RecSys2013: Yelp Business Rating Prediction

Препоръчващи системи 2016

Изготвили:

Теофана Гаджева Ф.Н. 24960

Валентин Змийчаров Ф.Н. 24952

1. Цел на разработваната система

Заданието за имплементираната система е от състезанието RecSys от 2013 година. В kaggle.com е създадено състезание (https://www.kaggle.com/c/yelp-recsys-2013), спрямо което са документираните резултати. Проблемът, който трябва да се реши е да се предскаже рейтингът, който даден потребител би дал за даден бизнес.

1.1 Тренировъчни данни

Тренировъчните данни са в 4 файла в json формат както следва:

- **11 475 бизнеса** със следните атрибути:
  - id
  - адрес
  - отворен ли е
  - категории
  - град
  - брой ревюта
  - име
  - координати
  - звезди
- **8 282 checkin-а** със следните атрибути:
  - id на бизнес
  - брой отбелязвания, разпределени по дни от седмицата и часове
- **43 873 потребители** със следните атрибути:
  - id
  - име
  - брой ревюта
  - средно дадени звезди (рейтинг)
  - брой дадени гласове: funny/useful/cool (забавен/полезен/готин)
- **229 907 ревюта** със следните атрибути:
  - id
  - id на потребител
  - id на бизнес
  - дата
  - текст
  - звезди
  - гласове: funny/useful/cool (забавен/полезен/готин)

1.2 Тестови данни

Подобно на тренировъчните данни, тестовите също са в 4 файла в json формат, но липсват някои от атрибутите. В тези данни са включени бизнеси и

потребители, които не са включени в тренировъчните данни. Те са нови и затова за тях липсват информация за звезди. Данните са както следва:

- **2 797 бизнеса** със следните атрибути:
  - id
  - адрес
  - отворен ли е
  - категории
  - град
  - име
  - координати
- **1 796 checkin-a** със следните атрибути:
  - id на бизнес
  - брой отбелязвания, разпределени по дни от седмицата и часове
- **9 522 потребители** със следните атрибути:
  - id
  - име
  - брой ревюта
- **36 404 ревюта** със следните атрибути:
  - id
  - id на потребител
  - id на бизнес

1.3 Проблемът

Задачата е да се определят дадените звезди за всяко от ревютата в тестовите данни. Валидни са стойности между 0 и 5.

2. Възможни подходи

2.1 User-based collaborative filtering

При този подход се търси сходност между потребителите. Например ако се търси каква оценка е дал потребител **X** за бизнес **A**, се намират потребителите, които са най-сходни до **X** по отношение на оценяваните от тях бизнеси и са оценили бизнеса **A**. На база на техните оценки се определя най-вероятната оценка, която **A** би дал. Изпитва проблеми, когато потребител **A** не е оценявал нищо до този момент или бизнесът **X** не е бил оценяван от други потребители. Използва се имплементацията на **apache mahout**. *[1] [2]* Зададеният праг (threshold) е 0.5.

2.2 Item-based collaborative filtering

Подобно на предходния подход, при item-based collaborative filtering се търси сходност, но този път между бизнесите. Чрез изграждане на матрица на оценките за различните бизнеси се намират зависимости между отделните бизнеси. След това се намира бизнес, който е оценен от дадения потребителят **X** и има зависимост с бизнеса **A**. На база на това се определя оценката на **X** за **A**. Изпитва същите проблеми както при user-based collaborative filtering. Използва се имплементацията на **apache mahout**. *[1] [2]*

2.3 SVD

При този подход се образува матрица *Потребители X Бизнеси*. След различни преобразувания *[3]* се изчисляват липсващите стойности. Зададени са следните параметри: 30 свойства (feature), 0.065 ламбда, 100 итерации.

2.4 SVD++

SVD++ лежи на основата на SVD, но го надгражда. Разликите между двата алгоритъма са малки *[4]*. Зададени са следните параметри: 30 свойства (feature), 100 итерации.

