Matematyczne modele wykorzystywane w systemach rekomendacji.

Anita Kudaj

4 listopada 2018

Spis treści

Rozdział 1

Wstęp

Rozdział 2

Preliminaria

Rozdział 3

Modele tworzenia rekomendacji

3.0.1 Filtrowanie kolaboratywne (Collaborative filtering)

Słysząc od innych osób dobre opinie na temat ostatnio wydanej książki znanego pisarza, możemy zdecydować się na jej przeczytanie. Podobnie, poznając stwierdzenia, że ta sama książka jest katastrofą nie będziemy chcieli tracić pieniędzy na jej zakup oraz czasu na oddanie się lekturze. Możemy również otrzymać dwie sprzeczne opinie na temat tej samej pozycji. W każdym z przypadków słuchamy, analizujemy wartość uzyskanych ocen i ostatecznie podejmujemy na ich podstawie najbardziej odpowiednią dla nas decyzję.

Idea jaką rozważamy pod hasłem „filtrowanie kolaboratywne" mówi, że jeżeli użytkownicy A i B wykazują podobieństwo oraz użytkownik A zaopiniuje pewien przedmiot, którego użytkownik B jeszcze nie ocenił, to prawdopodobnie opinia użytkownika B będzie podobna do opinii użytkownika A.
W rozważnej metodzie wyróżniamy dwa podstawowe typy:

- filtorwanie kolaboratywne oparte na użytkowniku (user-based)
- filtrowanie kolaboratywne oparte na elementach (item-based)

Mianownikiem wspólnym w obu przypadkach jest fakt, że oceny jednych użytkowników są podstawą do tworzenia rekomendacji dla innych.

Filtrowanie kolaboratywne oparte na użytkowniku

W typie filtrowania opartym na użytkowniku zakładamy, że osoby z podobnymi preferencjami w przeszłości będą podobnie wybierały w przyszłości.

W celu dokładniejszego zrozumienia tego typu filtrowania został przedstawiony poniższy przykład. Zakładamy, że tabela przedstawia oceny czytelników dla kilku wybranych książek oraz, że każda z zapytanych osób mogła wystawić ocenę z zakresu 1 -10. Istotne jest, że nie wszyscy zapytani wystawili ocenę dla każdej z książek:

Książka/Czytelnik	Anna	Maciej	Bartek	Ewa	Sandra	Kacper
Wladza Absolutna	6	3		6	4	
Drzewo Anioła		6	6	5	6	
Proxima	7	7	8	7	8	9
Bastion	8	10	10	7	6	8
Teatr Świata	9	6	6	6	6	
Dżoker	5	7	7	5	4	2

Kluczowym krokiem do zasugerowania książki osobie poszukującej nowej lektury jest znalezienie podobnych jej czytelników. Aby to zrobić najpierw, przy użyciu ocen jakie zostały wystawione dla konkretnych książek, należy policzyć podobieństwo między poszczególnymi użytkownikami. Następnie dla wszystkich użytkowników rozważone zostają książki, które nie są ocenione przez nich ale są ocenione przez innych. W ten sposób zostają obliczone noty jakie wybrana osoba mogłaby zaproponować dla książek, których nie czytała, a które chcemy jej przedstawić. Najczęstszymi podejściami stosowanymi w tej metodzie do obliczania szukanego podobieństwa są Metryka Euklidesowa i Współczynnik Korelacji Pearsona. W tym przykładzie zastosujemy pierwsze ze wspomnianych rozwiązań. Używając wzoru:

$$d_e(x, y) = \sqrt{\sum_{i=1}^{n} \mid x_i - y_i \mid^2}$$

obliczone zostają szukane odległości:

(tabela z podobieństwami)

Bazując na podobieństwach między poszczególnymi użytkownikami, przez obliczenie średniej ważonej, zostaje przewidziana ocena jaką Bartek zaproponuje dla książki „Władza Absolutna". W poniższym równaniu wartość podobieństwa między Bartkiem i innymi użytkownikami została pomnożona przez ocenę jaką dany użytkownik wystawił dla książki „Władza Absolutna". Następnie, w celu normalizacji, wynik został podzielony przez sumę wartości wszystkich podobieństw.

(obliczenie)

Ostatecznie, gdy znane są oceny dla wszystkich książek dokonana zastaje rekomendacja dla naszego użytkownika.

