Politechnika Wrocławska

Wydział Informatyki i Zarządzania
kierunek studiów: Informatyka
specjalność: Projektowanie Systemów Informatycznych

Praca dyplomowa - magisterska

TITLE
TITLE EN

Katatzyna Biernat

słowa kluczowe:
KEYWORDS

krótkie streszczenie:
SHORT ABSTRACT

Promotor:	dr inż. Bernadetta Maleszka	.....	.....
	imię i nazwisko	*ocena*	*podpis*

*Do celów archiwalnych pracę dyplomową zakwalifikowano do:**

a) kategorii A (akta wieczyste)

b) kategorii BE 50 (po 50 latach podlegające ekspertyzie)

** niepotrzebne skreślić*

pieczątka wydziałowa

Wrocław 2016

Niniejszy dokument został złożony w systemie LaTeX.

Spis treści

Streszczenie

ABSTRACT PL

Abstract

ABSTRACT EN

Todo list

Rozdział 1

Cel pracy

Celem pracy jest zaproponowanie i zbudowanie hybrydowego algorytmu rekomendacji. Składowymi docelowego algorytmu są metody kolaboratywnego filtrowania oraz metody filtrowania z analizą treści.

Rozdział 2

Wstęp

Wraz z rozwojem Internetu zmienił się sposób dostępu do informacji. Kiedyś to użytkownik musiał walczyć pozyskanie wiedzy; dzisiaj to informacje walczą u uwagę użytkowników. W świecie zalanym wiadomościami koniecznym wydaje się być zastosowanie filtra, który odsieje interesującą i wartościową zawartość od tej niechcianej. Tak też z pomocą przychodzą zautomatyzowane mechanizmy rekomendacji.

Jednakże sama idea rekomendacji nie jest niczym nowym. Co więcej, zjawisko to możemy zaobserwować w naturze – na przykład wśród mrówek, które podążają wyznaczoną (rekomendowaną) ścieżką feromonową w poszukiwaniu pożywienia.

Ludzie od niepamiętnych czasów posiłkowali się opiniami innych aby ułatwić sobie dokonanie wyboru, od najbliższego grona znajomych do ekspertów i autorytetów.

Wraz z rozwojem nauk informatycznych problem rekomendacji stał się problemem interesującym badaczy. Za pierwszy system rekomendacji uznaje się *Tapestry* stworzony w laboratoriach Xerox Palo Alto Research Center w 1992 roku. Motywacją było odfiltrowanie rosnącej liczby niechcianej poczty elektronicznej [13].

Wkrótce później idea ta została rozszerzona przez takich graczy jak Amazon, Google, Pandora, Netflix, Youtube, Yahoo etc. aż do formy, jaką znamy dzisiaj: systemu, który sugeruje użytkownikom produkty, filmy, muzykę, strony internetowe na podstawie ich aktywności w sieci [36].

Wielkie koncerny internetowe stale poprawiają jakość swoich algorytmów rekomendacji. Najlepszym przykładem jest tutaj Netflix, który w październiku 2006 zorganizował ogólnodostępny konkurs na najlepszy algorytm. Zadaniem uczestników było ulepszenie algorytmu Cinematch. Już po siedmiu dniach od ogłoszenia konkursu trzy zespoły zdołały przebić Cinematch o 1.06% [26][28]. 18 września 2009 Netflix ogłosił, że zespół BellKor's Pragmatic Chaos poprawił Cinematch o 10,06% osiągając wynik $RMSE = 0.8567$. Tym samym wygrał nagrodę w wysokości $1,000,000 i zakończył konkurs [27][29].

Systemy rekomendacji ulepszane są nieustannie, o czym świadczy chociażby organizowana rokrocznie konferencja *ACM International Conference on Recommender Systems*. Tematyka ta poruszana jest także na konferencjach *European Conference on Information Retrieval*, *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* i wielu innych. Mimo dużego stopnia

zaawansowania wciąż istnieje pole manewru do ulepszania algorytmów rekomendacji i co za tym idzie zwiększanie zadowolenia użytkowników, które z kolei prowadzi do osiągania korzyści biznesowych.

Rozdział 3

Przegląd istniejących rozwiązań

Tradycyjnie wyróżniamy następujące techniki rekomendacji:

- **filtrowanie w oparciu o zawartość** (eng. content-based), technika koncentrująca się na atrybutach elementów. Użytkownikowi rekomendowane są elementy, które podobne są do tych wybieranych przez niego w przeszłości;
- **filtrowanie kolaboratywne** (eng. collaborative filtering), technika polegająca na odnajdywaniu użytkowników o podobnych gustach i sugerowaniu lubianych przez nich elementów aktualnie aktywnemu użytkownikowi;
- **filtrowanie demograficzne** (eng. demographic), technika koncentrująca się na sugerowaniu aktywnemu użytkownikowi elementów popularnych pośród użytkowników z tej samej okolicy bądź w podobnym przedziale wiekowym;
- **filtrowanie z analizą domeny wiedzy** (eng. knowledge-based), technika dobierająca kolejne elementy na podstawie określonej domeny wiedzy na temat tego, jak dany element spełnia potrzeby i preferencje użytkownika;
- **filtrowanie z analizą społecznościową** (eng. community-based), technika dobierająca rekomendacje dla użytkownika w zależności od preferencji innych użytkowników z jego sieci społecznościowej. W myśl zasady "powiedz mi kim są twoi przyjaciele a powiem ci kim jesteś";
- **hybrydowe systemy rekomendacji**, to kombinacja dowolnych powyższych technik.

Każda z tych technik ma swoje wady i zalety w zależności od kontekstu, w którym ma być stosowana[31].

3.1. Filtrowanie w oparciu o zawartość

Filtrowanie content-based opiera się na cechach elementów w systemie. Rekomendowane są obiekty, które podobne są do tych pozytywnie ocenionych wcześniej przez użytkownika[14]. W zależności od domeny pod uwagę mogą być brane słowa kluczowe, cechy takie jak rok wydania, reżyser, autor, kompozytor, gatunek itp.

3.1.1. Metody tworzenia profilu użytkownika

Profil użytkownika może być tworzony na dwa sposoby. Jeżeli użytkownik jawnie pozostawia informacje można mówić o podejściu aktywnym (explicit feedback). Do takich informacji należą: ocena konkretnych elementów, tzw. łapka w górę lub w dół, komentarz itp.

