Mikołaj Koszowski 274392
Projekt FSiT Etap II

Temat: Dyfuzyjny system rekomendacji filmów.

Systemy rekomendacji są nieodłącznym elementem większości dużych serwisów i sklepów internetowych, pozwalają one klientom wydajnie je przeszukiwać i odkrywać, a samym firmom lepiej zarabiać. Można wyróżnić dwa główne typy, rekomendacja ze względu na:

- **Zawartość** *(Content-based filtering)* bazuje on na opisie i cechach, tworzy model przedmiotu i preferencji użytkownika.
- **Kolaboracyjne** *(Collaborative filtering)* wykorzystuje on oceny innych użytkowników porównując stopień podobieństwa z danym użytkownikiem.

W użyciu jednak są systemy wykorzystujące oba typy filtrowania, YouTube wykorzystuje bardziej zawartość, natomiast Netflix czy Filmweb bazuje bardziej na porównywaniu użytkowników. Algorytm rekomendujący filmy na YouTube jest liczony w milionach linijek kodu, ponadto Google wykorzystuje sztuczną inteligencję rozpoznawanie mowy i analizę obrazu do wyciągnięcia dokładniejszej zawartości filmu, zbiera i analizuje informacje takie jak polubienia, komentarze, czas oglądania czy udostępnienia i wykorzystuje ogromną bazę użytkowników w swoim systemie rekomendacji. Netflix również wykorzystuje sieci neuronowe i inne metody uczenia maszynowego jednak ma on mniej informacji o swoich użytkownikach, spora część wiedzy jest domyślana na podstawie zachowania jak częstość oglądania następnych odcinków itd. System działa całkiem dobrze, ponieważ ponad 80% filmów i seriali oglądanych na tej platformie, jest odkrytych za pomocą systemu rekomendacyjnego. Filmweb wykorzystuje w swoim Gustomierzu algorytm SVD działający na macierzach i algorytm k-najbliższych sąsiadów służący do prognozowania wartości.

Dyfuzja to proces samorzutnego rozprzestrzeniania się cząsteczek lub energii w ośrodku o temperaturze większej od zera bezwzględnego. Wynika ona z ciągłych losowych zderzeń cząsteczek. Wyróżnia się dwa typy mikroskopowa związana z ruchami Browna, śledzeniem ruchu cząstek i makroskopową opisaną równaniami i biorącą pod uwagę cały układ.

Prawo zachowania masy mówi, że dla zamkniętych układów masa układu pozostaje stała. W postaci różniczkowej wygląda następująco:

$\nabla \vec{u}\,\rho + \frac{\partial \rho}{\partial t} = r$, gdzie **ρ**- gęstość/stężenie $\vec{\mathbf{u}}$- prędkość masy **r**- kreacja/anihilacja cząstek

Dzieląc prędkość na składowe związane z dyfuzją $\mathbf{u}^{\mathbf{diff}}$ i na $\mathbf{u}^{\mathbf{r}}$ związane z pozostałymi procesami zachodzącymi w płynących płynach takie jak adwekcja, czyli przenoszenie masy przez płynące cząsteczki, otrzymujemy ogólne równanie:

$$\nabla \vec{u}^{diff}\,\rho + \nabla \vec{u}^{r}\,\rho + \frac{\partial \rho}{\partial t} = r$$

Prawa Ficka definiują one matematycznie dyfuzję makroskopową.

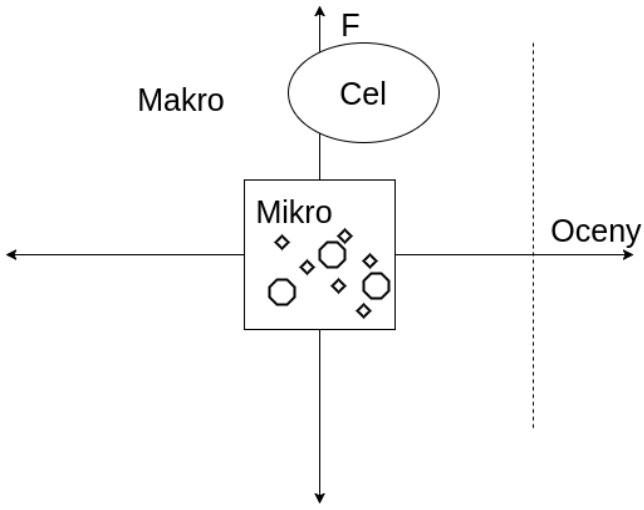
I. Definicja strumienia dyfuzji
$J = -D\nabla\rho$, D- współczynnik proporcjonalności

