Uniwersytet Warszawski
Wydział Matematyki, Informatyki i Mechaniki

Grzegorz Szpak
Nr albumu: 319400

Systemy rekomendacji bazujące na metodzie wspólnej filtracji

Praca licencjacka
na kierunku MATEMATYKA

Praca wykonana pod kierunkiem
prof. Andrzeja Skowrona / dr. Marcina Szczuki
Instytut Matematyki / Instytut Informatyki

Lipiec 2016

Oświadczenie kierującego pracą

Potwierdzam, że niniejsza praca została przygotowana pod moim kierunkiem i kwalifikuje się do przedstawienia jej w postępowaniu o nadanie tytułu zawodowego.

Data Podpis kierującego pracą

Oświadczenie autora (autorów) pracy

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data Podpis autora (autorów) pracy

Streszczenie

W pracy przedstawiono analizę systemów rekomendacyjnych opartych na metodzie wspólnej filtracji. Zaprezentowane zostały dwa podstawowe typy wspólnej filtracji - filtracja oparta o użytkowników oraz filtracja oparta o przedmioty. Autor skupia się na implementacji tej metody opartej na algorytmie k - najbliższych sąsiadów. Przedstawione zostają różne warianty poszczególnych składowych algorytmu (wybór funkcji podobieństwa, rozmiar sąsiedztwa, sposób ustalania oceny na podstawie ocen sąsiadów) i ich wpływ na jakość systemu.

Słowa kluczowe

eksploracja danych, systemy rekomendacyjne, wspólna filtracja, algorytm k - najbliższych sąsiadów

Dziedzina pracy (kody wg programu Socrates-Erasmus)

11.1 Matematyka

Klasyfikacja tematyczna

68T05 Learning and adaptive systems

Tytuł pracy w języku angielskim

Recommender systems based on collaborative filtering

Spis treści

Rozdział 1

Wprowadzenie

1.1. Systemy rekomendacyjne i ich znaczenie

Każdego dnia stawiani jesteśmy przed szeregiem wyborów. Którą książkę przeczytać? Jaki film obejrzeć? Którą stronę internetową odwiedzić? Jako że ilość informacji w otaczającym nas świecie rośnie dużo szybciej niż nasze możliwości poznawcze, konieczne stało się stworzenie technologii, która będzie w stanie wybrać informacje najbardziej dla nas użyteczne. Taką rolę we współczesnych aplikacjach pełnią systemy rekomendacyjne. Bazując na dotychczasowej aktywności użytkownika, systemy rekomendacyjne starają się przewidzieć jego ocenę wobec danego przedmiotu. Terminy "ocena" i "przedmiot" są oczywiście ogólne i w zależności od zastosowania mogą reprezentować różne zjawiska - od wystawienia rzeczywistej oceny filmowi, przez kupno przedmiotu w sklepie internetowym, po fakt dołączenia do grupy na portalu społecznościowym.

Stosowanie systemów rekomendacji przynosi korzyści zarówno dla klienta, jak i dla sprzedawcy (właściciela serwisu). Użytkownik od razu otrzymuje treść, którą jest zainteresowany - bez konieczności przeglądania serwisu. Korzyści sprzedawcy z kolei można by mnożyć: od lepszego zrozumienia potrzeb użytkownika, przez zwiększoną różnorodność sprzedawanych produktów, po wzrost lojalności klientów - wszystko to przekłada się na pomnożenie zysków. Z tego powodu systemy rekomendacji stosowane są przez największe firmy:

- Netflix[1] TODO
- Google News[2] TODO
- Amazon[3] TODO

1.2. Typy systemów rekomendacyjnych

W zależności od zastosowanych metod, systemy rekomendacyjne dzieli się na 4 kategorie:

1. **Systemy bazujące na wspólnej filtracji (ang. *collaborative filtering, CF*)**

   Podejście wspólnej filtracji polega na wyznaczaniu preferencji danego użytkownika bazując na preferencjach użytkowników o podobnym guście. Podobieństwo to jest wyznaczane na podstawie wcześniejszych zachowań użytkownika. Ze względu na prostotę

[1] https://www.netflix.com
[2] https://news.google.com
[3] https://www.amazon.com

implementacji i wysoką jakośćpredykcji, wspólne filtrowanie jest najbardziej popularną i najczęściej stosowaną technikę rekomendacji. Dodatkową zaletą tej metody jest jej elastyczność i możliwość stosowania w wielu dziedzinach - system oparty na wspólnej filtracji jest w stanie udzielać dokładnych rekomendacji bez potrzeby "zrozumienia" samego przedmiotu. Niepotrzebne są dodatkowe dane na temat użytkowników czy przedmiotów - wystarczy sama historia ocen.

Wady wspólnej filtracji są natomiast następujące:

- wprowadzanie nowego użytkownika lub przedmiotu do systemu (ang. *cold start problem*)
- ciężko jest uzyskać sensowne rekomendacje dla użytkownika o unikalnych preferencjach
- rzadkość danych - pojedynczy użytkownik ocenia zwykle niewielki ułamek dostępnych przedmiotów
- w niektórych wersjach *collaborative filtering* problemem może być też skalowalność.

W pozostałej części pracy autor skupi się na różnych wariantach *collaborative filtering* i spróbuje w oparciu o tę technikę zbudować jak najbardziej wydajny system rekomendacji.