Jednakże nawet jeżeli użytkownik nie jest skory do zostawiania tego typu śladów, to i tak można wiele na jego temat wywnioskować korzystając z podejścia pasywnego (implicit feedback). System bierze wówczas pod uwagę aktywność użytkownika taką jak: historia zakupów, historia przeglądarki a nawet ruchy myszką. W przypadku serwisu z muzyką czy filmem cenną informacją będzie fakt, czy użytkownik wysłuchał lub obejrzał dany materiał do końca czy też wyłączył go po paru sekundach. [23][19]

3.1.2. Zalety podejścia content-based

Do zalet filtrowania w oparciu o zawartość należy niezależność użytkownika. Podczas budowania rekomendacji brany pod uwagę jest tylko jego profil; aktywność innych aktorów w systemie nie wpływa na wynik końcowy. Inną przewagą jest przejrzystość – każda propozycja jest w pełni uzasadniona, gdyż opiera się na działaniach użytkownika w przeszłości (podczas gdy w przypadku filtrowania kolaboratywnego mamy do czynienia z czarną skrzynką). Ponadto, tego typu algorytm ma możliwość zaproponowania elementu, który nie był nigdy wcześniej oceniany przez nikogo. Zapobiega to zjawisku długiego ogona [23].

3.1.3. Najczęściej spotykane problemy

Aby rekomendacja była skuteczna użytkownik powinien ocenić jak najwięcej elementów. Problematyczni są zatem użytkownicy, którzy dopiero co dołączyli do serwisu oraz tacy, którzy nie są aktywni i rzadko zostawiają po sobie ślad [24].

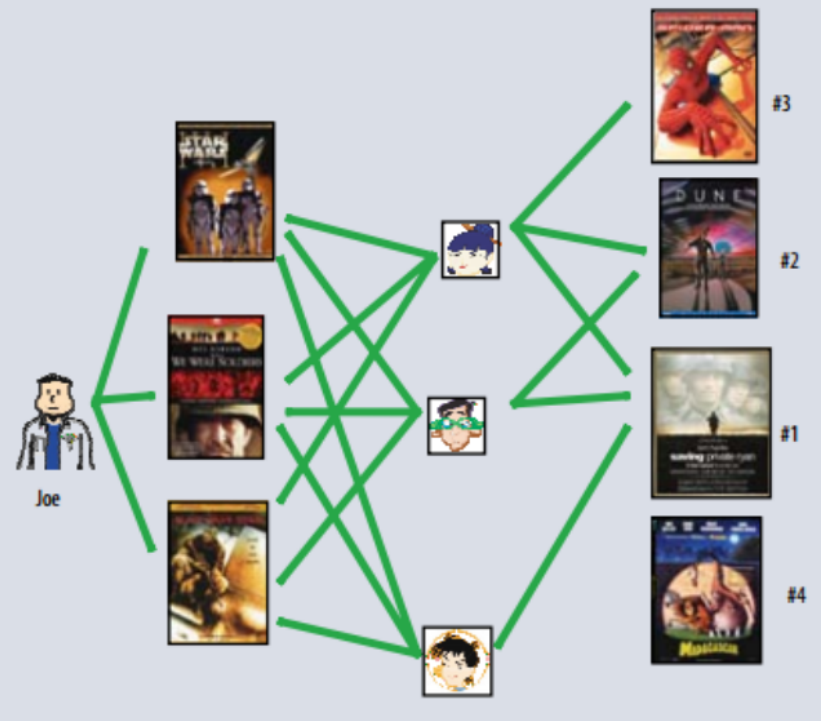
Podejście content-based jest podatne na pułapkę tzw. bańki informacyjnej. Jeżeli w systemie rekomendującym produkcje kinowe użytkownik do tej pory oceniał jedynie filmy akcji, to mało prawdopodobne jest, że algorytm zaproponuje mu ciekawy dramat obyczajowy. Nowe propozycje nie są zaskakujące[23].

3.2. Filtrowanie kolaboratywne

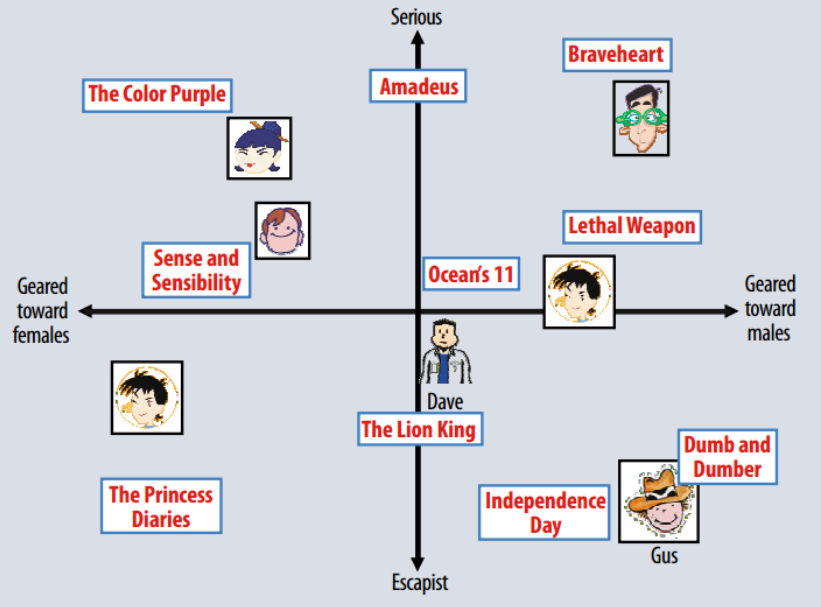
Filtrowanie kolaboratywne opiera się o założenie, że ludzie o zbliżonym guście dokonują podobnych wyborów. Użytkownicy o zbliżonym guście to osoby, które oceniły konkretne elementy w podobny sposób[31][35][14].

W przypadku filtrowania kolaboratywnego można wyróżnić dwa główne podejścia: oparte o regułę sąsiedztwa (ang. *neighborhood*) oraz oparte o model (ang. *model-based*), wykorzystujące modele ukrytych parametrów[20][19].

Filtrowanie w oparciu o regułę sąsiedztwa koncentruje się na związkach element-element bądź użytkownik-użytkownik[19]. Rysunek 3.1 pokazuje regułę sąsiedztwa skoncentrowaną na relacji użytkownik-użytkownik. Joe ocenił trzy filmy. System odnajduje innych użytkowników, którzy ocenili te trzy pozycje podobnie jak Joe. Każdy z nich pozytywnie ocenił film „Saving Private Ryan", zatem jest to pierwsza rekomendacja dla Joe.



Rys. 3.1: Filtrowanie kolaboratywne metodą sąsiedztwa, zorientowane na użytkownika[20].



Rys. 3.2: Filtrowanie kolaboratywne z wykorzystaniem modelu ukrytych parametrów[20].

Ideą podejścia model-based jest zbadanie i modelowanie zależności element-użytkownik wraz z czynnikami reprezentującymi ukryte własności elementów i użytkowników. Taki

model jest następnie uczony przy użyciu dostępnych danych. W rezultacie można z niego odczytać przewidywaną ocenę elementu dla konkretnego użytkownika[6][19].