2.5 Хибридна препоръчваща система

В желанието си да подобрим резултатите от изброените алгоритми решихме да добавим още параметри, чрез които да изчислим оценката за дадено ревю. След анализ на данните се спряхме на 3 параметъра *[5]*:

1. **Дали даден бизнес е отворен в момента (w1)** - Изчисляваме средната оценка за всички отворени и затворени бизнеси. При изчисляване на предполагаем рейтинг задаваме този критерий с различно тегло - ако бизнесът е отворен му задаваме оценка средна за всички отворени и аналогично ако е затворен - средната за всички затворени
2. **Брой ревюта на даден бизнес (w2)** - Преценихме, че популярността на един бизнес (колко пъти е ревюиран) може да има пряко отношение към това как даден потребител би го оценил. Добавяме този критерий с различно тегло и отново го приравняваме към средните стойности от тренировъчните ревюта.
3. **Брой отбелязвания (*checkin*) за даден бизнес (w3)** - Подобно на предходния критерий - променяме предполагаемата оценка за даден бизнес в зависимост от броя отбелязвания за него.

2.6 Проблеми

По време на имплементацията срещнахме проблеми, предимно свързани с липсата на данни. За много от тестовите ревюта потребителите и/или бизнесите не бяха споменавани в тренировъчните данни. Освен това за тренировъчните ревюта липсва дата на извършване на оценката. Няма информация потребителите от кой град са и съответно няма как да се направи връзка дали даден бизнес е от същия град като потребителя.

1. Когато се налагаше да оценяваме ревюта, при които потребителят и бизнесът нямат налични оценки, взимаме сума от трите общи параметъра за цялата система (*Описани в секция 2.5*): 0.33 * w1 + 0.33 * w2 + 0.33 * w3.
2. Когато се налагаше да оценяваме ревюта, при които само потребителят няма оценки, избираме средната оценка за бизнеса до момента.
3. Когато се налагаше да оценяваме ревюта, при които само бизнесът няма оценки, изибраме средната оценка, давана от потребителя до момента.

3. Експерименти, избрано решение и постигнати резултати

Първо изпробвахме кой от четирите алгоритъма ще даде най-добър резултат без добавяне на други параметри. Резултатите, които системата генерира са чрез RMSE (root-mean-square error):

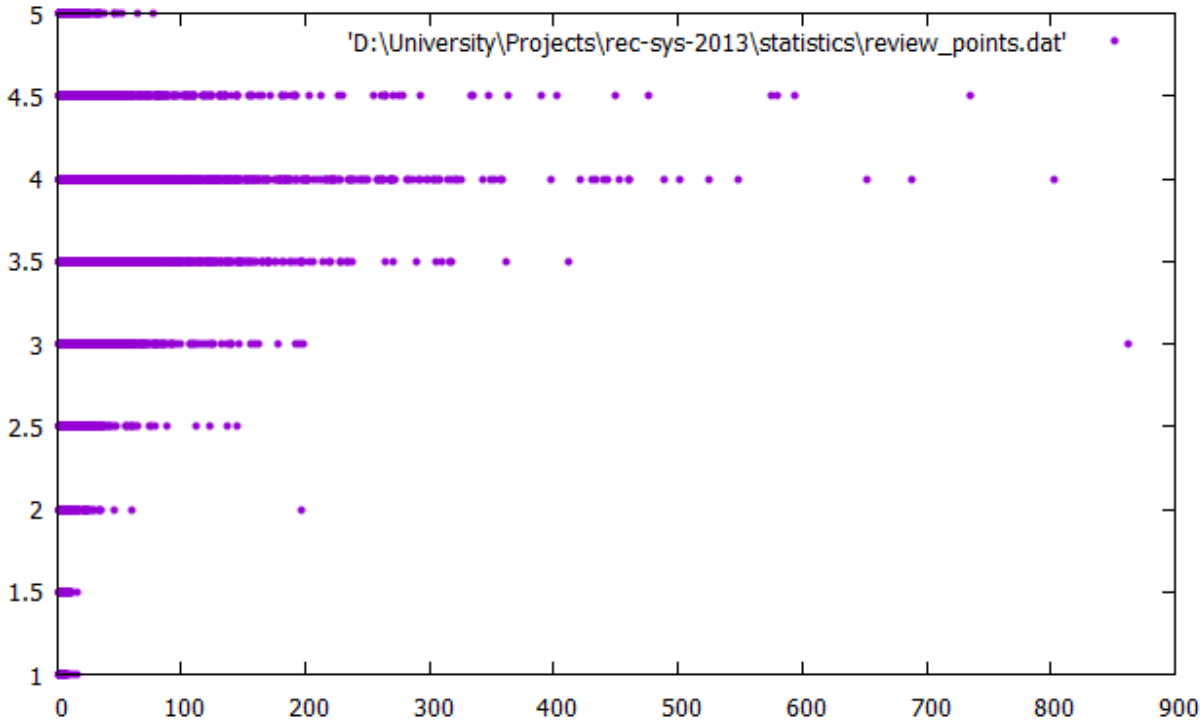
Класификатор	RMSE	Място
User-based	**1.30467**	**132**
Item-based	1.31005	134
SVD	1.36046	150
SVD++	1.42361	154