2. **Systemy oparte na zawartości (ang. *content - based*))**

   To podejście koncentruje się na charakterystyce samych przedmiotów. Przedmiot reprezentowany jest przez zbiór cech. Cechami filmu mogą być na przykład: gatunek, rok produkcji, obsada aktorska, reżyser. Na podstawie wartości atrybutów wyznaczane jest podobieństwo między przedmiotami, następnie użytkownikowi prezentowane są przedmioty podobne do dotychczas polubionych. Przykładowo, jeśli użytkownik często oglądał filmy sensacyjne wyreżyserowane przez Martina Scorsese, system będzie rekomendował inne filmy tego reżysera.

   Metoda bazująca na zawartości także jest szeroko stosowana. Do jej zalet zaliczyć można:

   - niski poziom skomplikowania
   - w przeciwieństwie do wspólnej filtracji, rekomendacja dla danego użytkownika nie zależy od ocen innych użytkowników
   - transparentność - rekomendacje mogą być w łatwy sposób wytłumaczone na podstawie modelu (które cechy przedmiotu zadecydowały o jego wyborze)
   - możliwość rekomendacji przedmiotów nieocenionych jeszcze przez żadnego użytkownika.

   Tego typu podejście wymaga jednak często wiedzy eksperckiej przy tworzeniu atrybutów przedmiotów. Z racji swojej natury, słabo sprawdza się też przy rekomendacji przedmiotów o innej charakterystyce niż dotychczas ocenione przez użytkownika.

3. **Systemy bazujące na osobowości**

   To stosunkowo nowe podejście, zostało zaproponowane przez Ricardo Buettnera w [3]. Polega na stworzeniu profilu osobowościowego użytkownika na podstawie jego zachowań w sieciach społecznościowych i prezentowaniu treści dopasowanych do tego profilu.

4. **Systemy hybrydowe**

   Polegają na łączeniu wyżej opisanych technik. Systemy hybrydowe mogą być implementowane w wielu wersjach: przez wykorzystanie wspólnej filtracji oraz rekomendacji opartej na zawartości osobno, a następnie połączeniu wyników, przez wykorzystanie dodatkowej wiedzy o użytkownikach czy przedmiotach w metodzie collaborative filtering czy przez zunifikowanie różnych podejść w jeden model. Łączenie różnych podejść może często wyraźnie podnieść jakość rekomendacji. Wadą tego typu systemów jest zwykle wysoki poziom skomplikowania.

1.3. The Netflix Prize

Do znacznego rozwoju badań nad systemami rekomendacyjnymi przyczynił się konkurs The Netflix Prize, zorganizowany przez internetową wypożyczalnię filmów Netflix w 2006 roku. Konkurs polegał na przewidywaniu ocen, jakie użytkownicy wystawiliby filmom, na podstawie historycznych danych. Opublikowany zbiór treningowy zawierał 100,480,507 ocen wystawionych przez 480,189 użytkowników 17,770 filmom. Aby otrzymać główną nagrodę - $1,000,000 - należało poprawić wynik algorytmu firmy Netflix (o nazwie Cinematch) o co najmniej 10%. Zawody zostały rozstrzygnięte dopiero w roku 2009, kiedy algorytm zespołu BellKor's Pragmatic Chaos poprawił wynik algorytmu Cinematch o 10.6%. Zwycięski algorytm był algorytmem hybrydowym opartym między innymi na wspólnej filtracji, rozkładzie macierzy według wartości osobliwych oraz wzmacnianych drzewach decyzyjnych (ang. *gradient boosted decision trees*). [TODO bibliografia]

Rozdział 2

Formalizacja problemu rekomendacji

2.1. Podstawowe pojęcia

Dwoma głównymi pojęciami wyorzystywanymi w systemach rekomendacyjnych są *Użytkownik* oraz *przedmiot*. Użytkownicy mają preferencje wobec niektórych przedmiotów, nazywane *oceną*. Preferencje są często reprezentowane jako zbiór trójek (*użytkownik, przedmiot, ocena*). Jak wspomniano wcześniej, to, czym dokładnie jest *ocena*, zależy od charakteru serwisu. Ocena może być liczbą całkowitą pochodzącą z określonego przedziału (przykładowo: pięciopunktowa skala ocen w serwisie Netflix) lub mieć postać binarną (używane w serwisach społecznościowych "lubię" / "nie lubię"). Unarne oceny, jak na przykład informacja, czy użytkownik kupił dany przedmiot, są z kolei zwykle wykorzystywane w sklepach internetowych.

2.2. Oznaczenia

Niech $U = \{u_1, u_2, \cdots, u_k\}$ oznacza zbiór użytkowników, $I = \{i_1, i_2, \cdots, i_l\}$ - zbiór przedmiotów, a $R = \{r_1, r_2, \cdots, r_m\}$ - zbiór ocen.

Definicja 2.2.1 *Zbiór $S \subset U \times I \times R$ tworzy macierz o wymiarach $|U| \times |I|$, w której element r o współrzędnych (x, y) odpowiada trójce (u_x, i_y, r). Taka macierz nazywana będzie* ***macierzą użyteczności*** *lub* ***macierzą ocen****.*

Przykład macierzy użyteczności prezentuje tabela 2.1:

Tabela 2.1: Przykładowa macierz ocen

	Matrix	Szklana Pułapka	Titanic	Forrest Gump	Wall-E
Użytkownik A	5	1	?	2	2
Użytkownik B	1	5	2	5	5
Użytkownik C	2	?	3	5	4
Użytkownik D	4	3	5	3	?

Parom (u, i), dla których użytkownik u nie ocenił jeszcze przedmiotu i, odpowiadają puste pola w macierzy ocen. Te pola są zaznaczone symbolem "?".