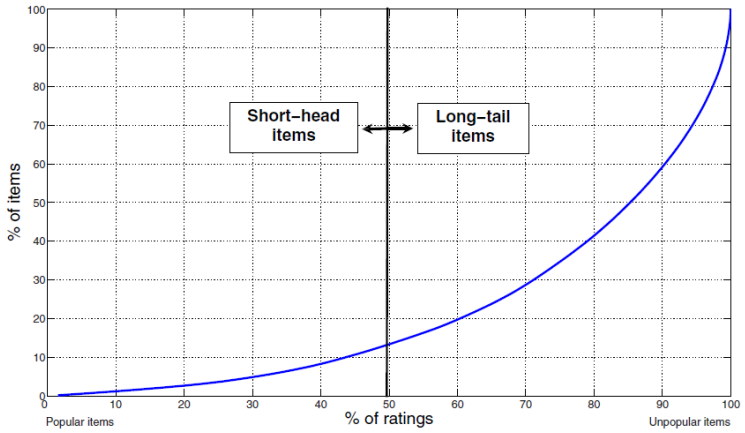
Rysunek 3.2 pokazuje w sposób uproszczony podejście oparte o model. W układzie współrzędnym oznaczeni są użytkownicy wedle swoich preferencji oraz konkretnych cech (np. płeć) a także filmy, które stanowią odpowiedź na dany zestaw preferencji/cech [20].

3.2.1. Najczęściej spotykane problemy

Jednym z problemów klasycznego podejścia do kolaboratywnego filtrowania jest brak uwzględnienia dynamiki zmian w gustach użytkowników. Ten sam użytkownik na przestrzeni kilku lat lub miesięcy może zupełnie inaczej ocenić ten sam film bądź piosenkę. Rozwiązaniem jest dodanie czynnika czasu podczas obliczania wag kolejnych ocen. [4][17][20].

Innym problemem jest tzw. zimny start (eng. cold start). Polega on na tym, że użytkownicy nowi w systemie ocenili zbyt mało elementów, aby można było zbudować dla nich dobre rekomendacje[39][33].

Powszechnym zjawiskiem jest tzw. efekt długiego ogona. Rysunek 3.3 przedstawia jak rozkłada się procentowa ilość ocen danych elementów w zależności od ich popularności. Jeżeli algorytm rekomendacji nie wspiera mniej popularnych elementów, to istnieje ryzyko, że użytkownicy nie otrzymają możliwości eksplorowania nowych, niszowych materiałów[33][3].



Rys. 3.3: Problem długiego ogona: 50% ocen dotyczy 10-12% najpopularniejszych elementów w systemie[33].

Systemy rekomendacji wykorzystujące filtrowanie kolaboratywne nie są skalowalne. Złożoność rośnie proporcjonalnie do ilości użytkowników i elementów. Wielkie koncerny internetowe takie jak Twitter wykorzystają klastry i maszyny z bardzo dużą ilością pamięci aby zachować płynność działania serwisu [10].

3.3. Popularne serwisy wykorzystujące algorytmy rekomendacji

W przeciągu ostatnich lat algorytmy rekomendacji zagościły na bardzo wielu popularnych serwisach internetowych z różnych domen. Poniższa lista prezentuje garstkę wybranych stron.

3.3.1. Rekomendacja muzyki

- **YouTube** – serwis powstały w 2005 roku, pozwalający na bezpłatne umieszczanie, odtwarzanie, ocenianie i komentowanie filmów. Od 2006 roku przejęty przez Google. YouTube buduje profil użytkownika w oparciu o jego aktywność w serwisie. Brane pod uwagę są polubienia (łapka w górę), subskrypcje, udostępnianie a także informacje czy użytkownik obejrzał film do końca czy tylko pewien jego procent. Techniki rekomendacji stosowane przez serwis to przede wszystkim asocjacyjna eksploracja danych i licznik wspólnych odwiedzin danego wideo w czasie trwania pojedynczej sesji [5].
- **LastFM** – internetowa radiostacja oferująca rozbudowany mechanizm rekomendacji piosenek "Audioscrobbler".
- **Pandora** – spersonalizowane radio internetowe wykorzystujące projekt Music Genome Project. Każda piosenka przeanalizowana jest pod kątem maksymalnie 450 cech; na tej podstawie budowane są rekomendacje[1].

3.3.2. Rekomendacja filmów

- **Netflix** – amerykańska platforma oferująca strumieniowanie filmów i seriali. Działający od 2007 roku gigant oferuje rozbudowany system rekomendacji Cinematch[30].
- **Filmweb** – polski serwis poświęcony filmom i jednocześnie druga największa baza filmowa na świecie. Oferuje system rekomendacji Gustomierz, który umożliwia poznawanie nowych filmów w guście użytkownika[7].
- **Internet Movie Database (IMDb)** – największa internetowa baza filmów. Baza zawiera $3,837,014$ pozycji, które są oceniane w skali od 1 do 10 przez użytkowników[16].

3.3.3. Platformy typu e-commerce

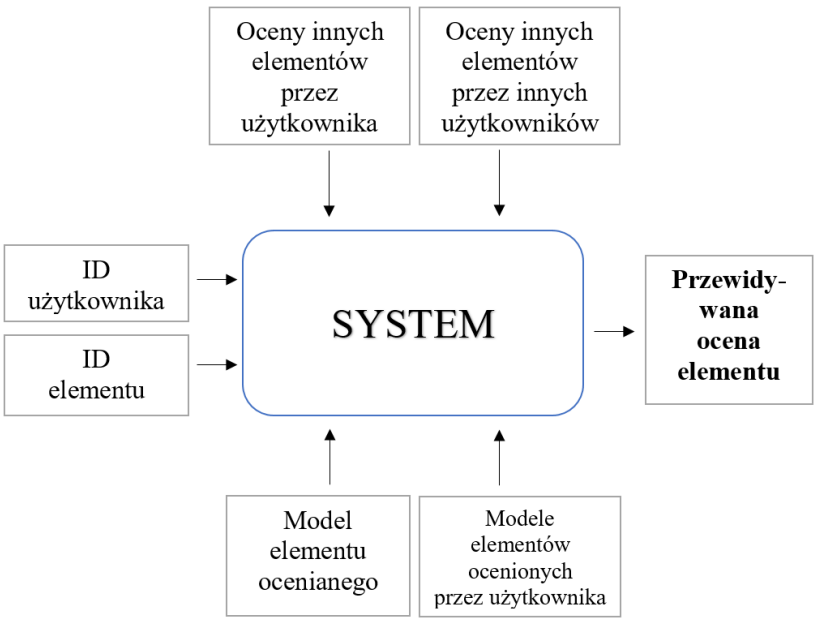
- **Allegro** – polski portal aukcyjny. Swoim użytkownikom oferuje panel rekomendacji. Prezentowane produkty wybierane są w oparciu o to co dotychczas kupował i oglądał użytkownik[2].
- **Amazon** – największy na świecie sklep internetowy typu B2C. Amazon w swoich mechanizmach rekomendacji wykorzystuje algorytmy filtrowania kolaboratywnego typu item-to-item[22].

Rozdział 4

Model systemu

Głównym założeniem systemu zaproponowanego przez autorkę jest połączenie zalet kolaboratywnego filtrowania i filtrowania w oparciu zawartość minimalizując jednocześnie wady obu podejść.