Filtrowanie kolaboratywne oparte na elementach

W przypadku filtrowania kolaboratywnego opartego na elementach wartości podobieństwa między użytkownikami zostaje zastąpiona przez wartości podobieństwa między elementami.

W tym przypadku podstawowe założenie mówi, że jeżeli użytkownik wybrał element A w przeszłości oraz element B jest podobna do A, to użytkownik będzie skłonny wybrać również element B.

Podobnie jak w przypadku opartym na użytkowniku również i w tym należy wykonać dwa kroki. W kroku pierwszym zostaje obliczone prawdopodobieństwo występujące między elementami. Następnie na podstawie ocen wydanych już przez użytkownika dla podobnych elementów przewidziana zostaje ocena dla elementu nieocenionego.

Najczęstszą miarą podobieństwa w tym przypadku jest podobieństwo kosinusów. Miara ta wyraża podobieństwo miedzy n-wymiarowymi wektorami poprzez kąt między nimi w przestrzeni wektorowej. W tym przypadku wektorami są kolumny przedmiotów.

$$sim(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| \star |\vec{b}|}$$

Warto dodać, że im mniejsza wartość kąta tym większe jest podobieństwo.

Chcąc zastosować ten rodzaj filtrowania kolaboratywnego dla poprzedniego przykładu i przewidzieć ocenę użytkownika dla pewnej wybranej książki należy zdefiniować wszystkie książki podobne do wybranej za pomocą podobieństwa kosinusów. W kolejnym kroku przez policzenie sumy ważonej ocen dla książek podobnych do wybranej, które zostały wcześniej ocenione przez czytelnika, zostaje przewidziana ocena dla książki. Poniższa tabela przedstawia rozważane podobieństwa:

(tabela powiązań między książkami)

Znając rozważane podobieństwa można przewidzieć ocenę książki „Drzewo Anioła" licząc sumę ważoną ocen nadanych przez Kacpra podobnym książkom. Podobieństwo

książki „Drzewo Anioła” i każdej z innych książek ocenianych przez Kacpra zostaje wymnożone przez oceny, które nadał konkretnym pozycjom. Następnie sumę wszystkich wyników i dzielimy przez sumę wszystkich podobieństw.

(równanie)

Wady i zalety filtrowania kolaboratywnego

Mając przed sobą dwa typy filtrowania kolaboratywnego możemy zadać pytanie o efektywność, czy precyzyjność tego rozwiązania.
Poniżej kilka wniosków i informacji, opartych na rozważaniach Christian Desrosiers i George Karypis [**44**][**file:///C:/Users/akuda/Downloads/NbrRSsurvey2011.pdf**] [**strona 7**] [**A comprehensive survey of neighborhood-based recommendation methods**][**Christian Desrosiers, George Karypis**] które pozwalają dostrzec zalety i wady tego rozwiązania.

Zalety:

- Opisywane podejście tworzenia rekomendacji jest intuicyjne i łatwe implementacji zarówno w przypadku metody opartej na użytkownikach jak i metody opartej na elementach.
- Metody filtrowania kolaboratywnego pozwalają ponadto na zwięzłe i intuicyjne wyjaśnienie obliczeń prognostycznych, które wykonujemy.
- W rozważanych metodach filtrowania nie są wykorzystywane informacje o zawartości produktów, czy informacje o profilu użytkownika. Kiedy wiec wzrośnie liczba ocen dla konkretnego produktu zmianie ulegnie jedynie wartość podobieństwa między elementami.

Z drugiej strony:

- Filtrowanie kolaboratywne jest kosztowne obliczeniowo, ponieważ wykorzystywane są tu informacje o użytkownikach, produktach oraz ocenach produktów przez użytkowników.
- Podejście to zawodzi, kiedy istnieje potrzeba stworzenia rekomendacji dla nowego użytkownika o którego ocenach nie ma informacji.
- Zarówno metoda oparta na użytkownikach, jak i metoda oparta na elementach jest mało wiarygodna kiedy zasób danych na którym bazujemy jest mały.

- Kiedy brakuje nam informacji o podobieństwach między użytkownikami lub elementami nie jesteśmy w stanie odnaleźć rekomendacji bazując tylko na informacji o ocenach.

