицировать оценку исходя из близости распределения действий пользователя к параметрам вакансии с поправкой на общее число кликов пользователя.

Клики пользователя и статус объявления



Ранкер основан на предыдущих кликах пользователя:

- 1. Хороший признак: пользователь кликал на вакансию с таким же заголовком, уровнем карьеры, индустрии, дисциплины.
- 2. Отношение негативных/позитивных действий.
- 3. (Только для офлайн-этапа) Пользователь недавно был активен на сайте.
- 4. (Только для офлайн-этапа) Пользователь уже кликал на эту вакансию.

Ранкер, оптимизирующий функционал задачи:

- (Только для офлайн-этапа) Желание пользователя сменить работу
- Премиум-пользователь
- VIP-вакансия

Локальная валидация



На офлайн-этапе использовалась последние P дней так, чтобы кол-во пользователей и вакансии в валидации совпадало с тестом.

Целевые пользователи и вакансии: те, которые совершили/получили хотя бы одно действие.

Минусы:

- В валидации не было событий с типом «интерес рекрутера» (на деле этого события не было в онлайн-этапе).
- Совпадение с Public Leaderboard примерно в 70% случаев.

Онлайн-этап: валидация на всех пользователях и вакансиях с начала онлайн-этапа, которые совершили/получили хотя бы одно действие.

Улучшения



Submission History

Score	Rank (at time of submission)	Labbel	Time
32493.00	6	30	2017-03-16T13:53:57.000+01:00
29844.00	7	29	2017-03-15T23:15:12.000+01:00
31273.00	6	28	2017-03-15T14:04:52.000+01:00
27870.00	6		2017-03-14T10:50:11.000+01:00
27999.00	5		2017-03-13T20:09:33.000+01:00
25561.00	4		2017-03-12T20:47:03.000+01:00
22914.00	5	23	2017-03-12T19:58:56.000+01:00
19715.00	8	21	2017-03-12T14:47:44.000+01:00
21066.00	6		2017-03-12T14:15:58.000+01:00
289.00	16	sample	2017-03-12T13:41:40.000+01:00
20580.00	5	le-19	2017-03-12T12:00:46.000+01:00
2968.00	16	fix 10	2017-03-12T11:11:36.000+01:00
20675.00	4		2017-03-11T22:54:10.000+01:00
20073.00	4		2017-03-11T21:41:14.000+01:00
18274.00	5	all types interactions	2017-03-11T20:38:50.000+01:00
19250.00	4		2017-03-11T19:22:21.000+01:00
-5552.00	16		2017-03-11T18:27:23.000+01:00
1955.00	14		2017-03-10T22:59:52.000+01:00

Описание	Результат	Место
Интересы по тэгам + заголовок из профиля	19250	4
Разделение по типам кликов + использования уровня	20073	4
карьеры		
Совпадение заголовка, уровня карьеры, индустрии,	21066	6
дисциплины		
Подбор весов	25561	4
Учет ВИП-статуса	27999	5
Добавление IDF	31273	6
Штраф за низкое соотношение кликов/удалений	32493	6

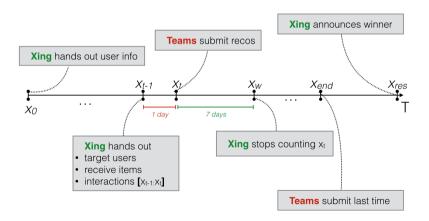


Official, April 16th

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

Онлайн-этап





Первая неделя



Rank	Team	Score		
1	layer6.ai	4391		
2	Get all the data	4261		
3	Lunatic Goats	4037		
4	rho	3784		
5	chome	3637		
6	leavingseason	3601		
7	Donau	3104		
8	poem in rain	2651		
9	RecoPassion	2202		
10	Endeavour	985		
11	Avito	983		
12	Taoist	717		
13	Degree of Belief	508		

Вторая неделя



Rank	Team	Score
1	chome	5408
2	layer6.ai	5408
3	rho	5236
4	Avito	4815
5	Lunatie Goats	4757
6	leavingseason	4364
7	Get all the data	3737
8	poem in rain	3556
9	Donau	3223
10	YunOS	2526
11	Taoist	2187
12	Degree of Belief	1876
13	Endeavour	1598
14	Druid	933
15	Hushpar	326
16	RecoPassion	315

Третья неделя



Rank	Team	Score
1	layer6.ai	5555
2	Lunatic Goats	4984
3	leavingseason	4809
4	Get all the data	4645
5	rho	4300
6	chome	4240
7	Avito	3895
8	Donau	3839
9	poem in rain	3007
10	YunOS	2918
11	Druid	2167
12	Endeavour	1740
13	RecoPassion	1578
14	Degree of Belief	1447
15	Hushpar	1343
16	Taoist	946
17	Amethyst	395

