17.12.2019

Algorytmy numeryczne
Zadanie 3

Wojciech Rosiński
Nr indeksu: 240425

1. Opis zadania

Zadanie polegało na implementacji metody stosowanej w systemach rekomendacji oraz przeprowadzeniu jej testów w oparciu o przykładowe dane o produktach ze sklepu Amazon. Głównym założeniem zadania było wykorzystanie, zaimplementowanej na potrzeby poprzedniego projektu, klasy MyMatrix, która pozwala wykonywać operacje na macierzach oraz wyznaczać rozwiązania układów równań za pomocą metody eliminacji Gauss'a. Dodatkowo, wszelkie obliczenia potrzebne do realizacji zadania miały być przeprowadzone na zmiennych typu Double. Natomiast do celów testowych należało użyć trzech wybranych samodzielnie podzbiorów danych:

- (S) Mały (od 10 do 100 produktów),
- (M) Średni (od 100 do 1 000 produktów),
- (B) Duży (od 1 000 do 10 000 produktów).

2. Poprawność wczytywania danych

Wczytywanie danych odbywa się za pomocą specjalnie przygotowanej metody **loadProductData()**, która pozwala określić, z jakiej grupy i w jakiej ilości produkty chcemy wczytać. Dodatkowo, musimy określić minimum wystawionych opinii, jakie produkt musi posiadać, aby został wczytany. Metoda rozwiązuje również problem wielokrotnej oceny tego samego produktu przez danego użytkownika poprzez wczytanie jedynie ostatniej, wystawionej przez niego oceny.

Po wczytaniu danych do programu, uzupełniane są listy produktów i użytkowników za pomocą metody **fillUsersAndProductsLists()**. Metoda sprawdza czy kolejny dodawany produkt lub użytkownik istnieje już w odpowiedniej liście, co pozwala na uniknięcie duplikatów. Co więcej, program dostarcza metodę **filterUsersList()**, która sortuje użytkowników malejąco po ilości ocenionych produktów (uprzednio wczytanych), a następnie redukuje ilość tychże użytkowników do wartości podanej jako parametr, pozostawiając pierwszych X użytkowników (tych z największą liczbą ocenionych produktów).

Po wypełnieniu list produktów i użytkowników oraz ich odpowiedniej korekcie program za pomocą metody **fillRatingsMatrix()** wypełnia macierz ocen, w której liczba użytkowników stanowi ilość wierszy, a liczba produktów jest ilością kolumn. W ten sposób wartość macierzy o indeksie „i" oraz kolumnie „j" stanowi ocenę użytkownika „i" na temat produktu „j".

3. Testy

Zgodnie z założeniem zadania, testy przeprowadzono na trzech grupach danych: 20 produktów, 200 produktów i 2 000 produktów. Każdy z wczytanych produktów posiada minimum 100 wystawionych ocen. W każdej z trzech grup danych liczba użytkowników biorących udział w testach wynosi 100. Natomiast wartość parametru lambda jest określona na poziomie 0.1. Wszystkie pomiary są uśrednione za pomocą średniej arytmetycznej poprzez wykonanie 40 prób działania algorytmu na każdym ze zbiorów danych.

3.1. Przydatność implementowanej metody dla testowanych danych i obranych parametrów

Implementowana metoda wydaje się przydatna, ponieważ po losowym wyzerowaniu wartości (zielony kolor) w macierzy ocen i wykonaniu algorytmu wyliczona macierz rekomendacji przedstawia oceny zbliżone do uprzednio wyzerowanych. Poniżej przedstawione są kolejno: macierz ocen 10-ciu użytkowników dla 10-ciu produktów oraz wyliczona macierz rekomendacji tych produktów dla parametru D = 3 oraz λ = 0.1.

	0	1	2	3	4	5	6	7	8	9
0	3.0	5.0	3.0	**4.0**	5.0	4.0	3.0	4.0	3.0	4.0
1	4.0	4.0	4.0	**5.0**	4.0	5.0	4.0	5.0	4.0	5.0
2	5.0	**3.0**	5.0	5.0	3.0	5.0	5.0	**5.0**	5.0	5.0
3	3.0	4.0	5.0	4.0	**4.0**	4.0	3.0	4.0	3.0	4.0
4	5.0	5.0	3.0	5.0	5.0	5.0	5.0	5.0	5.0	**5.0**
5	**2.0**	5.0	5.0	5.0	5.0	5.0	2.0	5.0	2.0	5.0
6	3.0	5.0	5.0	5.0	5.0	5.0	3.0	5.0	3.0	5.0
7	5.0	5.0	5.0	5.0	5.0	5.0	5.0	**5.0**	5.0	5.0
8	5.0	**4.0**	5.0	5.0	4.0	5.0	5.0	5.0	5.0	5.0
9	5.0	5.0	2.0	4.0	5.0	4.0	5.0	1.0	5.0	**1.0**