Porównanie filtrowania kolaboratywnego opartego na użytkownikach i filtrowania kolaboratywnego opartego na elemtach:

Warty rozważenie jest również fakt wyboru między rekomendacją opartą na użytkowniku, a rekomendacją opartą na elementach. Według Christian Desrosiers, George Karypis [**44**][**file:///C:/Users/akuda/Downloads/NbrRSsurvey2011.pdf**] [**strona 7**] [**A comprehensive survey of neighborhood-based recommendation methods**][**Christian Desrosiers, George Karypis**] jest kilka obszarów, które należy rozważyć przed ostatecznym wyborem toku postępowania:

Precyzyjność: Metodę wybieramy w zależności od stosunku między użytkownikami a przedmiotami w rozważanych danych. Mianowicie, jeżeli rozważany zbiór zawiera dużą liczbę użytkowników i jednocześnie mała liczbę elementów preferowanym rozwiązaniem jest metoda oparta na elementach.
Sprawność: Złożoność rozważanych algorytmów zależy od stosunku między liczbą użytkowników, a liczbą elementów. Przyjmując O, U, E jako liczbę odpowiednio ocen, użytkowników i elementów zdefiniujmy $p = O/U$ i $q = O/E$. Wtedy też złożoność metody opartej na użytkownikach wyrażona zostaje przez p^2/E, a złożoność metody opartej na elementach przez q^2/U.
Stabilność: Rozważając ten aspekt przed wyborem metody należy rozważyć co rośnie szybciej – liczba użytkowników, czy liczba elementów. Jeżeli liczba elementów wydaje się bardziej statyczna wtedy też lepszym wyborem jest metoda oparta na elementach i odwrotnie.
Uzasadnienie: Pod tym względem lepszym wyborem będzie system rekomendacji oparty na elementach. W przypadku bowiem potrzeby wyjaśnienia naszej rekomendacji przedstawienie listy rozważanych elementów jest łatwiejsze niż przedstawienie listy użytkowników.
Serendipity: Patrząc pod kontem możliwości wyszukiwania zaskakujących rekomendacji lepszym wyborem byłby system oparty na użytkowniku. Pozwala on bowiem dojść do znacznie ciekawszych wniosków niż system oparty na elementach.

3.0.2 Systemy rekomendujące oparte na treści(Content-based recommender systems:

System rekomendacji filtrowania kolaboratywnego opiera się na informacji o ocenach przyznanych poszczególnym elementom w rozważanym zbiorze. Analizując przypadek osoby, która przyznała ocenę 5 dla wybranej książki można stwierdzić, że użytkownik ten miał na uwadze wiele czynników, na przykład: zawartą historię, gatunek, przedstawienie postaci, styl pisania autora.

Systemy rekomendacji oparte na treści ukierunkowane są na spersonalizowany poziom użytkownika, rozważają jego indywidualne preferencje oraz treść produktu. Opierają się na obliczaniu podobieństw. Wykorzystywane są tutaj metody uczenia maszynowego, takie jak klasyfikacja.

W typie rekomendacji opartym na treści stworzenie rekomendacji i wygenerowanie listy elementów, które mogę być odpowiednie użytkownikowi poprzedzone jest dwoma ważnymi krokami. Pierwszy krok to wygenerowanie informacji na temat naszego produktu, natomiast drugi to wygenerowanie profilu użytkownika oraz rozpoznanie cech produktu dla niego odpowiednich.

Generowanie profilu produktu opiera się na znalezieniu cech go opisujących. Najczęściej spotykaną formą opisu produktów jest przedstawienie ich w przestrzeni wektorowej, gdzie wiersze są nazwami produktów, a wartości poszczególnych cech reprezentowane są w kolumnach. Warto zauważyć, że krok generowania profilu opiera się głownie na wybraniu najbardziej istotnych w rekomendacji atrybutów oraz ocenie ich względnej ważności względem produktu.

Do generowania takiego profilu używany jest algorytm TFIDF (z ang. TF – term frequency, IDF - inverse document frequency), który pozwala policzyć względną ważność powiązania cechy z przedmiotem. Algorytm ten został dokładnie opisany w rozdziale [numer rozdziału].

