Uniwersytet Warszawski
Wydział Matematyki, Informatyki i Mechaniki

Grzegorz Szpak
Nr albumu: 319400

Systemy rekomendacyjne oparte o wspólną filtrację

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 internetowych 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.

Przydatność systemów rekomendacyjnych obrazuje najlepiej tak zwany "fenomen długiego ogona" (ang. *long tail phenomenon*). + ile w google netflix

1.2. Fenomen "długiego ogona"

1.3. Typy systemów rekomendacyjnych

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

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

   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 implementacji i wysoką wydajność, 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.

   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.

1.4. The Netflix Prize

Do znacznego rozwoju badań nad systemami rekomendacyjnymi przyczynił się konkurs The Netflix Prize, zorganizowany przez internetową wypożyczalnię filmów Netflix[1] 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

[1] https://www.netflix.com

algorytmem hybrydowym opartym między innymi na wspólnej filtracji, rozkładzie macierzy według wartości osobliwych oraz wzmacnianych drzewach decyzyjnych (ang. *gradient boosted decision trees*).

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.

Definicja 2.1.1 *Zbiór wszystkich trójek* (*Użytkownik, przedmiot, ocena*) *tworzy macierz, którą nazwiemy* ***macierzą użyteczności*** *lub* ***macierzą ocen****.*

Przykład macierzy użyteczności prezentuje tabela 1 TODO

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 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, a r_u i r_i - wektorami ocen (odpowiednio) danego użytkownika oraz danego przedmiotu.

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):

$$MAE(f, S_{test}) = \frac{1}{S_{test}} \sum_{(u,i) \in S_{test}} |f(u,i) - r_{ui}| \tag{2.1}$$

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

$$RMSE(f, S_{test}) = \sqrt{\frac{1}{S_{test}} \sum_{(u,i) \in S_{test}} (f(u,i) - r_{ui})^2} \tag{2.2}$$

W dalszej części pracy autor skupi się na problemie rekomendacji sformułowanym jako problem regresji.

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

3.1. Filtracja w oparciu o użytkowników

Najczęściej używaną wersją metody wspólnej filtracji jest podejście oparte na algorytmie najbliższych sąsiadów. Schemat działania algorytmu przy wyznaczaniu r_ui polega kolejno na:

1. TODO
2. Wyborze k użytkowników o

Schemat algorytmu prezentuje rys 1 TODO

3.1.1. Miary podobieństwa

3.1.2. Filtracja w oparciu o przedmioty

Rozdział 4

Wpływ funkcji podobieństwa na jakość predykcji

Rozdział 5

Wyznaczanie sąsiedztwa w algorytmie kNN

Rozdział 6

Metody wyznaczania oceny na podstawie ocen sąsiadów

Rozdział 7

Podsumowanie

W pracy przedstawiono pierwszą efektywną implementację blabalizatora różnicowego. Umiejętność wykonania blabalizy numerycznej dla danych „z życia" stanowi dla blabalii fetorycznej podobny przełom, jak dla innych dziedzin wiedzy stanowiło ogłoszenie teorii Mikołaja Kopernika i Gryzybór Głombaskiego. Z pewnością w rozpocznynającym się XXI wieku będziemy obserwować rozkwit blabalii fetorycznej.

Trudno przewidzieć wszystkie nowe możliwości, ale te co bardziej oczywiste można wskazać już teraz. Są to:

- degryzmolizacja wieńców telecentrycznych,
- realizacja zimnej reakcji lambliarnej,
- loty celulityczne,
- dokładne obliczenie wieku Wszechświata.

7.1. Perspektywy wykorzystania w przemyśle

Ze względu na znaczenie strategiczne wyników pracy ten punkt uległ utajnieniu.

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.

[Fif00] Filigran Fifak, Gizbert Gryzogrzechotalski, *O blabalii fetorycznej*, Materiały Konferencji Euroblabal 2000.

[Fif01] Filigran Fifak, *O fetorach σ-ρ*, Acta Fetorica, 2001.

[Głomb04] Gryzybór Głombaski, *Parazytonikacja blabiczna fetorów — nowa teoria wszystkiego*, Warszawa 1904.

[Hopp96] Claude Hopper, *On some Π-hedral surfaces in quasi-quasi space*, Omnius University Press, 1996.

[Leuk00] Lechoslav Leukocyt, *Oval mappings ab ovo*, Materiały Białostockiej Konferencji Hodowców Drobiu, 2000.

[Rozk93] Josip A. Rozkosza, *O pewnych własnościach pewnych funkcji*, Północnopomorski Dziennik Matematyczny 63491 (1993).

[Spy59] Mrowclaw Spyrpt, *A matrix is a matrix is a matrix*, Mat. Zburp., 91 (1959) 28–35.

[Sri64] Rajagopalachari Sriniswamiramanathan, *Some expansions on the Flausgloten Theorem on locally congested lutches*, J. Math. Soc., North Bombay, 13 (1964) 72–6.

[Whi25] Alfred N. Whitehead, Bertrand Russell, *Principia Mathematica*, Cambridge University Press, 1925.

[Zen69] Zenon Zenon, *Użyteczne heurystyki w blabalizie*, Młody Technik, nr 11, 1969.