II. Określa zmianę stężenia z czasem
$\frac{\partial\rho}{\partial t} = D\,\Delta\rho$
Otrzymuje się z powyższych praw i założeń że:
r=0 – brak kreacji/anihilacji
$\vec{u}^{r}$=0 – brak adwekcji
D=const

Parametry modelu

Po przejrzeniu literatury zdecydowano na dyfuzję w gazie i model hybrydowy łączący dwa podejścia: mikroskopowe ruchy Browna gdzie cząsteczki filmu i użytkownika będą śledzone wraz z interakcjami w które będą wchodzić i makroskopowe gdzie będzie zachodzić dyfuzja na osi pionowej, która będzie odpowiedzialna za generalny trend ruchu cząsteczek użytkownika w górę sprzęgnięta ze stężeniami dla mikro podukładu. Między cząsteczkami filmów nie będą oddziaływać siły będą się one tylko zderzały. Sprzęgnięta to znaczy losowy ruch cząsteczek, przez coraz bardziej równomierne rozmieszczenie wszystkich filmów powoduje, że wypadkowa siła oddziaływania między cząsteczką użytkownika a cząsteczkami filmów będzie malała.

Generacja modelu użytkownika ze wszystkich dostępnych ocen zostaną wyciągnięte listy 10 najpopularniejszych reżyserów i 100 najpopularniejszych tagów określających zawartość. Wszystko brane z odpowiednią wagą zależną od oceny filmu i różnicy między oceną użytkownika, a oceną społeczności.



Generowanie gazu będzie się odbywać przed symulacją, cząsteczki filmu będą rozłożone na osi związanej z ich oceną i osi związanej z siłą oddziaływania z użytkownikiem, która może być odpychająca i przyciągająca. Ilość cząsteczek będzie zależała od parametru związanego z popularnością danego filmu. Brana pod uwagę będzie również ocena i logarytm stosunku dochodu do budżetu wszystko z odpowiednią wagą, co sprawi że zmniejszą się interakcje z filmami które przyniosły straty. Jeżeli tagi czy reżyser będą się pokrywały z wygenerowanymi listami siła również zostanie odpowiednio zwiększona. Środek układu będzie odpowiednio unormowany do średniej oceny użytkownika, których rozkład zazwyczaj jest podobny do rozkładu normalnego.

Proponowane klasy

movie

- id
- x,y i v_x, v_y
- r-promień, m-masa
- F- siła oddziaływań

user

- x,y i v_x, v_y
- r-promień, m-masa
- lista interakcji – przechowuje id filmów

Planuje się zwiększyć masę i promień cząsteczki użytkownika do dwu-krotności tego co cząsteczka filmu, żeby wykazywała większą bezwładność przy zderzeniach, przez co poruszała się mniej chaotycznie. Im większa liczba filmów będzie symulowana tym lepsze wyniki powinien dawać algorytm, więc planuję go zaimplementować w C++ ze względu na jego wydajność.

↓ ↓ ↓ Errata ↓ ↓ ↓

Struktura Baz Danych

(Kolumny mające znaczenie dla projektu)

ratings.csv (zbiór ocen użytkowników)

- userId, movieId, rating

movies_metadata.csv

- movieId, budget, revenue, popularity, vote_average

credits.csv

- movieId, crew{director, etc}

keywords.csv

- movieId, keywords[{keyId, name}]

Oddziaływania międzycząsteczkowe

Cząsteczki filmu imitują gaz doskonały i nie ma między nimi oddziaływań innych niż zderzenia. Cząsteczka użytkownika nie wpływa również na cząsteczkę filmów.

Natomiast cząsteczka filmu działa na cząsteczkę użytkownika siłą F/r^2, gdzie parametr F wygląda:

$$F = A \cdot avg + B \cdot \ln\left(\frac{revenue}{budget}\right) + C \cdot \Delta ratio_{director} + D \cdot \Delta ratio_{tag} + fluctuations$$

A,B,C,D to jeszcze nieustalone stałe , ale przewiduje się A<B<<C,D.

Branie logarytmu z ilorazu równemu wielokrotności zysków w stosunku do budżetu sprawi że składnik będzie ujemny dla nierentownych filmów, co sprawi że sama siła również będzie mniejsza, a jeżeli cały parametr F stanie się ujemny to siła z przyciągającej zamieni się w odpychającą.