Do każdego użytkownika u przypisany jest zbiór $I_u \subseteq I$ ocenionych przez niego przedmiotów. Analogicznie, $U_i \subseteq U$ oznaczać będzie zbiór użytkowników, którzy ocenili przedmiot i. Niech dalej M reprezentuje macierz użyteczności, gdzie r_{ui} jest oceną udzieloną przedmiotowi i przez użytkownika u. Niech r_u oznacza $|I|$ - wymiarowy wektor ocen użytkownika u (brak oceny reprezentowany jest przez 0), a $\bar{r}_u$ - średnią z ocen wystawionych przez tego użytkownika, tj.

$$\bar{r}_u = \frac{\sum_{i \in I_u} r_{ui}}{|I_u|}. \tag{2.1}$$

Podobnie c_i reprezentować będzie $|U|$ - wymiarowy wektor ocen wystawionych przedmiotowi i, zaś $\bar{c}_i$ - jego średnią notę.

2.3. Formalna definicja problemu

Przy ocenie jakości systemu rekomendacyjnego definiuje się zwykle ([1], [2] TODO) dwa problemy:

1. Zadanie rekomendacji - mając danego użytkownika u, wyznaczyć n przedmiotów najbardziej pasujących do gustu tego użytkownika.

2. Zadanie regresji - niech $S_{train} \subseteq U \times I \times R$ będzie zbiorem ocen dostępnych w systemie, a $S_{test} \subseteq U \times I$ - zbiorem par (*użytkownik, przedmiot*), dla których preferencje chcemy określić. Celem jest nauczenie funkcji $f : U \times I \to R$, która będzie minimalizować pewną funkcję błędu $e : (U \times I \to R) \times (U \times I) \to \mathbb{R}$.

   Dwie najczęściej używane funkcje straty to średni bezwzględny błąd (ang. *Mean Absolute Error*, MAE):

   $$MAE(f, S_{test}) = \frac{1}{S_{test}} \sum_{(u,i) \in S_{test}} |f(u,i) - r_{ui}| \tag{2.2}$$

   oraz pierwiastek błędu średniokwadratowego (ang. *Root Mean Squared Error*, RMSE):

   $$RMSE(f, S_{test}) = \sqrt{\frac{1}{S_{test}} \sum_{(u,i) \in S_{test}} (f(u,i) - r_{ui})^2} \tag{2.3}$$

   TODO normalized mean squared error

Zaletą obydwu funkcji jest zachowanie skali, z której pochodzą oceny - przykładowo w pięciopunktowej skali, reprezentowanej przez liczby całkowite z przedziału $[1, 5]$, MAE o wartości 0.7 oznacza, że algorytm średnio mylił się o 0.7 punktu.

Rozdział 3

Wspólna filtracja

Główna idea, na której bazuje metoda *collaborative filtering*, polega na tym, że ocena użytkownika u dla przedmiotu i będzie podobna do oceny innego użytkownika v, którego dotychczasowy schemat oceniania był zbliżony do ocen u. Analogicznie, preferencje użytkownika u wobec przedmiotów i oraz j będą podobne, jeśli pozostali użytkownicy oceniali oba przedmioty w analogiczny sposób.

Algorytmy wspólnej filtracji można podzielić na dwa typy:

- wspólna filtracja oparta na modelu - systemy oparte na tym podejściu wykorzystują dostępne oceny w celu wyuczenia modelu, który następnie będzie przewidywał nieznane wartości r_{ui}. W tego typu metodzie używa się na przykład rozkładu macierzy użyteczności według wartości osobliwych, maszyn wektorów wspierających, sieci Bayesowskie (TODO: bibliografia) oraz algorytm Expectation - Maximisation.

- wspólna filtracja oparta na sąsiedztwie - w tej metodzie ustala się funkcję podobieństwa, a następnie używa się jej do wyznaczenia zbioru *sąsiadów* - użytkowników bądź przedmiotów charakteryzujących się podobnym schematem oceniania. Ostateczna ocena wyznaczana jest tylko na podstawie ocen sąsiadów.

Szczegółową analizę filtracji opartej na modelu można znaleźć w TODO. W dalszej części pracy autor skupi się na podejściu bazującym na sąsiedztwie.

3.1. Filtracja w oparciu o użytkowników

Podejście bazujące na użytkownikach było pierwszą zautomatyzowaną metodą wspólnej filtracji. Zostało wprowadzone pierwszy raz w ramach *GroupLens Research Project*[1] [TODO bibliografia].

Załóżmy, że dana jest funkcja podobieństwa $s_{users} : U \times U \to \mathbb{R}$. Schemat wyznaczania oceny użytkownika u dla przedmiotu i jest następujący:

1 **foreach** $u' \in U \setminus \{u\}$ **do**
2 | Oblicz $s_{users}(u, u')$;
3 **end**
4 Wyznacz $N_i(u)$ - zbiór sąsiadów, którzy ocenili i;
5 Wyznacz predykcję oceny $\hat{r}_{ui}$ na podstawie $\{r_{vi} \mid v \in N_i(u)\}$;

[1] Projekt badawczy prowadzony na Uniwersytecie Minnesota

Najczęściej spotykaną metodą wyznaczania zbioru sąsiadów $N_i(u)$ (linia 4) jest wybranie k użytkowników o największej wartości funkcji podobieństwa, gdzie k jest parametrem algorytmu. Podstawową metodą wyznaczania predykcji (linia 5) jest z kolei wzięcie średniej oceny ze zbioru sąsiadów:

$$\hat{r}_{ui} = \frac{1}{|N_i(u)|} \sum_{v \in N_i(u) r_{vi}} r_{vi} \tag{3.1}$$

3.1.1. Wyznaczanie podobieństwa między użytkownikami

Wybór odpowiedniej funkcji podobieństwa s_{users} jest najważniejszą częścią algorytmu wspólnej filtracji. Odgrywa ona podwójną rolę:

- umożliwia wyznaczenie odpowiedniego zbioru $N_i(u)$
- w niektórych podejściach wartości funkcji s_{users} pozwalają na dokładniejsze wyznaczenie ostatecznego wyniku (więcej w rozdziale 5)

Poniżej znajduje się przegląd funkcji podobieństwa najczęściej opisywanych w literaturze.

