|  |  |  |  |
| --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  | **Софийски университет „Св. Кл. Охридски”**  Факултет по математика и информатика |  |   **Курсов Проект**  на тема: |
|
| A Bottom-Up Approach to Job Recommendation System  (RecSys Competition 2016) |

Студенти: **Стоян Ефтимов Ефтимов - 25741**

**Хризантема Стефанова Станчева - 25652**

Курс: **Първи**

Специалност: **Изкуствен интелект**

Преподавател: **доц. Милен Чечев**

Дата: **София, 29.06.2018 г.**

**Съдържание**

[**Въведение**](#_30j0zll) **4**

[**Данни**](#_yyif02i5uaml) **4**

[Users](#_2jfr4j8tun6m) 5

[Items](#_h8qpogqcoby0) 5

[Interactions](#_1nm2o5g5uvmk) 5

[Training and test set](#_33pjg1psdlxi) 5

[**Реализации(Методологии)**](#_vhz5x6qt2xxv) **6**

[3.1 Използване на данните Interactions](#_6tctwxqnpr0f) 6

[3.2 Collaborative Filtering](#_7ro44dhhgbeo) 6

[3.2.1 K-Means Clustering:](#_b9xp8tgkt77n) 7

[3.2.2 Cosine Similarity:](#_xcs884u7ctiv) 7

[3.3 Scoring](#_9iaoc8fppizl) 8

[**Експерименти**](#_kp5ggkucpwyt) **9**

[**Резултати/Анализ**](#_is7x79s48wfg) **9**

[**Възможни подобрения**](#_5ec4nppzft5e) **9**

[**Библиография**](#_1t3h5sf) **10**

# Въведение

Повечето уебсайтове днес се стремят да предоставят качествени препоръки, за да увеличат и задържат своите клиенти. С цел да се осигури по-добро обслужване и да се подобри ефективността е от решаващо значение доставчиците на онлайн услуги да предскажат намерението на потребителите и да им предоставят подходящо съдържание от огромния обем информация. Този процес обикновено се нарича система за препоръки, където профилирането на потребителите, съдържанието на елементите, както и историческите данни за дейността на потребителя, се усвояват, за да се моделират намеренията или интересите на потребителите.

В тази статия са представени подходите за разработване на система за препоръки за работа за уеб сайт за социални контакти в кариерата - XING. Възприетият подход е отдолу нагоре: започва се със задълбочено проучване на данните и постепенно се изграждат по-малките компоненти на системата. Разгледани са традиционните подходи на системите за препоръки като Collaborative Filtering. Моделът, който постига най-добри резултати е базиран на специално дефинирана скоринг функция. Тази работа се основава на предизвикателството, организирано от конференцията ACM RecSys 2016 и по конкретно на решението на Sonu K. Mishra,Manoj Reddy [1], които са завършили на 20то място в официалното класиране.

# Данни

Съществуват четири масива от данни, предоставени от конкурса в допълнение към списък с 150K целеви потребители. За всеки целеви потребител, трябва да се генерират 30 препоръки.

## Users

Този набор от данни съдържа информация, получена от потребителските профили. Предоставените полета са: потребителски идентификатор, роли на работа, кариера ниво, дисциплина, индустрия, различни области за професионален опит, регион, държава, образование и т.н. В този файл има 40 хиляди записа.

## Items

Този набор от данни съдържа информация за обявите, публикувани на уебсайта. Предоставените полета са: идентификационен номер на позицията, роли на работа, дисциплина, индустрия, вид заетост, местоположение и т.н. В този файл има 168 хиляди записа.

## Interactions

Файлът за взаимодействията съдържа информация за действията, извършвани от потребителите върху различни обяви. Предоставените полета са: потребителски идентификатор, идентификационен номер на позиция, тип на взаимодействие и времева маркировка. В този файл има 547 хиляди записа. Типът на взаимодействие се състои от следните стойности:

1. потребителят е кликнал върху елемента

2. потребителят е отбелязал елемента

3. потребителят е кликнал върху бутона за формуляр за кандидатстване

4. потребителят е изтрил препоръката

В оригиналната формулировка на задачата съществува и допълнителен сет с данни Impressions, който ние не използваме.

## Training and test set

Тъй като няма ясно дефиниран набор от тренировъчни данни трябва да го създадем. Тъй като задачата е да се предвидят взаимодействията на потребителите в седмицата след края на наличните данни, моделът може да бъде обучен върху всички данни, с изключение на последната седмица, а данните от последната седмица могат да бъдат използвани за оценяване на модела.

# Реализации(Методологии)

## 3.1 Използване на данните Interactions

Този подход е съсредоточен изцяло върху набора от данни за Interactions, за да идентифицираме елементите, с които даден потребител е взаимодействал в миналото. След това препоръката за всеки тестов потребител става въз основа на елементите, с които той е взаимодействал положително в миналото. Идеята е, че ако потребителят е взаимодействал положително с конкретна позиция за работа, тогава е по-вероятно да взаимодейства отново позитивно с нея. Елементите за позиция за работа със стойност на взаимодействие 3 са на върха, следвани от 2 и 1. Елементите с тип на взаимодействие 4 се игнорират, тъй като сигнализират, че потребителят не харесва дадения елемент, и не искаме да показваме елементи, които потребителят е изтрил. Проблем, който възниква е, че данните за взаимодействията са много редки и не включват информация за всички тестови потребители. Друго предизвикателство е, че не всеки потребител е взаимодействал с най-малко 30 елемента, което може да доведе до препоръчване на по-малко от 30 елемента за всеки потребител. Това може да доведе до по-нисък резултат, тъй като уеб сайтът препоръчва до 30 елемента за всеки потребител.