Aby dokładnie przyjrzeć się metodom opartym na treści rozważmy, podobnie jak poprzednio, przykład oparty na książkach:

Książka	Gatunek
Władza Absolutna	Powieść kryminalna
Drzewo Anioła	Proza współczesna
Proxima	Powieść fantastycznonaukowa
Bastion	Horror
Teatr Świata	Literatura faktu
Dżoker	Powieść kryminalna

Tym razem stworzenie rekomendacji wymaga większej liczby faktów.
Przez wykorzystanie algorytmu TFIDF stworzymy profil każdej z książek. Pierwszym etapem algorytmu jest stworzenie macierzy „term frequency", której wypełnienie przedstawia odniesienie każdego z podanych terminów do każdej z książek. Załóżmy, że 1 oznacza iż książka reprezentuje cechy danego gatunku, natomiast 0 oznacza brak takich cech.
(tabela frequency)
Następnym krokiem jest stworzenie „ inverse dokument frequency" przez wykorzystanie poniższej formuły:

$$IDF = \log \frac{x}{y}$$

x - całkowita liczba dokumentów
y - częstotliwość dokumentu
W rozważanym przypadku x to liczba książek, natomiast y to całkowita liczba wystąpień „term frequency", uzyskana dla wszystkich dokumentów.

(tabela przeliczona)

Ostatnim krokiem jest stworzenie macierzy TFIDF przez zastosowanie poniższej formuły:

$$TF * IDF$$

(tabela po przemnożeniu)

Po wygenerowaniu profilu przedmiotu należy wygenerować profil użytkownika. W tym kroku stworzona zostaje macierz preferencji dopasowana do treści produktu. Definiując bowiem cechy użytkownika wspólne z treścią produktu zostaje wygenerowany efektywniejszy sposób porównania użytkowników i przedmiotów, a w efekcie możliwym staje się obliczenie podobieństwo między nimi.

Rozważając poniższy zbiór danych, który przedstawia informacje o czytelnikach i książkach. W poniższym zestawieniu 1 oznacza, że dana osoba przeczytała książkę, natomiast puste miejsce, że nie podjęła się lektury.

Książka/Czytelnik	Anna	Maciej	Bartek	Ewa	Sandra	Kacper
Władza Absolutna	1	1		1	1	
Drzewo Anioła		1	1	1	1	
Proxima	1	1	1	1	1	1
Bastion	1	1	1	1	1	1
Teatr Świata	1	1	1	1	1	
Dżoker	1	1	1	1	1	1

Następnym krokiem jest tworzenie profilu użytkownika, który zostanie użyty do porównania z profilem przedmiotu. Profil użytkownika powinien zawierać więc jego preferencje dotyczące cech danego przedmiotu, w tym przypadku będą to preferencje dotyczące gatunku książki. Iloczyn skalarny zbudowany na TFIDF i macierzy preferencji użytkowników przedstawi powinowactwo każdego z użytkowników do każdego z rozważanych gatunków książki.

(macierz powinowactwa)

Kolejnym krokiem po wygenerowaniu profilu przedmiotu i profilu użytkownika jest przedstawienie w jakim stopniu każdy z użytkowników będzie zainteresowany każdą z książek. Do tego wykorzystane zostanie, wcześniej już wspomniane, podobieństwo kosinusów.

(macierz zainteresowań)

Wady i zalety systemu rekomendującego opartego na treści

Podobnie jak wcześniej opisana metoda filtrowania tak i metoda oparta na treści posiada swoje wady i zalety.

Zalety systemu:

- Jest metodą ukierunkowaną na indywidualne rozważnie każdego użytkownika dzięki czemu jest lepsza niż filtrowanie kolaboratywne, w którym rekomendacja zostaje dokonana przez wzięcie pod uwagę ogół użytkowników.

- Dokładność systemu rekomendacji opartego na treści jest wyższa w porównaniu do podejścia kolaboratywnego działającego na ocenach.
- Systemy te są lepsze od filtrowania kolaboratywnego pod kątem tworzenia rekomendacji dla nowych użytkowników. Po otrzymaniu takiego użytkownika, którego preferencje nie są znane stworzenie rekomendacji nie jest w tym przypadku problemem. Istnieje bowiem możliwość szybkiego ustalenia cech elementu najbardziej dla niego odpowiednich.

Rozdział 4

Eksperymenty / cześć praktyczne

Rozdział 5

Podsumowanie

Korzystając z [1, Rozdział 3, akapit 4]

Bibliografia

[1] Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems: introduction and challenges. In *Recommender systems handbook*, pages 1–34. Springer, 2015.