Miary podobieństwa bazujące na korelacji

Inne miary podobieństwa

- **Współczynnik korelacji liniowej Pearsona**

  Współczynnik Pearsona określa poziom zależności liniowej między zmiennymi losowymi. Estymator współczynnika korelacji liniowej dla wektorów prób losowych $x, y \in \mathbb{R}^n$ jest zdefiniowany następująco:

  $$pearson_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{3.2}$$

  gdzie $\bar{x}$, $\bar{y}$ oznaczają wartości średnie z tych prób, tj $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$. Współczynnik korelacji pearsona przyjmuje wartości z przedziału $[-1, 1]$. Wartości bliskie 1 oznaczają silną zależność liniową między x a y, wartości bliskie zera - brak liniowej zależności, natomiast wartości bliskie -1 - ujemną liniową zależność.

  W kontekście systemów rekomendacyjnych, współczynnik korelacji określa się jak poniżej:

  $$pearson\_corr(u, v) = \frac{\sum_{i \in I_u \cap I_v}(r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in I_u \cap I_v}(r_{ui} - \bar{r}_u)^2}\sqrt{\sum_{i \in I_u \cap I_v}(r_{vi} - \bar{r}_v)}} \tag{3.3}$$

  Jest to najbardziej popularna funkcja do określania podobieństwa między użytkownikami. Wadą korelacji Pearsona jest zwracanie wysokiego podobieństwa dla par użytkowników, którzy ocenili niewielką liczbę takich samych przedmiotów. Jednym z możliwych sposobów obejścia tego problemu jest ustalenie progu T i przeskalowanie wyniku, przykładowo przez pomnożenie go przez $min(\frac{|I_u \cap I_v|}{T}, 1)$.

- **Zmodyfikowana korelacja Pearsona**

  Niektóre systemy interpretują medianę ze skali ocen r_m jako neutralne nastawienie użytkownika wobec przedmiotu. Ekstrand i in. zaproponowali w [3] (TODO bibliografia)

modyfikację współczynnika korelacji Pearsona, w której wektor ocen jest normalizowany przy użyciu owej mediany zamiast średniej:

$$median\_centered\_corr(u,v) = \frac{\sum_{i \in I_u \cap I_v}(r_{ui} - r_m)(r_{vi} - r_m)}{\sqrt{\sum_{i \in I_u \cap I_v}(r_{ui} - r_m)^2}\sqrt{\sum_{i \in I_u \cap I_v}(r_{vi} - r_m)}} \tag{3.4}$$

- **Współczynnik korelacji rang Spearmana**

  Podczas gdy współczynnik Pearsona wykorzystuje bezpośrednie wartości ocen r_{ui} do wyznaczenia korelacji, metoda Spearmana polega na przyznaniu tym ocenom rang. Wektor ocen r_u jest przekształcany w wektor rang w następujący sposób: najwyżej oceniony przedmiot otrzymuje rangę 1, kolejne przedmioty otrzymują wyższe rangi. Przedmiotom z tą samą oceną przyznaje się średnią rangę dla ich pozycji. Oznaczając przez k_{ui} rangę przyznaną przedmiotowi i w kontekście użytkownika u, otrzymujemy wzór:

$$spearman\_rank\_corr(u,v) = \frac{\sum_{i \in I_u \cap I_v}(k_{ui} - \bar{k}_u)(k_{vi} - \bar{k}_v)}{\sqrt{\sum_{i \in I_u \cap I_v}(k_{ui} - \bar{k}_u)^2}\sqrt{\sum_{i \in I_u \cap I_v}(k_{vi} - \bar{k}_v)}} \tag{3.5}$$

  Główną zaletą współczynnika Spearmana jest uniknięcie problemu normalizacji ocen (więcej o normalizacji w rozdziale 5). Obliczanie rang wymaga jednak dodatkowego kosztu.

Inne miary podobieństwa

- **Podobieństwo cosinusowe**

  Cosinus kąta między wektorami x, y w przestrzeni euklidesowej $(V, < \cdot >)$ wyraża się przez:

$$cos(x,y) = \frac{< x, y >}{\|x\| \, \|y\|} \tag{3.6}$$

  Dla wektorów z przestrzeni $\mathbb{R}^n$ ze standardowym iloczynem skalarnym, im większa wartość bezwzględna cosinusa, tym mniejszy kąt między nimi, co z kolei odpowiada intuicyjnie większemu "podobieństwu" między wektorami.

  Ta intuicja leży u podstaw wykorzystania miary cosinusowe jako funkcji podobieństwa, definując ją jak poniżej:

$$cosine(u,v) = \frac{r_u \cdot r_v}{\|r_u\|_2 \, \|r_v\|_2} \tag{3.7}$$

  Wadą wydawać się może reprezentowanie nieznanych ocen przez 0, co często upodabnia brak oceny do oceny negatywnej. Co więcej, używanie tej funkcji podobieństwa nie bierze pod uwagę statystycznych różnic między ocenami użytkowników (różnych średnich czy wariancji ocen). Mimo wymienionych wad, podobieństwo cosinusowe daje zaskakująco dobre wyniki w praktyce, co pokaże rozdział 4.