Таблица 1: Резултати от различните алгоритми

Вижда се, че User-based (**w**) подхода генерира най-добри резултати, затова се спряхме на него. След това го комбинирахме с посочените в *Секция 2.5* параметри и експериментирахме с различни тегла. Анализирахме средните стойности за трите параметъра:

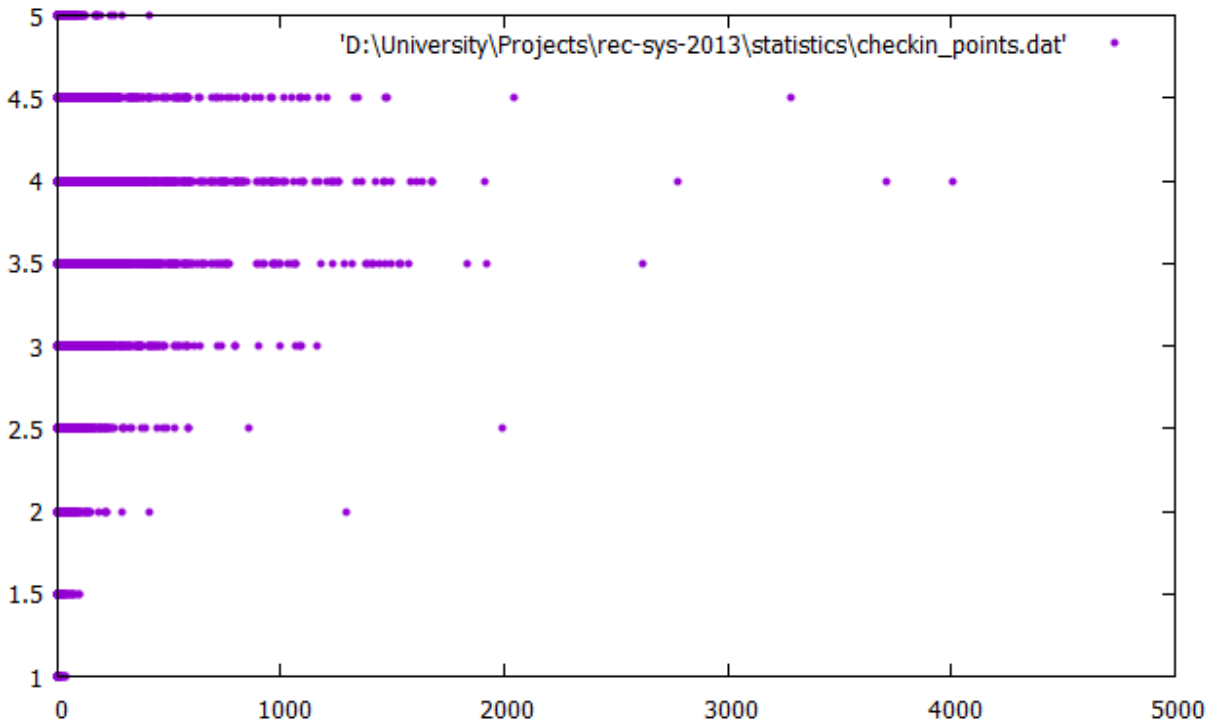
- **w1** - Средният рейтинг за отворените бизнеси е 3.70, за затворените - 3.46.