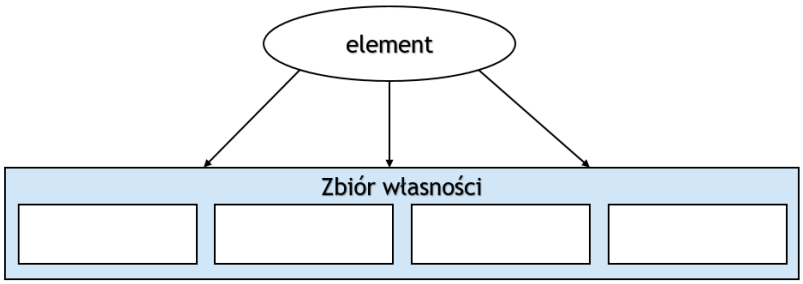
Rys. 4.1 przedstawia czarnoskrzynkowy model systemu. Danymi wejściowymi są numery identyfikacyjne użytkownika dla którego ma być zbudowana rekomendacja oraz elementu, dla którego ma być przewidziana ocena. System pobiera model elementu a także modele wszystkich innych elementów, które użytkownik ocenił w przyszłości. Jednocześnie pobierane są informacje o tym jak użytkownicy systemu ocenili inne elementy. Wynikiem wyjściowym jest predykcja – jak aktywny użytkownik oceni element.



Rys. 4.1: Model czarnoskrzynkowy

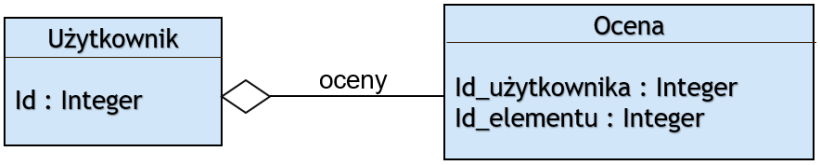
Założeniem systemu jest uniwersalność, zatem model elementu jest uogólniony i dostosowuje się w zależności do domeny, w której system jest wykorzystywany. Rys.

4.2 przedstawia reprezentację elementu w systemie. W zbiorze wartości mogą znaleźć się takie pozycje jak lista aktorów, reżyser (w przypadku filmów), gatunek, wykonawca (w przypadku muzyki), typ produktu lub cena (w przypadku systemów typu e-commerce).



Rys. 4.2: Uogólniony model elementu

Każdy użytkownik systemu jest anonimowy. Nie jest znana jego płeć, wiek, pochodzenie itp. System nie przechowuje także informacji właściwych mediom społecznościowym takich jak relacje między użytkownikami (przyjaźnie, śledzenie). Wiadomym jest jedynie jakie elementy zostały ocenione i jak zostały ocenione. Rys. 4.3 przedstawia reprezentację użytkownika w systemie.



Rys. 4.3: Uogólniony model elementu

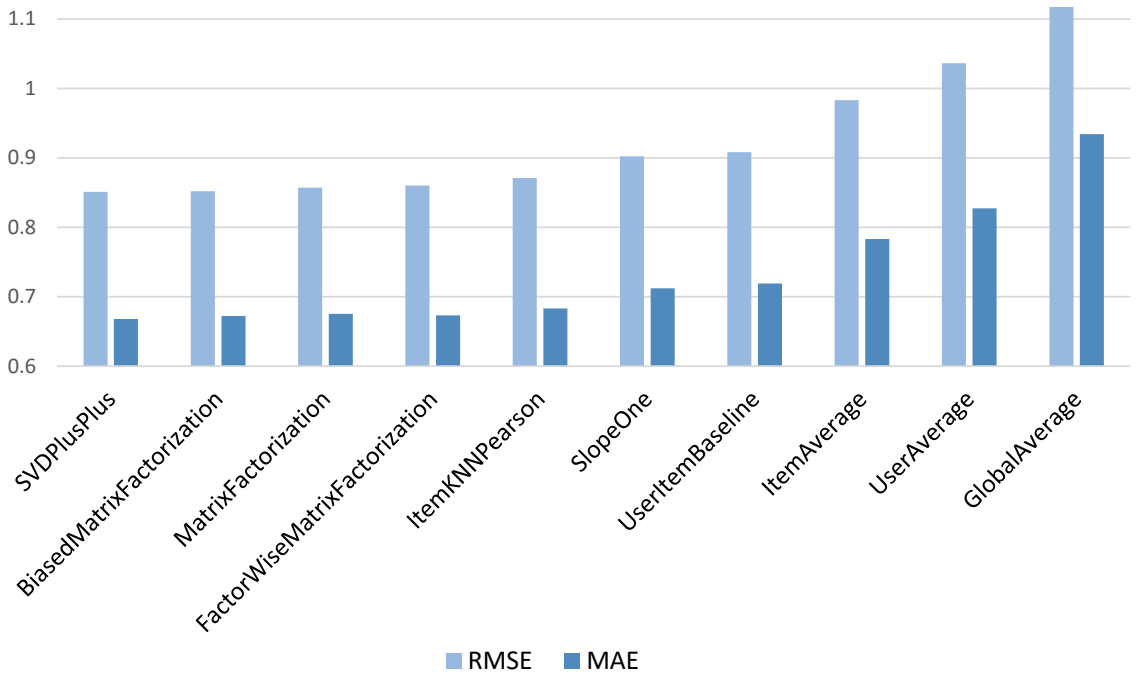
Rozdział 5

Algorytmy

5.1. Filtrowanie kolaboratywne

Implementacja algorytmów collaborative-filtering, które wykorzystane zostały w projekcie pochodzą biblioteki MyMediaLite [8][9].

Przed podjęciem decyzji dotyczącej wyboru algorytmu kolaboratywnego filtrowania zostały przeanalizowane testy na bazie MovieLens M1[11]. Testy przeprowadzone zostały z pięciokrotną walidacją krzyżową. Rys. 5.1 przedstawia wyniki (im mniejsza wartość RMSE i MAE tym lepiej).