- **Podobieństwo Jaccarda**

  Współczynnik podobieństwa Jaccarda (inaczej: *indeks Jaccarda*) mierzy podobieństwo między dwoma zbiorami i jest zdefiniowany jako iloraz mocy przecięcia zbiorów i mocy sum tych zbiorów. W kontekście systemów rekomendacji wylicza się go zatem następująco:

$$jaccard(u,v) = \frac{|I_u \cap I_v|}{|I_u \cup I_v|} \tag{3.8}$$

Jako że indeks Jaccarda ignoruje wartości ocen, jego oczywistą wadą jest fakt utraty informacji przy skalach ocen innych unarna.

- **Rozszerzone podobieństwo Jaccarda**

  Jest to wersja podobieństwa Jaccarda dla wektorów z $\mathbb{R}^n$:

$$extended\_jaccard(x, y) = \frac{x \cdot y}{\|x\|_2^2 + \|y\|_2^2 - x \cdot y}, \tag{3.9}$$

  czyli używane jako miara podobieństwa użytkowników będzie miało postać:

$$extended\_jaccard(u, v) = \frac{r_u \cdot r_v}{\|r_u\|_2^2 + \|r_v\|_2^2 - r_u \cdot r_v}, \tag{3.10}$$

- **Podobieństwo euklidesowe**

  Odległość euklidesowa między użytkownikami określona jest w naturalny sposób przez:

$$d_2(u, v) = \sqrt{\sum_{i \in I_u \cap I_v} (r_{ui} - r_{vi})^2}. \tag{3.11}$$

  Są różne możliwości przekształcenia tej miary odległości w funkcję podobieństwa (idącą w $[0, 1]$, gdzie 1 oznacza największe podobieństwo). Do najczęściej używanych należą:

$$euclidean1(u, v) = \frac{1}{1 + d_2(u, v)} \tag{3.12}$$

  oraz

$$euclidean2(u, v) = e^{-d_2(u,v)} \tag{3.13}$$

  TODO: przewaga drugiej Te miary może być uogólniona poprzez użycie odległości Minkowskiego:

$$d_p(u, v) = \left(\sum_{i \in I_u \cap I_v} |r_{ui} - r_{vi}|^p \right)^{\frac{1}{p}} \tag{3.14}$$

- **Różnica średniokwadratowa (ang. *mean squared difference, MSD*)**

  Podobną, bazującą na odległości między ocenami funkcją podobieństwa jest różnica średniokwadratowa. Definiuje się ją jako odwrotność średniego kwadratu różnicy między ocenami użytkowników u i v:

$$MSD(u, v) = \frac{|I_u \cap I_v|}{\sum_{i \in I_u \cap I_v} (r_{ui} - r_{vi})^2} \tag{3.15}$$

  Różnica średniokwadratowa, podobnie jak podobieństwo euklidesowe, nie odzwierciedla ujemnej korelacji między użytkownikami. Posiadanie informacji o takiej korelacji może czasami zwiększyć jakość predykcji (TODO bibliografia).

3.2. Filtracja w oparciu o przedmioty

Obok wielu zalet, wspólna filtracja bazująca na użytkownikach ma jedną poważną wadę - brak skalowalności. Wyznaczanie zbioru $N_i(u)$ jest operacją o złożoności $\mathcal{O}(|U|)$ (lub gorszej - bezpośrednie wyliczanie wartości funkcji podobieństwa s wymaga czasu $\mathcal{O}(|U||I|)$). W dobie serwisów mających setki milionów użytkowników potrzebne było bardziej skalowalne podejście.

Naturalnym pomysłem wydawało się zastosowanie tej samej metody, zmodyfikowanej przez zastąpienie użytkowników przedmiotami. Zamiast szukania podobieństw między zachowaniami użytkowników, do wyznaczania rekomendacji wykorzystywane są podobieństwa między schematami oceniania przedmiotów. Jeśli dwa przedmioty i i j cieszyły się podobną opinią wśród użytkowników, można uznać je za podobne i wykorzystać ocenę, jaką użytkownik u wystawił przedmiotowi i do predykcji oceny dla przedmiotu j. Warto zauważyć, że ta wersja metody *collaborative filtering* jest podobna do rekomendacji opartej na zawartości, z tą różnicą, że do wyznaczania podobieństwa używa się preferencji użytkowników zamiast cech przedmiotów.

Zamiast funkcji s_{users} mamy zatem funkcję $s_{items} : I \times I \to \mathbb{R}$, na której podstawie wyznaczamy zbiór $N_u(i)$ przedmiotów najbardziej podobnych do i ocenionych przez u. Porównanie złożoności czasowej i obliczeniowej obu sposobów wspólnej filtracji przedstawia tabela 3.1.

Tabela 3.1: Porównanie złożoności pamięciowej i obliczeniowej obu wersji wspólnej filtracji

	Pamięć	Czas uczenia	Czas zapytania
Użytkownicy	$\mathcal{O}(\|U\|^2)$	$\mathcal{O}(\|U\|^2\|I\|)$	$\mathcal{O}(\|U\|)$
Przedmioty	$\mathcal{O}(\|I\|^2)$	$\mathcal{O}(\|I\|^2\|U\|)$	$\mathcal{O}(\|I\|)$

Zakładając, że $|U| \gg |I|$ (co jest prawdą w przypadku większości serwisów), łatwo zauważyć, że "przedmiotowe" podejście do wspólnej filtracji pozwala w znaczący sposób zmniejszyć złożoność zarówno czasową, jak i pamięciową algorytmu. Warto przy tym zaznaczyć, że macierz użyteczności jest często macierzą rzadką, co znacznie przyspiesza obliczanie wartości funkcji podobieństwa i co za tym idzie, czyni metodę wspólnej filtracji opartą na sąsiedztwie jeszcze bardziej skalowalną.