## 3.2 Collaborative Filtering

Collaborative filtering е традиционен подход за автоматични прогнози (филтриране) на интересите на потребителя. Съществуват два възможни сценария:

1. Прилика между потребители (User-User Similarity): Да предположим, че има двама човека търсещи работа, които имат подобна образователна квалификация, подобен опит и живеят в същия регион. Ако един от тях, кандидатства за работа X, тогава е много вероятно и другият да кандидатства за същата работа.

2. Прилика между позиции (Item-Item Similarity): От друга страна, нека дадена длъжност Y е подобна на X по отношение на местонахождение, опит и основни изисквания. Ако потребител кандидатства за работа X, тогава е много вероятно той да кандидатства и за работа Y.

Съществуват две предизвикателства при употребата на този подход.

1. Sparsity: Тъй като наборът от данни за взаимодействие е много рядък, има по-малко факти от изводите, които трябва да се направят.

2. Определение за сходство: Трябва да има конкретно определение на приликата. Сходството не трябва да бъде изчислително скъпо предвид огромното количество данни, с които се занимаваме.

За да се определи дали двама потребители или два елемента са подобни са разгледани два подхода.

### 3.2.1 K-Means Clustering:

Изпълняваме алгоритъма K-Means за клъстеризиране на потребителски данни и данните за позиции за работа. Потребителите (или позициите), принадлежащи към същия клъстер, се считат за сходни, а тези в различните клъстери се считат за несъответстващи. Недостатъкът на този подход е, че качеството зависи от броя на клъстерите, който трябва да бъде избран ръчно.

### 3.2.2 Cosine Similarity:

Сходството S на два елемента се определя количествено от произведението на на съответните им вектори, нормализирани по дължината на векторите, както е показано в уравнението, където *Ux* и *Iy* представляват характеристичните вектори на потребител *x* и елемент *y*.

Първо се разделят потребителите и елементите в клъстери. При генериране на препоръки за потребител се разглеждат само потребителите, принадлежащи към един и същ клъстер, и се пресмята колко те са сходни според косинусовата метрика. Ако двама потребители са в един и същ клъстер, тяхното сходство се изчислява като косинусовата прилика на съответните вектори на елементите. От друга страна, се предполага, че потребителите, принадлежащи към различни клъстери, имат сходство 0. По този начин броят на изчисленията е значително намален.

Ако искаме да оценим как потребител u ще взаимодейства с елемент i. Взаимодействието на потребител u с точка i е средната стойност на взаимодействията на потребителите от същия клъстер с елемент i.

Другата възможност е да пресметнем средната стойност на взаимодействията на потребителя u с другите елементи от клъстера на елемент i.

## 3.3 Scoring

Този подход се базира на данните за потребителите и позициите за работа, използвайки полетата от тези набори от данни. Идеята е, че за генериране на препоръки за даден потребител, можем да пресметнем точки за всеки елемент(позиция за работа) и след това да класираме тези елементи според резултата. Резултатът се състои от положителните взаимодействия в миналото ако има такива. Други компоненти на резултата са припокриване на позициите за работа, нивото на кариерата, индустрията и дисциплината на потребителите. Всички тези компоненти се претеглят по различен начин за да се сумират до общия брой точки. Функцията за оценяване изглежда така:

*Score* = *w1*(Interaction score) + *w2*(no. of overlaps in user job-roles and item-titles) + *w3* (I[career level match]) + *w4* (I[discipline IDs match]) + *w5*(I[industry IDs match] + w6 (I[regions match])

Въз основа на естеството на данните теглата са определени евристично и са съответно:

*w1 = 100, w2 = 10, w3 = 12, w4 = 10, w5 = 5, w6 = 2.*

Това дава сравнително добри резултати. Алтернативно решение е теглата да бъдат научени с линейна регресия.

# Експерименти

Тъй като данните са достъпни само за участници регистрирани в състезанието, експериментите се осъществяват посредством системата предаване на решения. Поради липсата на официално предоставени тестови данни трудно бихме могли да оценим ефективността на проведените експерименти.

# Резултати

Първоначалните модели, които се занимават най-вече само с експлицитната информация, дадена в масивите от данни за взаимодействия, се представят сравнително добре, като се има предвид тяхната простота. Въпреки че подходите за Collaborative Filtering са често срещани в системите за препоръки, те не се справят много добре. Причините са високите размерности на наборите от данни и sparsity-то им. По-напредничавите модели представляват силни препоръчващи системи. Евристичният механизъм за оценяване е "лек" и дава добри резултати.

# Възможни подобрения

Можем да бъде направен анализ на поведението на потребителите, като се използва времевата информация от данните за взаимодействия. Това включва анализ на сесията на потребителя. На базата на няколко взаимодействия, можем да се предскажат следващите действия на потребител в текущата сесия.

# Библиография

[1] <https://www.researchgate.net/profile/Sonu_Mishra2/publication/308741629_A_bottom-up_approach_to_job_recommendation_system/links/5a83d44745851504fb3a7e24/A-bottom-up-approach-to-job-recommendation-system.pdf>

[2] <https://github.com/mishrasonu1993/JobRecSys>

[3] <https://github.com/marcowplm/recsys-2016>

[4] <http://2016.recsyschallenge.com/>