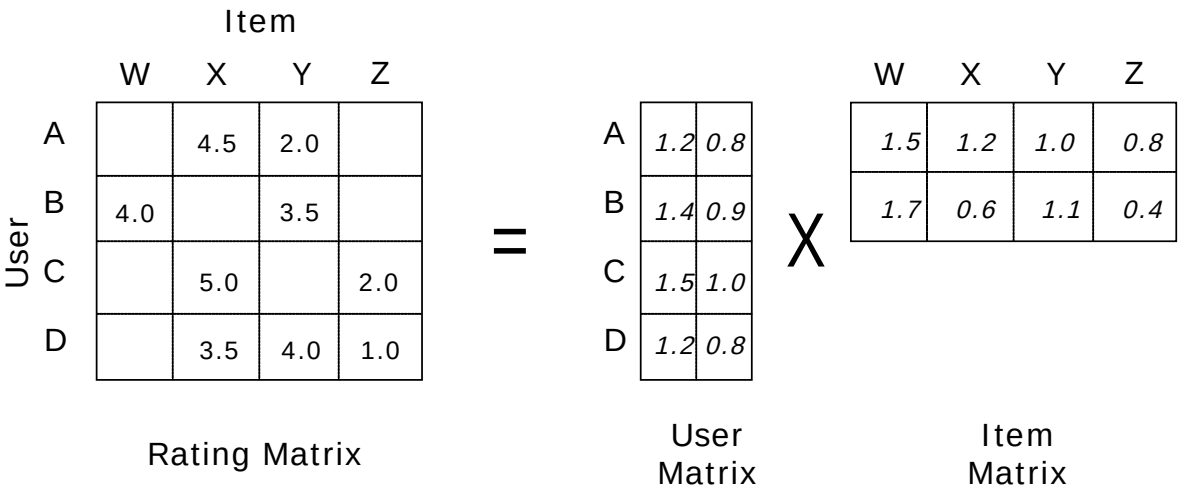
Rys. 5.1: Test algorytmów filtrowania kolaboratywnego [25]

Najefektywniejsze okazały się algorytmy SVD++, Biased Matrix Factorization i Matrix Factorization bazujące na podejściu model-based.

5.1.1. Matrix Factorization

Metoda matematyczna

Algorytm *Matrix Factorization* bazuje na matematycznej metodzie rozkładu macierzy na czynniki



Rys. 5.2: Faktoryzacja macierzy [32]

Początkowo dana jest niekompletna macierz zawierająca oceny, jakie użytkownicy wystawili konkretnym elementom (*Rating Matrix*). Celem metody jest odnalezienie wartości, jakie można wstawić w puste miejsca, czyli przewidzenie jaką ocenę dany użytkownik wystawi nieocenionemu jeszcze elementowi.

Tworzone są więc odrębne macierze dla użytkowników i elementów zawierające ukryte własności. Każdy element powiązany jest z wektorem $q_i \in \mathbb{R}^f$ a każdy użytkownik z wektorem $p_u \in \mathbb{R}^f$. Wartości czynników ukrytych determinują stopień zainteresowania daną cechą (w przypadku macierzy użytkowników) bądź stopień, w jakim dany element posiada tę cechę (w przypadku macierzy elementów).

Iloczyn skalarny $q_i^T p_u$ przedstawia relację pomiędzy użytkownikiem a elementem. Na tej podstawie można wnioskować ocenę r_{ui}, jaką użytkownik może wystawić elementowi: $r_{ui} = q_i^T p_u$ i w konsekwencji estymować zainteresowanie użytkownika danym elementem [20]..

Głównym wyzwaniem jest odnalezienie wartości macierzy użytkownika i elementu, które po przemnożeniu przez siebie dadzą kompletną macierz ocen. W przypadku omawianych algorytmów stosowana jest metoda stochastycznego gradientu prostego.

Przebieg algorytmu

Przebieg algorytmu składa się z dwóch faz. W pierwszej fazie inicjowany jest model (Algorytm 1). Daną wejściową jest macierz zawierająca dotychczasowe oceny wszystkich elementów przez wszystkich użytkowników w systemie. Na wyjściu otrzymywane są dwie nowe macierze reprezentujące ukryte własności użytkowników i elementów. Na tym etapie są one wypełnione wartościami losowymi.

Algorytm 1 Matrix Factorization – Inicjacja modelu

$N \leftarrow$ liczba użytkowników
$M \leftarrow$ liczba elementów
$F \leftarrow$ liczba ukrytych własności
$ratings \leftarrow$ Macierz NxM zawierająca dotychczasowe oceny wszystkich elementów przez wszystkich użytkowników
$user\_factors \leftarrow$ Macierz NxF reprezentująca ukryte własności użytkowników
$item\_factors \leftarrow$ Macierz MxF reprezentująca ukryte własności elementów
for each $uf \in user\_factors$, $if \in item\_factors$ **do**
  wstaw losową wartość za pomocą transformacji Boxa-Mullera
end for
for each $user, item \in ratings$ **do**
  if $ratings_{user,item} = NULL$ **then**
    wstaw 0 do wiersza $user\_factors_{user}$ i $item\_factors_{item}$
  end if
end for
return $user\_factors, item\_factors$

W fazie drugiej następuje uczenie metodą stochastycznego gradientu prostego (Algorytm 2). Wynikiem tej fazy są macierze reprezentujące ukryte własności użytkowników i elementów. Mnożąc je ze sobą uzyskiwana jest przewidywana ocena każdego z elementów przez użytkowników.

Przed rozpoczęciem uczenia ustalane są parametry: parametr regulujący (ang. *regularization*), tempo uczenia (ang. *learn rate*), parametr zanikania (ang. *decay*) i liczba iteracji. W trakcie trwania głównej pętli parametr regulujący pozostaje niezmienny. Służy on zapobieganiu zjawiska nadmiernego dopasowania (ang. *overfitting*). Tempo uczenia jest przy każdym przebiegu pętli mnożone przez parametr zanikania, dzięki czemu można kontrolować jak kolejne przebiegi pętli wpływają na finalny wynik.

Ostatnim parametrem ustalanym przed główną pętlą jest skośność globalna (ang. *global bias*), która jest średnią wszystkich znanych ocen.

W pętli uczenia wykonywane są następujące operacje: dla każdej pary użytkownik – element budowana jest przewidywana ocena poprzez obliczenie iloczynu skalarnego odpowiednich wartości z macierzy wartości ukrytych. Ocena ta jest modyfikowana poprzez dodanie globalnej skośności a następnie porównywana z faktyczną oceną elementu przez użytkownika. Tak uzyskany błąd służy do wyliczenia delty. Macierze wartości ukrytych uaktualniane są o wyliczoną deltę.

Pod koniec każdej iteracji aktualizowane jest tempo uczenia.

Algorytm 2 Matrix Factorization – Faza uczenia

global_bias ← średnia wszystkich ocen
X ← liczba iteracji
regularization ← parametr regulujący
current_learnrate ← tempo uczenia
decay ← paramert zanikania
for each $x \in X$ **do**
  for each $user, item \in ratings$ **do**
    $predicion = global\_bias + IloczynSkalarny(user\_factors_{user}, item\_factors_{item})$;
    $error = ratings_{user,item} - prediction$
    //dopasowanie własności ukrytych:
    for each $f \in F$ **do**
      $delta_u = error * item\_factors_{item,f} - regularization * user\_factors_{user,f}$
      $delta_i = error * user\_factors_{user,f} - regularization * item\_factors_{item,f}$
      $user\_factors_{user,f}$ += $current\_learnrate * delta_u$
      $item\_factors_{item,f}$ += $current\_learnrate * delta_i$
    end for
  end for
  current_learnrate *= *decay*
end for
return $user\_factors, item\_factors$

5.1.2. Biased Matrix Factorization

Algorytm *Biased Matrix Factorization* jest modyfikacją wyżej opisanego algorytmu *Matrix Factorization*. Podobnie jak jego pierwowzór składa się z dwóch faz. W pierwszej fazie dodatkowo inicjowane są dwa dodatkowe wektory: skośność użytkowników (ang. *user bias*) i skośność elementów (ang. *item bias*).