3.2.1. Wyznaczanie podobieństwa między przedmiotami

Nietrudno zauważyć, że wymienione w sekcji 3.1.1 funkcje podobieństwa mogą być z powodzeniem uogólnione i zastosowane przy wyznaczaniu podobieństwa między przedmiotami. Przykładowo, współczynnik korelacji liniowej Pearsona (równanie 3.3) dla przedmiotów można zdefiniować jako:

$$pearson\_corr(i,j) = \frac{\sum_{u \in U_i \cap U_j}(r_{ui} - \bar{c}_i)(r_{uj} - \bar{c}_j)}{\sqrt{\sum_{u \in U_i \cap U_j}(r_{ui} - \bar{c}_i)^2}\sqrt{\sum_{u \in U_i \cap U_v}(r_{uj} - \bar{c}_j)}} \tag{3.16}$$

Analogicznie zdefiniować można pozostałe miary.

Miarę podobieństwa przeznaczoną specjalnie dla przedmiotów zaproponowali Sarwar i in. w [4]. (TODO bibliografia) Bazuje ona na obserwacji, iż różnice w skalach ocen poszczególnych użytkowników często są znacznie wyraźniejsze niż w przypadku przedmiotów. Dlatego

przy obliczaniu podobieństwa między przedmiotami większy sens może mieć normalizacja oceny względem średniej oceny *użytkownika* zamiast *przedmiotu*. Formalnie ta miara, zwana **dopasowanym podobieństwem cosinusowym (ang. *Adjusted Cosine Similarity*)** określona jest jako:

$$adjusted\_cosine(i,j) = \frac{\sum_{u \in U_i \cap U_j}(r_{ui} - \bar{r}_u)(r_{uj} - \bar{r}_u)}{\sqrt{\sum_{u \in U_i \cap U_j}(r_{ui} - \bar{r}_u)^2}\sqrt{\sum_{u \in U_i \cap U_v}(r_{uj} - \bar{r}_u)}} \tag{3.17}$$

W niektórych przypadkach tak określona funkcja podobieństwa daje lepsze wyniki niż zwykle używane miary (TODO bibliografia).

Rozdział 4

Wpływ miary podobieństwa na jakość predykcji

W poniższym rozdziale przedstawione zostaną wyniki eksperymentu, mającego na celu ukazanie, jak wybór metody wspólnej filtracji oraz funkcji podobieństwa wpływa na jakość systemu rekomendacji.

4.1. Opis eksperymentu

4.1.1. Opis zbioru danych

W pracy wykorzystany został zbiór ocen filmów MovieLens[1]. Zbiór składa się ze 100,000 ocen udzielonych 1682 filmom przez 943 użytkowników, przy czym każdy użytkownik ocenił co najmniej 20 filmów. Skala ocen jest pięciopunktowa. Obok samych ocen, zbiór zawierał podstawowe informacje na temat filmów (m.in. tytuł, gatunek, datę wydania) oraz użytkowników (wiek, płeć, zawód, kod pocztowy). Dane zbierane były w ramach wspomnianego wcześniej projektu GroupLens przez 7 miesięcy.

4.1.2. Sposób przeprowadzenia eksperymentu

Jakość predykcji sprawdzana była przy użyciu pięciowarstwowej walidacji krzyżowej. Zbiór ocen był podzielony na 5 równolicznych, rozłącznych podzbiorów - skorzystano z gotowego podziału dostarczonego razem ze zbiorem danych przez zespół GroupLens. Następnie kolejno każdy z podzbiorów służył jako zbiór testowy, a cztery pozostałe podzbiory tworzyły zbiór treningowy. Ostateczny wynik jest średnią z wyników pięciu przeprowadzonych analiz.

Do oceny jakości predykcji użyto funkcji straty opisanych w rozdziale 2 - MAE oraz RMSE.

Zbiór sąsiadów wyznaczany był przez wzięcie 10 najbardziej podobnych użytkowników bądź przedmiotów. Ostateczna ocena wyliczana była jako średnia z ocen sąsiadów (zgodnie ze wzorem 3.1).

Aby uzyskać punkt odniesienia, dla obydwu podejść został również dodany wynik predyktora bazowego (*baseline*) - w przypadku użytkowników zwracał on średnią ocenę dla danego użytkownika ($r_{ui} = \bar{r}_u$), a w przypadku przedmiotów - średnią ocenę wystawioną przedmiotowi ($r_{ui} = \bar{c}_i$).

[1] https://movielens.org/

4.2. Wyniki

Wyniki dla metody bazującej na użytkownikach w zależności od postaci funkcji podobieństwa s_{users} prezentuje tabela 4.1.

Tabela 4.1: Wyniki dla podejścia opartego na użytkownikach

	MAE	RMSE
baseline	**0.828**	**1.031**
pearson_corr	0.805	1.012
median_centered_corr	0.787	0.999
spearman_rank_corr	0.807	1.014
cosine	0.806	1.014
jaccard	0.810	1.018
extended_jaccard	0.806	1.018
*euclidean*1		
*euclidean*2		
MSD		

Wyniki dla metody bazującej na przedmiotach w zależności od postaci funkcji podobieństwa s_{items} prezentuje tabela 4.2.