Комбинация решений



За 2 недели до конца удалось настроить линейную комбинацию решений. Василий Лексин построил решение на основе Local Collective Embeddings: Рассказ об этом решении (Avito Data Science meetup: Рекомендации). Улучшение результата на локальной валидации: +8.1%

Четвертая неделя



Rank	Team	Score
1	layer6.ai	2034
2	leavingseason	1638
3	rho	1466
4	chome	1462
5	Lunatic Goats	1421
6	Endeavour	1372
7	Degree of Belief	1334
8	Avito	1234
9	Get all the data	1197
9	Get all the data Donau	1197 1184
10	Donau	1184
10	Donau Druid	1184 1124
10 11 12	Donau Druid YunOS	1184 1124 1062
10 11 12 13	Donau Druid YunOS poem in rain	1184 1124 1062 1027
10 11 12 13	Donau Druid YunOS poem in rain Hushpar	1184 1124 1062 1027 509

Пятая неделя



Rank	Team	Score
1	JKU-Alpha	1232
2	leavingseason	533
3	layer6.ai	532
4	Get all the data	506
5	chome	475
6	Lunatic Goats	436
7	poem in rain	368
8	Donau	352
9	Donau Avito	352 302
9	Avito	302
9	Avito Endeavour	302 266
9 10 11	Avito Endeavour Degree of Belief	302 266 264
9 10 11 12	Avito Endeavour Degree of Belief Druid	302 266 264 257
9 10 11 12 13	Avito Endeavour Degree of Belief Druid YunOS	302 266 264 257 155

Финальные результаты



Week 18		Week 19			Week 20			Week 21			Week 22				
Rank Team	Score	Rank	Team	Score	Rank	Team	Score	Rank	Team	Score	Rank	Team	Score	Score	Ran
1 layer6.ai	4391	2	layer6.ai	5408	1	layer6.ai	5555	1	layer6.ai	2034	3 la	ayer6.ai	532	10963	
3 Lunatic Goats	4037	5	Lunatic Goats	4757	2	Lunatic Goats	4984	5	Lunatic Goats	1421	6 I	unatic Goats	436	9741	. :
5 chome	3637	1	chome	5408	6	chome	4240	4	chome	1462	5 c	home	475	9648	. :
4 rho	3784	3	rho	5236	5	rho	4300	3	rho	1466	17 r	ho	0	9536	
6 leavingseason	3601	6	leavingseason	4364	3	leavingseason	4809	2	leavingseason	1638	210	eavingseason	533	9173	
2 Get all the data	4261	7	Get all the data	3737	4	Get all the data	4645	9	Get all the data	1197	40	Get all the data	506	8906	(
11 Avito	983	4	Avito	4815	7	Avito	3895	8	Avito	1234	9.4	vito	302	8710	
7 Donau	3104	9	Donau	3223	8	Donau	3839	10	Donau	1184	81	Oonau	352	7062	
8 poem in rain	2651	8	poem in rain	3556	9	poem in rain	3007	13	poem in rain	1027	7 p	oem in rain	368	6563	
14 YunOS	0	10	YunOS	2526	10	YunOS	2918	12	YunOS	1062	13 3	/unOS	155	5444	10
9 RecoPassion	2202	16	RecoPassion	315	13	RecoPassion	1578	18	RecoPassion	0	18 F	RecoPassion	0	3780	1:
10 Endeavour	985	13	Endeavour	1598	12	Endeavour	1740	6	Endeavour	1372	10 F	Indeavour	266	3338	1:
13 Degree of Belief	508	12	Degree of Belief	1876	14	Degree of Belief	1447	7	Degree of Belief	1334	11 [Degree of Belie	264	3323	1:
16 Druid	0	14	Druid	933	11	Druid	2167	11	Druid	1124	12 [Druid	257	3291	14
12 Taoist	717	11	Taoist	2187	16	Taoist	946	16	Taoist	400	14 T	aoist	74	3133	15
15 Hushpar	0	15	Hushpar	326	15	Hushpar	1343	14	Hushpar	509	16 F	Iushpar	8	1852	16
18 JKU-Alpha	0	18	JKU-Alpha	0	18	JKU-Alpha	0	17	JKU-Alpha	121	1 J	KU-Alpha	1232	1353	
17 Amethyst	0	17	Amethyst	0	17	Amethyst	395	15	Amethyst	439	15 A	Amethyst	37	834	18

Офлайн-стадия



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

Что не получилось



- 1. Учет особенностей функционала задачи
- 2. Использование сложной модели
- 3. Получение обратной связи по изменению качества работы модели на онлайн-этапе
- 4. У организаторов: провести конкурс без «накладок».