Inaczej jest też obliczana skośność globalna:

$$global\_bias = \frac{\frac{a - r_{min}}{r_{max} - r_{min}}}{1 - \frac{a - r_{min}}{r_{max} - r_{min}}}, \tag{5.1}$$

gdzie

a to średnia wszystkich ocen
r_{min} to minimalna ocena w systemie
r_{max} to maksymalna ocena w systemie

Faza druga wygląda podobnie jak w przypadku algorytmu *Matrix Factorization*, jednak są uwzględniane dodatkowe parametry i wykonywane dodatkowe kroki.

Algorytm 3 Biased Matrix Factorization – Faza uczenia

$global\_bias \leftarrow$ średnia wszystkich ocen
$X \leftarrow$ liczba iteracji
$regU, regI, BiasReg \leftarrow$ parametry regulujące dla użytkownika, elementu i ogólny
$current\_learnrate \leftarrow$ tempo uczenia
$BiasLearnRate \leftarrow$ tempo uczenia skośności
$decay \leftarrow$ paramert zanikania
for each $x \in X$ **do**
  for each $user, item \in ratings$ **do**
    $score = global\_bias + user\_bias_{user} + item\_bias_{item} + IloczynSkalarny(user\_factors_{user}, item\_factors_{item})$
    $sig\_score = \frac{1}{1+\exp(-score)}$
    $predicion = rating_{min} + sig\_score + (rating_{max} - rating_{min})$
    $error = ratings_{user,item} - prediction$
    $gradient\_common = err * sig\_score * (1 - sig\_score) * (rating_{max} - rating_{min})$
    //dopasowanie skośności:
    $user\_bias_{user}$ += $BiasLearnRate * current\_learnrate * (gradient\_common - BiasReg * RegU * user\_bias_{user})$
    $item\_bias_{item}$ += $BiasLearnRate * current\_learnrate * (gradient\_common - BiasReg * RegI * item\_bias_{item})$
    //dopasowanie własności ukrytych:
    for each $f \in F$ **do**
      $delta_u = gradient\_common * item\_factors_{item,f} - RegU * user\_factors_{user,f}$
      $delta_i = gradient\_common * user\_factors_{user,f} - RegI * item\_factors_{item,f}$
      $user\_factors_{user,f}$ += $current\_learnrate * delta_u$
      $item\_factors_{item,f}$ += $current\_learnrate * delta_i$
    end for
  end for
  $current\_learnrate$ *= $decay$
end for
return $user\_factors, item\_factors$

5.1.3. SVD++

opisać SVD++

Algorytm SVD++ to rozszerzenie popularnej metody SVD. SVD, czyli dekompozycja głównych składowych (ang. *Singular Value Decomposition*) jest metodą matematyczną służąca do redukcji wymiaru macierzy. Klasyczne SVD nie działa w przypadku gdy macierz wejściowa jest niekompletna. W przeszłości problem ten był rozwiązywany poprzez wpisywanie wartości w puste miejsca, co jednak jest mało efektywne obliczeniowo i może zaburzać finalny wynik [34][20].

zatem w takiej postaci nie nadaje się do wykorzystania w algorytmach rekomendacji. Dopiero rozszerzenie, SVD++, pozwala na przewidywanie ocen elementów przez użytkowników.

5.2. Filtrowanie z analizą zawartości

Algorytmy content-based budują rekomendację na podstawie ocen, jakie zostały dotychczas wystawione przez użytkownika. Analizowane są cechy elementów i ich wartości oraz określana jest ich siła wpływu na finalną ocenę.

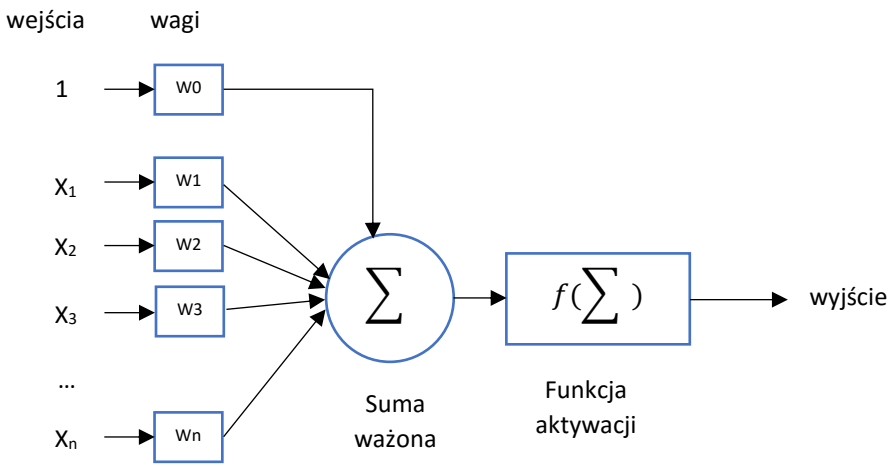
W tym celu dla każdego użytkownika tworzona jest sieć neuronowa, która uczy się jego preferencji.

W projekcie wykorzystana została implementacja sieci neuronowych z biblioteki AForge.NET Framework [18].

5.2.1. Konstrukcja sieci neuronowej

Struktura perceptronów

Sieć neuronowa składa się z trzech warstw neuronów (perceptronów). W każdej warstwie wszystkie neurony mają konstrukcję na jak rys. 5.3



Rys. 5.3: Schemat perceptronu

Do neuronu przekazywany jest zestaw wartości w postaci wektora x. Następnie obliczana jest suma ważona tych wartości w zależności od nadanych wag w. Wynik tej operacji przekazywany jest do funkcji aktywacji neuronu. Jeżeli funkcja przyjmie wartość wyższą lub równą niż określony próg aktywacji, to perceptron zostanie pobudzony (zwróci wartość 1). Proces ten obrazuje równanie 5.2.

$$N(x_1, x_2, x_3, \ldots, x_n) = \begin{cases} 1 & \text{jeśli } f(w_0 + \sum_{i=1}^{n} w_i x_i) \geqslant \eta \\ 0 & \text{jeśli } f(w_0 + \sum_{i=1}^{n} w_i x_i) < \eta \end{cases}, \tag{5.2}$$

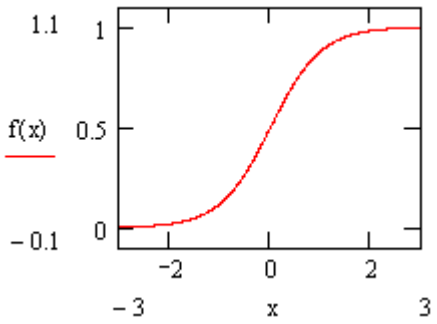
gdzie

w to wagi kolejnych wejść
x to wartości przekazywane do wejść
$f(u)$ to funkcja aktywacji neuronu
η to próg aktywacji neuronu

Na potrzeby algorytmu rekomendacji zdecydowano się przyjąć sigmoidalną funkcję aktywacji neuronu (równanie 5.3). Funkcja przyjmuje wartości z zakresu $[0, 1]$.

$$f(x) = \frac{1}{1 + \exp(-\alpha x)}\,. \tag{5.3}$$

Wykres funkcji wygląda jak na rys. 5.4.



