ACM RecSys Challenge 2016

Решение команды Avito (7 место)

Василий Лексин, Андрей Остапец

Avito.ru

9 июля 2016 г.

- **Users** — данные по пользователям сайта XING.com: jobroles, career_level, discipline, industry, location, experience, education.
- **Items** — данные по вакансиям: title, career_level, discipline, industry, location, employment_type, tags, created_at, active_during_test.
- **Impressions** — какие вакансии и когда рекомендовались пользователю.
- **Interactions** — по каким вакансиям и когда пользователь совершал действия: clicked, bookmarked, replied, deleted.

Интервал дат: 2015-08-19 – 2015-11-09

Impressions

- 201M уникальных user-item-week сочетаний.
- 2.7M уникальных пользователей.
- 846K уникальных вакансий.

Interactions

- 8.8M events: **clicked** – 7.2M, deleted – 1.0M, **replied** – 422K, **bookmarked** – 206K.
- 785K уникальных пользователей.
- 1.03M уникальных вакансий.
- 2.8M из 6.9M (40%) пар user-item содержатся в impressions.

Постановка задачи

Описание данных: target users и items

150K пользователей, которым нужно порекомендовать, из них:

- у 39.7K (26.5%) нет событий;
- у 59.5K (39.6%) менее 1 события;
- у 70.6K (47.1%) менее 2-х событий.

327K активных вакансий, из них:

- у 129K (39.5%) нет событий;
- у 164K (50.1%) менее 1 события;
- у 188K (57.6%) менее 2-х событий.

Для каждого из 150K целевых пользователей отобрать по 30 вакансий, на которые пользователь наиболее вероятно совершил одно из действий: clicked, bookmarked or replied в следующую после 2015-11-09 неделю.

$$score = \sum_{all\_users} 20(P2 + P4 + R + S) + 10(P6 + P20),$$

где Pk – precision at k; R – recall; S – user success (хотя бы одно положительное действие на любой из 30 вакансий).

Решение команды

Пример пользовательской сессии

CV

experience_n	experience	discipline_id_user	country_use	region_user	jobroles
5 or more entr	10-15 years	Sales & Commerce	Germany	Bavaria	['962959', '283291', '**502342**']

SESSIONS

created_at	impression	discipline_id_item	country_ite	region_item	title	tags
09-01 1:27	1	Production & Manufac	Germany	Baden-Württer	['620383', '1118975']	['102823', '1335184', '624061', '73234', '1604815', '7862074
09-01 1:27	1	Other Disciplines	Germany	Hamburg	['4572761', '3543754', '19689	['993979', '2426818', '792504', '4425481', '494116', '976257
09-01 1:27	1	Other Disciplines	Germany	Berlin	['18091']	['4198994', '4182900', '4354582', '1399193', '1377742', '358
09-02 0:21	1	Health, Medical & Soci	non_dach	not specified	['165415', '1986087', '2585795	['2426818', '3726822', '792504', '1830721', '184797', '32562
09-04 20:46	0	IT & Software Develop	Germany	Brandenburg	['655030']	['1491612', '972718', '2426818', '2383555', '4483314', '43211
09-08 20:16	1	Other Disciplines	Germany	Hamburg	['2915824', '4035399', '15676	['2110329', '503870', '2426818', '1437930', '2245760', '3592
09-08 22:40	1	Sales & Commerce	Germany	not specified	['3418410', '3413328']	['686709', '2036672', '3794933', '**502342**', '3413328', '11785
09-08 22:41	1	IT & Software Develop	non_dach	not specified	['3408137']	['2632767', '1491612', '2245760', '689679', '1565617', '43211
09-09 22:58	1	Administration	Germany	Berlin	['4141254', '1118975']	['4162864', '1491612', '1565617', '689679', '4204056', '1545
09-09 23:00	0	Production & Manufac	Germany	Lower Saxony	['4454260', '**502342**']	['543177', '4160943', '2501578', '4329775', '3085937', '2342
09-09 23:01	0	Other Disciplines	Germany	Lower Saxony	['1567693', '568776']	['1178568', '1248479', '370640', '2342166', '94890', '3794933
09-09 23:02	1	Health, Medical & Soci	non_dach	not specified	['1567693']	['1178568', '1565617', '1491612', '1601282', '2380081', '494
09-09 23:08	0	Finance, Accounting &	Austria	not specified	['2865345', '3294368']	['3391339', '3176219', '4499767', '2426818', '798840', '3729
09-09 23:08	0	Production & Manufac	Germany	Lower Saxony	['494116']	['3176219', '159096', '3391339', '4499767', '494116', '37295
09-09 23:08	0	Health, Medical & Soci	Switzerland	not specified	['128836', '1836819']	['1798728', '128836', '675557', '2976021']
09-09 23:09	0	Production & Manufac	Austria	not specified	['2846960', '76751', '4227194'	['2846960', '2632767', '3872048', '3939477', '1469275', '169
09-09 23:09	0	Engineering & Technica	Germany	Berlin	['4141254']	['749243', '362736', '692505', '3669898', '624061', '494116']
09-10 0:59	1	Production & Manufac	Austria	not specified	['1118975', '3478136']	['**502342**', '4151211', '4439048', '3210328', '624061', '31375
09-10 0:59	1	Engineering & Technica	Germany	Bavaria	['128836', '76887']	['816406', '3347566', '**502342**', '4425481', '4160943', '16048
09-10 0:59	1	Engineering & Technica	Germany	North Rhine-W	['1119117', '3705605', '34781	['2896178', '1357922', '7031982', '1491612', '1830721', '133
09-10 1:00	0	Other Disciplines	Austria	not specified	['2915824', '4035399', '15676	['1565617', '1496767', '82994', '1625244', '1941434', '12318
09-10 1:00	1	Teaching, R&D	Germany	North Rhine-W	['1986087']	['3144475', '4245173', '3096790', '655817', '2969837', '4321
09-10 1:00	0	Other Disciplines	Germany	not specified	['2573697', '4035399', '44843	['3457262', '3658040', '2126708', '2110329', '2630003', '401
09-10 1:00	0	Engineering & Technica	Germany	Baden-Württer	['2140778', '3241763']	['1734724', '2000691', '4425481', '2111897', '577140', '9489
09-10 1:00	1	Management & Corpor	Germany	Bavaria	['494116', '1119117', '238737	['4245173', '1231885', '272304', '4140111', '4321623', '1830
09-10 1:00	1	Teaching, R&D	Germany	Bremen	['494116', '3376322']	['3376322', '2630003', '3371139', '1491612', '253918', '1768
09-10 1:01	0	Production & Manufac	Germany	Lower Saxony	['4454260', '**502342**']	['543177', '4160943', '2501578', '4329775', '3085937', '2342
09-10 1:01	0	Other Disciplines	Germany	North Rhine-W	['494116', '3408137']	['1565617', '1491612', '2245760', '2862074', '1231885', '183
09-10 1:02	0	Production & Manufac	Germany	Bavaria	['**502342**']	['1491612', '689679', '3945636', '2810778', '154542', '45335
09-10 1:02	0	Production & Manufac	Germany	Schleswig-Hols	['398644', '2865345']	['1565617', '4252882', '4162864', '1491612', '689679', '9034
09-10 1:02	1	Administration	Germany	Berlin	['4141254', '1118975']	['4162864', '1491612', '1565617', '689679', '4204056', '1545
09-10 1:02	0	Production & Manufac	Germany	Bavaria	['4013443']	['1288549', '339347', '2862074', '4425481', '4575422', '2426
10-16 22:33	1	Other Disciplines	Switzerland	not specified	['1986087', '1971244']	['1178568', '2245760', '1491612', '4162864', '689679', '3017