Schemat działania

0. Generacja listy filmów wraz z parametrami A i B na podstawie bazy movie_metadata.csv Indywidualne dla każdego użytkownika.(Na podstawie ratings.csv credits.csv i keywords.csv)
1. Policzenie obu parametru Δratio na podstawie oceny danego filmu i jego średniej oceny z bazy.
2. Połączenie policzonych parametrów.
3. Za pomocą powyższego pliku stworzenie listy cząsteczek filmów i umieszczenie ich w przestrzeni
(Oś X to G*średnia ocena filmu(G=const.), Oś Y położenie zależne od całkowitego parametru F)
4. Symulacja zderzeń i sił w układzie, zapisywanie zderzeń użytkownika do listy.
5. Wyciągnięcie z listy interakcji najczęściej występujących filmów.
6. Sprawdzenie czy użytkownik nie obejrzał już danego filmu.
7. Stworzenie listy najbardziej polecanych.
8. Wygenerowanie pliku html z prezentacją rekomendacji.

Ewaluacja

Na podstawie bazy ocen użytkowników (26mln ocen 270k unikalnych użytkowników)
Cel systemu: Być lepszym niż losowo polecane filmy.
Cel dodatkowy: Być lepszy niż polecanie kolejnych nieobejrzanych filmów z top 250.

W zależności od wyników można zastosować do oceny rekomendowanych filmów różne podejścia:

1) Generować rekomendacje na podstawie 90% ocen i porównywać rekomendacje z pozostałymi 10%. Ma to swoje minusy ponieważ przestrzeń filmów jest bardzo duża i są małe szanse, że system poda akurat te filmy które obejrzał użytkownik i będzie można porównać kolejność propozycji z posortowanymi obejrzanymi filmami.

2) Generowanie gazu z już obejrzanymi filmami, zwiększa szanse na możliwości porównywania, jednakże optymalizowanie systemu w ten sposób trochę wykrzywia cele stawiane przed modelem, który ma polecać nieobejrzane filmy. Nadal to około 100 ocen na kilkadziesiąt tysięcy filmów

3) Przeprowadzenie małego eksperymentu z 10 osób obejrzy 3 proponowane filmy i oceni jakość tych propozycji.

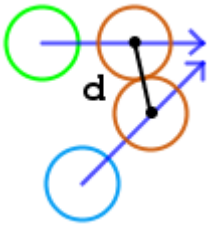
Podsumowanie parametrów wpływających na siły między cząstkami użytkownika i filmu

- ➔ lista 10 najpopularniejszych reżyserów
- ➔ lista 100 najpopularniejszych tagów
- ➔ ocena
- ➔ stosunek dochodów do budżetu

Wydajne liczenie zderzeń kul na 2-wymiarowej płaszczyźnie

1) Liczenie normy wektora kolizji.

$$\vec{n} = \frac{\overrightarrow{c_2 c_1}}{\| \overrightarrow{c_2 c_1} \|} = \frac{c_2 - c_1}{\sqrt{(c_{2x} - c_{1x})^2 + (c_{2y} - c_{1y})^2}}$$



2) Uwzględnienie stosunku prędkości i mas.

$$p = \frac{2(\vec{n}\vec{v}_1 - \vec{n}\vec{v}_2)}{m_1 + m_2}$$

3) Prędkość wynikowa.

$$\vec{w}_1 = v_1 - p m_1 \vec{n}$$
$$\vec{w}_2 = v_2 + p m_2 \vec{n}$$

Źródła

Baza danych 45 tysięcy filmów.
https://www.kaggle.com/rounakbanik/the-movies-dataset/data

Diffusion in Gases and Porous Media, R.E. Cunningham, R.J.J. Williams
Diffusion in Solids, A.S. Nowick, J.J.Burton
The Mathematics of Diffusion, J.Crank

https://en.wikipedia.org/wiki/Continuity_equation
https://en.wikipedia.org/wiki/Recommender_system
https://en.wikipedia.org/wiki/Netflix_Prize
http://www.filmweb.pl/help#I.6
http://home.agh.edu.pl/~dabrowa/files/4.Rownania-dyfuzji-wst-p-teoretyczny.pdf
http://ericleong.me/research/circle-circle/#dynamic-circle-circle-collision