- **w2** - Стойностите за рейтингите в зависимост от броя ревюта са представени на **Фигура 1**. В зависимост от това разделихме бизнесите на 4 категории:
    - **0-100 ревюта** - средна оценка *3.67*
    - **100-200 ревюта** - средна оценка *3.82*
    - **200-300 ревюта** - средна оценка *3.99*
    - **Повече от 300 ревюта** - средна оценка *4.04*



Фигура 1: Зависимост между брой ревюта и оценка

- **w3** - Стойностите за рейтингите в зависимост от броя отбелязвания (checkin) са представени на **Фигура 2**. В зависимост от това разделихме бизнесите на 4 категории:
    - **0-500 отбелязвания** - средна оценка *3.67*
    - **500-1000 отбелязвания** - средна оценка *3.82*
    - **1000-1500 отбелязвания** - средна оценка *3.85*
    - **Повече от 1500 отбелязвания** - средна оценка *3.76*



Фигура 2: Зависимост между брой отбелязвания и оценка

След проведени експерименти с различни тегла за стойностите на w, w1, w2, w3 резултатите са показани в **Таблица 2**:

Избор на тегла	RMSE	Място
0.88*w + 0.04*w1 + 0.04*w2 + 0.04*w3	1.29576	127
0.76*w + 0.08*w1 + 0.08*w2 + 0.08*w3	1.27863	125
0.55*w + 0.15*w1 + 0.15*w2 + 0.15*w3	**1.27375**	**122**
0.40*w + 0.20*w1 + 0.20*w2 + 0.20*w3	1.30951	133

Таблица 2: Резултати при различен избор на тегла за допълнителни параметри

Най-добри резултати се постигнаха при взимане на основния класификатор с тегло 0.55 и 0.15 за останалите 3 параметъра.

4. Бъдещи подобрения

С цел разширяване на проекта и постигане на по-добри резултати може да се добавят нови параметри при изграждане на оценката. Някои от атрибутите за потребители и бизнеси не влизат в употреба в текущата имплементация.

Друг напредък би могъл да се осъществи при изчисляване на данните за ревюта с липсващи данни. Експерименти с взимане на различни осреднени стойности биха могли да покажат по-добри резултати.

Проби свързани с промяна на параметрите при използваните алгоритми също биха могли да подобрят резултатите.

5. Литература

1. *John S. Breese, David Heckerman, and Carl Kadie, Empirical Analysis of Predictive Algorithms for Collaborative Filtering, 1998 Archived 19 October 2013 at the Wayback Machine.*
2. *Xiaoyuan Su, Taghi M. Khoshgoftaar, A survey of collaborative filtering techniques, Advances in Artificial Intelligence archive, 2009.*
3. *Sarwar, Badrul; Karypis, George; Konstan, Joseph A. & Riedl, John T. (2000). "Application of Dimensionality Reduction in Recommender System -- A Case Study" (PDF). University of Minnesota. Retrieved May 26, 2014.*
4. *Xavier Amatriain, (2015). What's the difference between SVD and SVD++?*
5. *F. Maxwell Harper, Funing Xu, Harmanpreet Kaur, Kyle Condiff, Shuo Chang, Loren Terveen, RecSys 2015, Putting Users in Control of their Recommendations*