Решение команды

Интересные наблюдения по данным

- Значительная доля пользователей и вакансий с малым количеством событий или без событий, значит нужен гибридный подход, учитывающий контент.
- Impressions слабо меняются во времени, 40% действий пользователя совершаются из impressions, то есть наличие пары user-item в impressions хороший признак.
- Географические признаки почему-то не работают.

- Выделяем последнюю полную неделю из interactions в контроль.
- Выделяем случайных 10K пользователей из тех, что совершал действия в эту неделю.
- Отбрасываем старые вакансии (created_at более месяца), которые никто не кликал.
- Хорошая корреляция с лидербордом.

Решение команды

Item-based collaborative filtering

Метрики сходства:

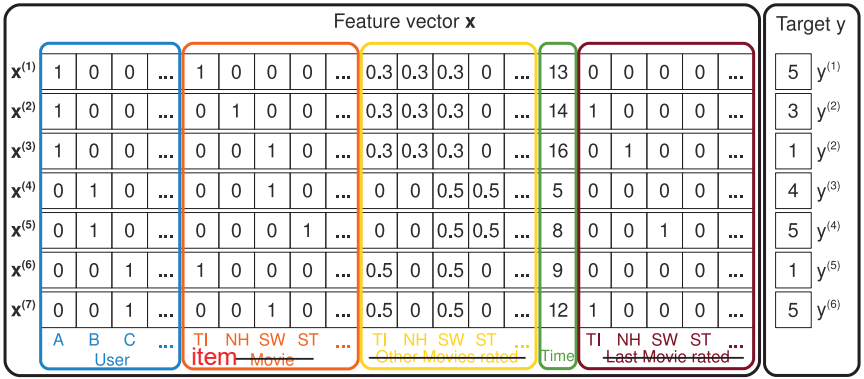
- Jaccard
- Cosine
- Pearson

Типы событий, на которых обучаемся:

- Interactions без deleted
- Только clicked
- Только impressions

Решение команды

Factorization Machines: общая идея



Users - все признаки (OneHotEncoder)

- jobroles
- career_level, discipline_id, industry_id
- country, region
- experience: n_entries_class, years, years_in_current
- edu: degree, fieldofstudies

Items - все признаки, кроме latitude и longitude (OneHotEncoder)

- title, tags
- career_level, discipline_id, industry_id
- country, region
- employment_type

- Количество латентных факторов
- Количество случайных негативных примеров на каждого пользователя
- Количество итераций
- Параметры регуляризации

Решение команды

Тематическая модель: Latent Semantic Indexing (LSI)

- Документ = пользователь (все токены вакансий, просмотренных пользователем + токены из CV) или вакансия (все токены вакансии).
- Строим TFIDF на токенах пользователей.
- Обучаем Latent Semantic Indexing модель (по сути SVD).
- Сходство пользователя и вакансии аналогично сходству двух документов.