Rys. 5.4: Wykres sigmoidalnej funkcji aktywacji perceptronu [18]

Struktura sieci i przebieg algorytmu

Pierwszym etapem algorytmu jest analiza cech elementów ocenionych przez użytkownika. Tworzona jest lista wszystkich występujących cech które powtarzają się minimum tyle razy, ile wynosi wartość parametru *minimumRepeatingFeatures*.

Następnie inicjowana jest sieć neuronowa. Ilość neuronów warstwy wejściowej jest równa ilości wyodrębnionych cech. Warstwa ukryta zawiera tyle neuronów ile jest to określone parametrem *hiddenLayerNeurons*. Warstwa wyjściowa składa się z tylko jednego neuronu (rys. 5.5).

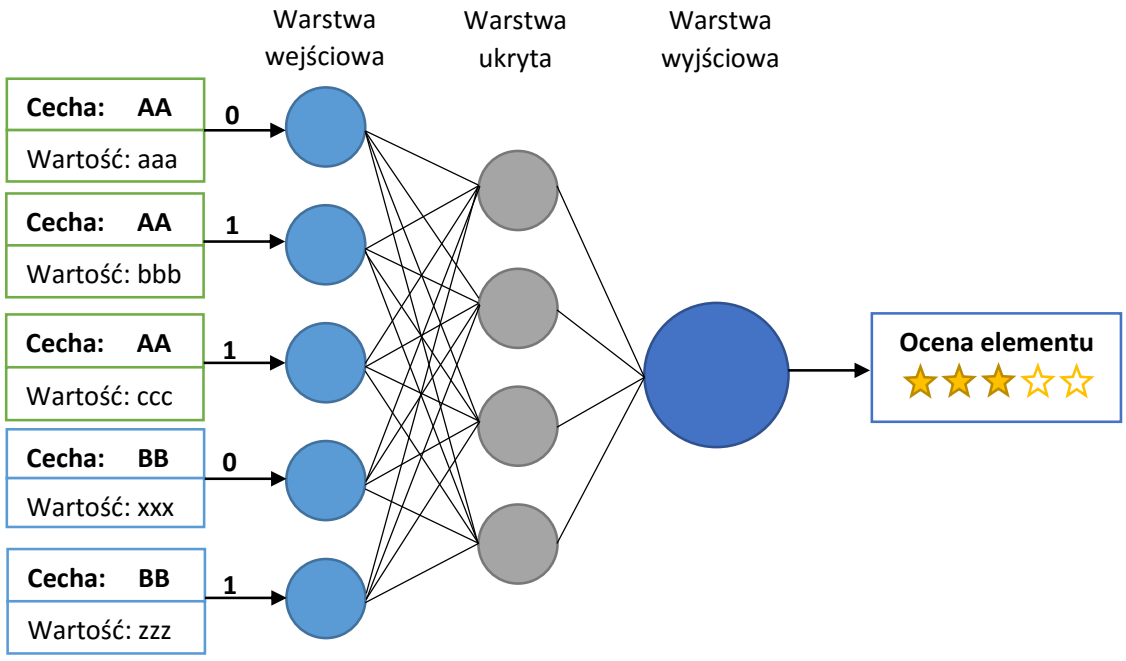
W kolejnym etapie dla każdego elementu tworzona jest mapa cech. Jeżeli element zawiera daną cechę o danej wartości przypisywana jest wartość 1. W przeciwnym razie wstawiane jest 0. Rys. 5.6 przedstawia przykładową mapę cech. Tak przygotowana lista przekazywana jest do sieci neuronowej.

Na wyjściu sieć zwraca przewidywaną ocenę elementu.

5.2.2. Uczenie sieci neuronowej

Aby sieć zwracała jak najlepsze wyniki musi wcześniej zostać nauczona preferencji użytkownika - tzn. dla każdej pary cecha-wartość powinna zostać odnaleziona odpowiednia waga.

W zależności od wybranej opcji sieć neuronowa może być uczona w wykorzystaniem propagacji wstecznej, algorytmu RPROP lub algorytmu genetycznego.



Rys. 5.5: Schemat sieci neuronowej

Cecha	Wartość	Czy zawiera?
Aktor	Julia Roberts	1
Aktor	Al Pacino	1
Aktor	Brad Pitt	0
...		
Reżyser	Francis Ford Coppola	1
Reżyser	Darren Aronofsky	0
...		

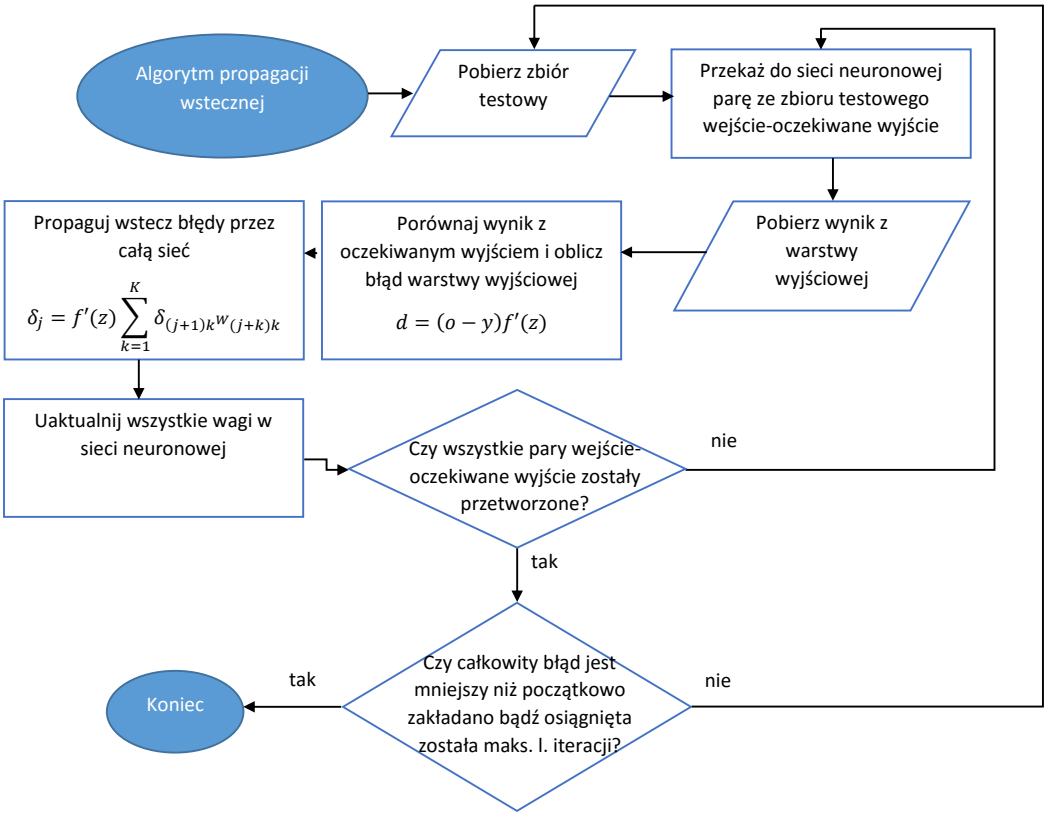
Rys. 5.6: Mapa cech elementu. Wiadomo, że element zawiera cechę „Aktor" o wartościach „Julia Roberts, Al Pacino" oraz cechę „Reżyser" o wartości „Francis Ford Coppola". Element nie zawiera cechy „Aktor" o wartości „Brad Pitt" ani cechy „Reżyser" o wartości „Darren Aronofsky" więc w te miejsca wstawiane jest 0.