Tabela 4.2: Wyniki dla podejścia opartego na przedmiotach

	MAE	RMSE
baseline	**0.803**	**1.000**
pearson_corr	5	1
median_centered_corr	1	5
spearman_rank_corr	2	?
cosine	4	3
jaccard		
extended_jaccard		
*euclidean*1		
*euclidean*2		
MSD		
adjusted_cosine		

4.3. Wnioski

Chuj, nie chce mi się

Rozdział 5

Optymalizacja algorytmu opartego na sąsiedztwie

W poniższym rozdziale przeanalizowane zostaną szczegóły algorytmu najbliższych sąsiadów - wybór zbioru sąsiadów oraz wyliczanie wyniku na ich podstawie. Zaprezentowane zostaną wyniki dalszych eksperymentów pokazujących, jak obydwa aspekty wpływają na jakość predykcji.

5.1. Wyznaczanie zbioru sąsiadów

Rozmiar zbioru N najbliższych sąsiadów oraz sposób tworzenia tego zbioru na podstawie funkcji podobieństwa mogą mieć znaczny wpływ na jakość systemu rekomendacji. Przykładowo, pierwszy system rekomendacji GropuLens przyjmował $N_i(u) = U \setminus \{u\}$ - jednak branie pod uwagę użytkowników o niskiej korelacji zmniejszało jakość predykcji [TODO bibliografia].

Naturalne wydaje się więc wybieranie sąsiadów na podstawie wartości funkcji podobieństwa. Przy większej liczbie użytkowników niemożliwe staje się jednak trzymanie całej macierzy podobieństwa w pamięci. Również z powodu zbyt wysokiej złożoności czasowej przeglądanie całego zbioru U lub I przy każdym zapytaniu byłoby nieakcpetowalne. Z tego powodu wyznaczanie zbioru N jest często podzielone na dwa etapy: etap preprocessingu, w którym dla każdego użytkownika bądź przedmiotu wylicza się zbiór kandydatów na sąsiadów, oraz etap właściwego wyboru sąsiedztwa.

5.1.1. Etap preprocessingu

W tym etapie dla każdego użytkownika u lub przedmiotu i wyznaczany jest podzbiór $U' \subset U$ (odpowiednio: $I' \subset I|$), przy czym $|U'| \ll |U|$, który, trzymany w pamięci podręcznej, będzie później służył do wyznaczania zbioru sąsiadów. Warto przy tym tak dobrać rozmiar zbioru U', aby dla możliwie dużej liczby zapytań o ocenę r_{uj} rozmiar zbioru $U' \setminus U_j$ przekraczał oczekiwany rozmiar N. Do najbardziej popularnych metod preprocessingu należą:

- wybór k' sąsiadów z największą wartością funkcji podobieństwa
- wybór sąsiadów, dla których wartość funkcji podobieństwa przekracza pewien ustalony próg T
- jeśli używana jest funkcja podobieństwa bazująca na korelacji (przykłady takich funkcji opsiane były w sekcji 3.1.1), możliwe jest odfiltrowanie sąsiadów z ujemną korelacją względem ocen danego użytkownika bądź przedmiotu.

5.1.2. Wybór sąsiedztwa

Mając przygotowany zbiór U' bądź I' kandydatów na sąsiadów, można przystąpić do właściwego wyznaczenia zbioru N. Podobnie jak w etapie wstępnego przetwarzania, może to być wykonane dwojako:

- poprzez wybranie wszystkich sąsiadów, dla których wartość funkcji podobieństwa przekracza próg T, będący parametrem algorytmu - to podejście zapewnia, że w zbiorze N znajdą się tylko "dostatecznie podobni" użytkownicy bądź "dostatecznie podobne" przedmioty. Odpowiedni wybór wartości progowej T zależy jednak od wybranej funkcji podobieństwa i określenie jej może być niełatwym zadaniem - stąd rzadko stosuje się tę metodę w praktyce.
- poprzez wybranie k "najbardziej podobnych" sąsiadów, gdzie k jest parametrem algorytmu.

Ze względu na swoją prostotę i dobre wyniki w praktyce, wybór k najbliższych sąsiadów jest najczęściej stosowanym podejściem we współczesnych systemach rekomendacji. Odpowiedni dobór parametru k może wyraźnie poprawić jakość rekomendacji. Poniższe dwa wykresy przedstawiają zależność MAE oraz RMSE od wartości k dla obu metod wspólnej filtracji. W obu przypadkach wybrano miary, które dawały najlepsze wyniki dla $k = 10$ - w przypadku podejścia bazującego na użytkownikach były to TODO oraz TODO, a w przypadku podejścia opartego na przedmiotach - TODO oraz TODO.

Lathia i in. [TODO bibliografia] zaproponowali też metodę, w której k będzie dobierane dynamicznie w celu zmniejszenia błędu predykcji -

5.2. Wyznaczanie oceny na podstawie ocen sąsiadów

Ostatnim krokiem w algorytmie wspólnej filtracji jest ustalenie predykcji oceny $\hat{r}_{ui}$ na podstawie ocen sąsiadów ze zbioru N.