Решение команды

Финальный набор моделей

SIM_jac	Item-based jaccard similarity
SIM_click	Item-based jaccard similarity on clicks
SIM_pearson	Item-based pearson similarity
SIM_imp	Item-based jaccard similarity on impressions
FR_f100_i25	Factorization, n_factors=100, iter=25
FR_f400_i70	Factorization, n_factors=400, iter=70
FR_f400_i50_no_side	Factorization, no side data
FR_imp	Factorization on impressions
TM	LSI topic model

Базовые модели:

FR_0	SIM_0	impression_prev	local score
1	0	0	76995
0	1	0	69622
0	0	1	104495
1	1	1	132505

- SIM_0 – item-based jaccard similarity
- FR_0 – factorization, 400 factors
- impression_prev – был ли impression этой вакансии пользователю

«Нулевая» версия:

FR_0	SIM_0	$impression\_prev$	$local\ score$
1	2	1	134285

Первая версия:

FR_0	SIM_0	$FR_0^{8.0} * SIM_0$	$impression\_prev$	$local\ score$
1	13	8	1	138073

Вторая версия:

FR_1	SIM_1	$FR_1^{8.0} * SIM_1$	$impression\_prev$	$local\ score$
1	13	8	1	140876

$FR_1 = FR\_f100\_i25$
$SIM_1 = SIM\_jac$

Третья версия:

FR_2	SIM_2	$FR_2^{8.0} * SIM_2$	$impression\_prev$	$local\ score$
1	13	8	1	143653

$FR_2 = 0.5*FR\_f100\_i25 + 0.5*FR\_f400\_i70$
$SIM_2 = 0.5*SIM\_jac + 0.5*SIM\_click$

local score (145841):

1.0*FR_3 + **15.0 *** ($FR_3^{8.0}$ * SIM_3) + **13.0 ***SIM_3 + **1.0 ***$impression\_prev$ - **0.5 ***$SIM\_pearson$ - **0.3 ***$FR\_f400\_i50\_no\_side$ + **0.5 *** ($FR\_imp^{2.0} * SIM\_imp$), где

$FR_3 = 0.5*FR\_f100\_i25 + 0.5*FR\_f400\_i70$

$SIM_3 = SIM\_click$

local score (146569):

1.0*FR_4 + **15.0 *** ($FR_4^{8.0}$ * SIM_4) + **13.0 ***SIM_4 + **1.0 ***$impression\_prev$ - **0.4 ***$SIM\_pearson$ - **0.3 ***$FR\_f400\_i50\_no\_side$ + **0.5 *** ($FR\_imp^{2.0} * SIM\_imp$) + **0.2 ***TM, где

$FR_4 = 0.5*FR\_f100\_i25 + 0.5*FR\_f400\_i70$

$SIM_4 = 0.4*SIM\_jac + 0.6*SIM\_t = 1$

Решение команды

Техническое подробности

- 1 сервер 56 ядер 256Gb
- Полное обучение + скоринг = 16 часов
- Библиотеки: graphlab, gensim

Leaderboard

История сабмитов

Score	Rank	Labbel	Time
554655	9	Topic model 100 factors	06/27/16
548366	8	Top 150 candidates from every model	06/25/16
543284	8	8 model set: 4FR + 4SIM + TM	06/24/16
537157	9	Topic model	06/23/16
530599	10	3 models set: FR + 2SIM + tuned coef.	06/23/16
497136	15	3 models set: FR + 2SIM	06/22/16
496241	1	FR with side data	03/20/16
397604	1	Impression prev feature	03/11/16
132790	1	Simple item-similarity recommender	03/10/16

Leaderboard

Финальный рейтинг

Rank	Team	Leaderboard Score	Full Score
1	YunOS-OneSearch	681707.38	2052185.54
2	mim-solutions	675985.03	2035964.16
3	DaveXster	665592.06	2005263.73
4	PumpkinPie	622408.55	1866477.77
5	milk tea	613125.21	1846420.12
6	mdr_rec	605048.58	1823472.31
7	**Avito**	**554654.72**	**1677898.52**
8	recometric	556133.18	1677233.84
9	nodalpoints	555483.39	1671812.08
10	lucky_dog	542213.51	1632828.82
21	XING_TELECOM	461000.32	1397030.74

Спасибо!

Приглашаем принять участие:

- Конкурс Avito-2016: Распознавание категории объявления (призовой фонд конкурса - 500 000 рублей)
- Data science meetup в Avito – 13 августа 2016г.