1. Macierz ocen

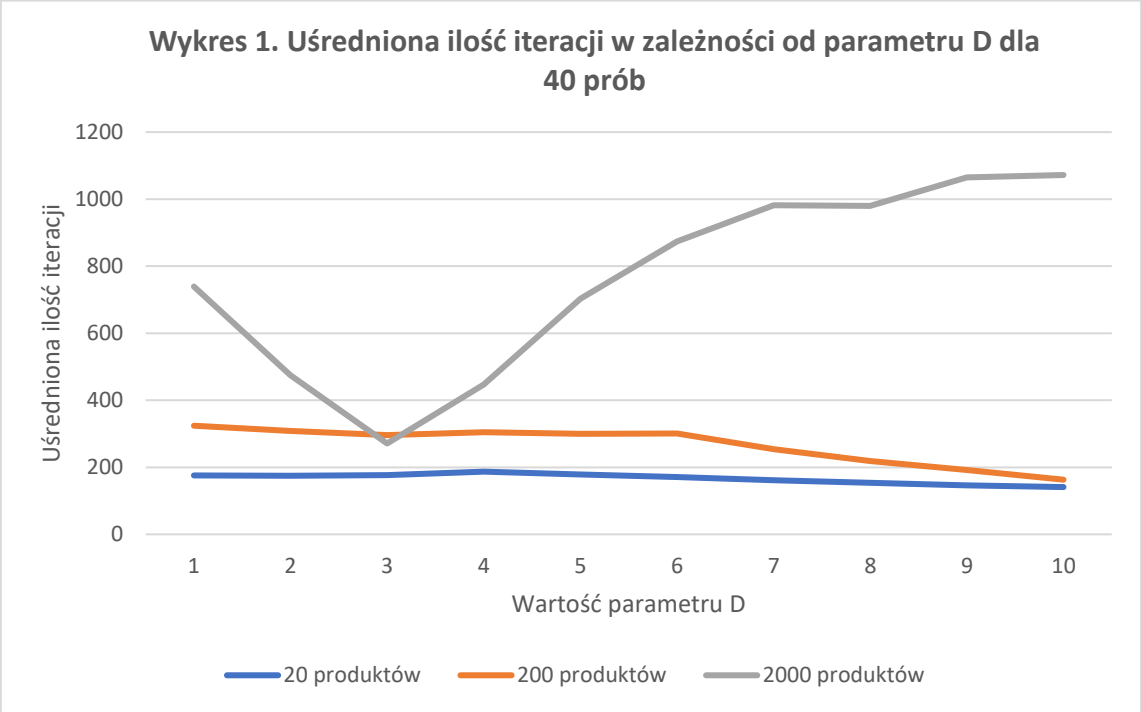
	0	1	2	3	4	5	6	7	8	9
0	3,0	4,9	3,4	**4,3**	4,9	4,2	3,0	3,5	3,0	4,0
1	4,0	4,2	4,5	**4,7**	4,0	4,7	4,0	4,9	4,0	4,7
2	5,0	**3,5**	5,0	4,8	3,1	4,9	5,0	**5,2**	5,0	5,0
3	3,0	3,9	4,2	4,1	**3,8**	4,2	2,9	4,4	2,9	4,2
4	5,0	5,0	3,9	5,0	4,9	5,0	5,0	4,2	5,0	**4,7**
5	**2,1**	5,0	5,0	4,8	5,1	4,8	2,1	5,1	2,1	4,9
6	3,0	5,0	4,9	5,0	5,0	5,0	3,0	5,1	3,0	5,0
7	4,9	5,0	4,6	5,3	4,8	5,3	4,9	**5,0**	4,9	5,1
8	5,0	**4,2**	4,7	5,0	4,0	5,1	5,0	5,2	5,0	5,0
9	5,0	5,0	1,6	3,9	5,1	3,9	5,0	1,6	5,0	**3,1**

2. Macierz rekomendacji

3.2. Tempo zbieżności w zależności od parametru D

Celem wyznaczenia tempa zbieżności algorytm wykonywał się do momentu, w którym różnica pomiędzy wartością funkcji celu pomiędzy bieżącą iteracją a poprzednią była mniejsza niż 0.01. Ilość iteracji jest uśredniona, ponieważ dla każdej grupy produktów i każdej z wartości parametru D przeprowadzono 40 prób.

Na wykresie widać, że dla małego i średniego zbioru danych tempo zbieżności rośnie wraz ze wzrostem parametru D. Natomiast dla dużego zbioru danych tempo zbieżności stopniowo rośnie do D = 3, następnie maleje wraz ze wzrostem parametru D.



3.3. Wpływ parametru D na jakość stworzonych rekomendacji i czas obliczeń.

Poniższe tabele przedstawiają uśredniony czas obliczeń [ms] i uśrednioną wartość funkcji celu w zależności od parametru D dla 40 prób wykonania się algorytmu. Na podstawie danych można stwierdzić, że dla wszystkich zbiorów danych wzrost parametru D powoduje spadek wartości funkcji celu. Dla małego zbioru danych czas obliczeń rośnie dla kolejnych wartości parametru D. Taka sama tendencja występuje dla średniego zbioru aczkolwiek od D = 9 widać spadek czasu obliczeń. Dla dużego zbioru czas obliczeń stopniowo spada do D = 3, osiągając minimum dla tego parametru, następnie rośnie.

20 produktów		
D	**Czas obliczeń**	**Wartość f. celu**
1	104	194,7367305345980
2	128	73,1905745558338
3	132	48,5054876408879
4	168	44,0814307207536
5	203	43,2710029406170
6	210	42,9701976737258
7	237	42,8599947539093
8	239	42,8016973241417
9	259	42,7807787623227
10	268	42,7741407673408

200 produktów		
D	**Czas obliczeń**	**Wartość f. celu**
1	792	9854.875934729662
2	885	6955.109348676245
3	1025	4746.63887308163
4	1338	3277.8080560828575
5	1522	2296.105303431261
6	1681	1628.577288534946
7	1666	1190.8767366050372
8	1777	910.5732823432609
9	1532	734.2909192992093
10	1447	632.6916023242381

2000 produktów		
D	**Czas obliczeń**	**Wartość f. celu**
1	15517	1025.3791671202307
2	12713	644.2562360862604
3	8893	387.2771049433344
4	17982	258.9932773977881
5	33082	193.01663663225025
6	47345	164.78045636885176
7	60459	152.67627221220155
8	68637	147.86678572394828
9	82487	145.73543147411397
10	92308	144.72278225659704