5.2.1. Średnia ważona

Podstawowe podejście - wyliczenie średniej arytmetycznej - ma tę wadę, że ignoruje wartości funkcji podobieństwa dla sąsiadów. Naturalnym pomysłem jest zatem zastąpienie średniej arytmetycznej średnią ważoną - dla użytkowników:

$$\hat{r}_{ui} = \frac{\sum_{v \in N_i(u)} s_{users}(u, v) r_{vi}}{\sum_{v \in N_i(u)} |s_{users}(u, v)|} \tag{5.1}$$

oraz dla przedmiotów:

$$\hat{r}_{ui} = \frac{\sum_{j \in N_u(i)} s_{items}(i, j) r_{uj}}{\sum_{j \in N_u(i)} |s_{items}(i, j)|} \tag{5.2}$$

Używanie średniej ważonej jest najczęściej stosowaną metodą wyliczania oceny [TODO bibliografia].

5.2.2. Normalizacja względem średniej (ang. *mean-centering*)

Jak wspominano we wcześniejszych rozdziałach, skala ocen u różnych użytkowników może być odmienna (niektórzy użytkownicy mogą mieć tendencję do dawania niższych ocen, inni z kolei mogą oceniać wyłącznie przedmioty, które im się podobały). Zamiast bezpośrednich

ocen r_{ui} proponuje się wtedy wykorzystanie ocen r'_{ui} znormalizowanych względem średniej oceny danego użytkownika: $r'_{ui} = r_{ui} - \bar{r}_u$. Intuicyjnie, znak r'_{ui} ma informować o preferencji użytkownika u wobec przedmiotu i - czy jest on nastawiony do i pozytywnie, czy negatywnie. Wzór na $\hat{r}_{ui}$ przyjmuje wtedy postać:

$$\hat{r}_{ui} = \bar{r}_u + \frac{\sum_{v \in N_i(u)} s_{users}(u,v)(r_{vi} - \bar{r}_v)}{\sum_{v \in N_i(u)} |s_{users}(u,v)|}. \tag{5.3}$$

Analogicznie, dla przedmiotów:

$$\hat{r}_{ui} = \bar{c}_i + \frac{\sum_{j \in N_u(i)} s_{items}(i,j)(r_{uj} - \bar{c}_j)}{\sum_{j \in N_u(i)} |s_{items}(i,j)|} \tag{5.4}$$

5.2.3. Standaryzacja Z (ang. *z-score normalization*)

Powyższe podejście może być dalej rozszerzone przez wzięcie pod uwagę odchylenia ocen danego użytkownika. Przykładowo, użytkownicy u_1 i u_2 mogą mieć średnią ocen 3, przy czym u_1 wystawiał wszystkie oceny z zakresu 1-5, a u_2 - tylko ocenę 3. Wtedy ocena 5 wystawiona przez u_2 oznacza większe zadowolenie z przedmiotu niż w przypadku użytkownika u_1.

Dzieląc odchylenie r_{ui} od średniej $\bar{r}_u$ przez odchylenie standardowe ocen danego użytkownika σ_u, dostajemy *standaryzację Z* oceny r_{ui}:

$$z\_score(r_{ui}) = \frac{r_{ui} - \bar{r}_u}{\sigma_u} \tag{5.5}$$

W ten sposób niwelowany jest efekt różnej rozpiętości ocen u użytkowników.
Stosując standaryzację Z, predykcję $\hat{r}_{ui}$ otrzymuje się przez

$$\hat{r}_{ui} = \bar{r}_u + \sigma_u \frac{\sum_{v \in N_i(u)} s_{users}(u,v) \frac{r_{vi} - \bar{r}_v}{\sigma_v}}{\sum_{v \in N_i(u)} |s_{users}(u,v)|}. \tag{5.6}$$

dla użytkowników oraz

$$\hat{r}_{ui} = \bar{c}_i + \sigma_i \frac{\sum_{j \in N_u(i)} s_{items}(i,j) \frac{r_{uj} - \bar{c}_j}{\sigma_j}}{\sum_{j \in N_u(i)} |s_{items}(i,j)|} \tag{5.7}$$

dla przedmiotów.

5.2.4. Eksperymenty i wnioski

Rozdział 6

Podsumowanie

W pracy dokonano szczegółowej analizy metody wspólnej filtracji opartej na algorytmie najbliższych sąsiadów. Na wstępie przedstawiono różne typy systemów rekomendacyjnych oraz ideę stojącą za wspólną filtracją. Opisano dwa podejścia do *collaborative filtering* - bazujące na użytkownikach oraz bazujące na przedmiotach. Przedstawiono też obszerny spis funkcji podobieństwa używanych w algorytmie najbliższych sąsiadów.

Zasadniczym celem autora było jednak pokazanie, że nawet przy tak nieskomplikowanym algorytmie, jakim bez wątpienia jest algorytm najbliższych sąsiadów, możliwe jest znaczne poprawienie wyników predykcji poprzez dogłębną analizę kolejnych kroków algorytmu. W kolejnych rozdziałach przeanalizowano:

1. wybór odpowiedniej funkcji podobieństwa (rozdział 4),
2. strategie konstruowania zbioru sąsiadów (sekcja 5.1),
3. metody wyznaczania ostatecznej predykcji na podstawie ocen sąsiadów (sekcja 5.2).

Dzięki kolejnym modyfikacjom algorytmu udało się ostatecznie poprawić jakość predykcji o niemal 10% w stosunku do modelu bazowego. Warto zaznaczyć, że mimo dokonanych zmian, zachowane zostały dwie zalety algorytmu, które zadecydowały o jego popularności - prostota implementacji i skalowalność.

Bibliografia

[Bea65] Juliusz Beaman, *Morbidity of the Jolly function*, Mathematica Absurdica, 117 (1965) 338–9.

[Blar16] Elizjusz Blarbarucki, *O pewnych aspektach pewnych aspektów*, Astrolog Polski, Zeszyt 16, Warszawa 1916.