Propagacja wsteczna

Algorytm propagacji wstecznej jest jedną z popularniejszych metod uczenia nadzorowanego jednokierunkowych sieci neuronowych.

Operując na zbiorze testowym (zwykle stanowi on 80% wszystkich danych) algorytm kolejno porównuje oczekiwane wyniki z tymi uzyskanymi z sieci neuronowej. Każdy przebieg pętli uczenia kończy się obliczeniem błędu sieci neuronowej d.

$$d = (o - y)f'(z)\,, \tag{5.4}$$



Rys. 5.7: Algorytm propagacji wstecznej

[12][37]

Algorytm RPROP

Algorytm Genetyczny

5.3. Algorytmy hybrydowe

Algorytmy hybrydowe

5.4. Analiza złożoności i poprawności

Analiza złożoności i poprawności

Rozdział 6

Ocena eksperymentalna

6.1. Opis metody badawczej

6.1.1. Miara oceny

W celu zbadania jakości algorytmów zostały zastosowane miary oceny: średnia kwadratowa błędów (RMSE) i średni błąd bezwzględny (MAE).

Średnia kwadratowa błędów

Średnia kwadratowa błędów (ang. RMSE – *root mean square error*) jest często wykorzystywaną miarą służącą zmierzeniu różnicy pomiędzy wartościami przewidywanymi a rzeczywistymi (obserwowanymi).

RMSE jest stosunkowo dobrą miarą dokładności ale tylko w celu porównania różnych modeli dla tego samego zestawu danych. RMSE jest zależne od skali, zatem nie sprawdza się najlepiej w przypadku porównywania ze sobą różnych zmiennych [15].

Średnią kwadratową błędów wylicza się ze wzoru:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}\,, \tag{6.1}$$

gdzie

$\hat{y}_i$ to wartość przewidywana
y_i to wartość rzeczywista

Im niższa wartość RMSE tym bardziej zbliżone są wartości przewidywane do rzeczywistych, zatem tym lepszy jakościowo jest model.

Średni błąd bezwzględny

Inną miarą mierzenia jakości modeli predykcyjnych jest MAE (ang. *mean absolute error*). Podobnie jak RMSE miara ta jest zależna od skali, zatem najlepiej sprawdza się w działaniu na tym samym zestawie danych [15].

Średni błąd bezwzględny wylicza się ze wzoru:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|\hat{y}_t - y_t|\,, \tag{6.2}$$

gdzie

$\hat{y}_i$ to wartość przewidywana
y_i to wartość rzeczywista

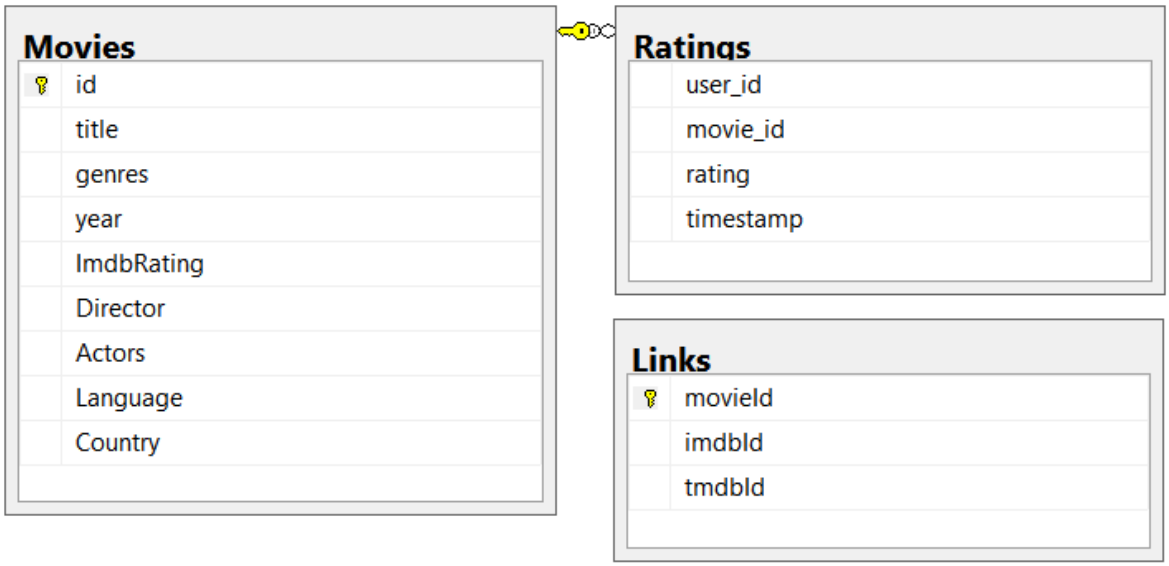
6.1.2. Zbiory danych

By uzyskać jak najbardziej miarodajne wyniki, badania zostały przeprowadzone na trzech różnych bazach danych z trzech różnych domen.

MovieLens

MovieLens [11] to baza zawierająca oceny filmów przez użytkowników portalu movielens.org. Baza zawiera 3706 filmów i 1000209 ocen wystawionych przez 6040 unikalnych użytkowników pomiędzy 25 kwietnia 2000 a 28 lutym 2003. Filmy oceniane są w skali od 1 do 5, gdzie 1 jest oceną najgorszą a 5 najlepszą.

Baza zawiera tabelę łączącą numery identyfikacyjne filmów z bazą IMDB.com. Korzystając z tego autorka pracy rozszerzyła oryginalną bazę filmów o informacje pobrane z IMDB.com. Ostateczny kształt bazy widoczny jest na rys. 6.1.



Rys. 6.1: Schemat bazy MovieLens

Yahoo Music

Opisać Yahoo Music

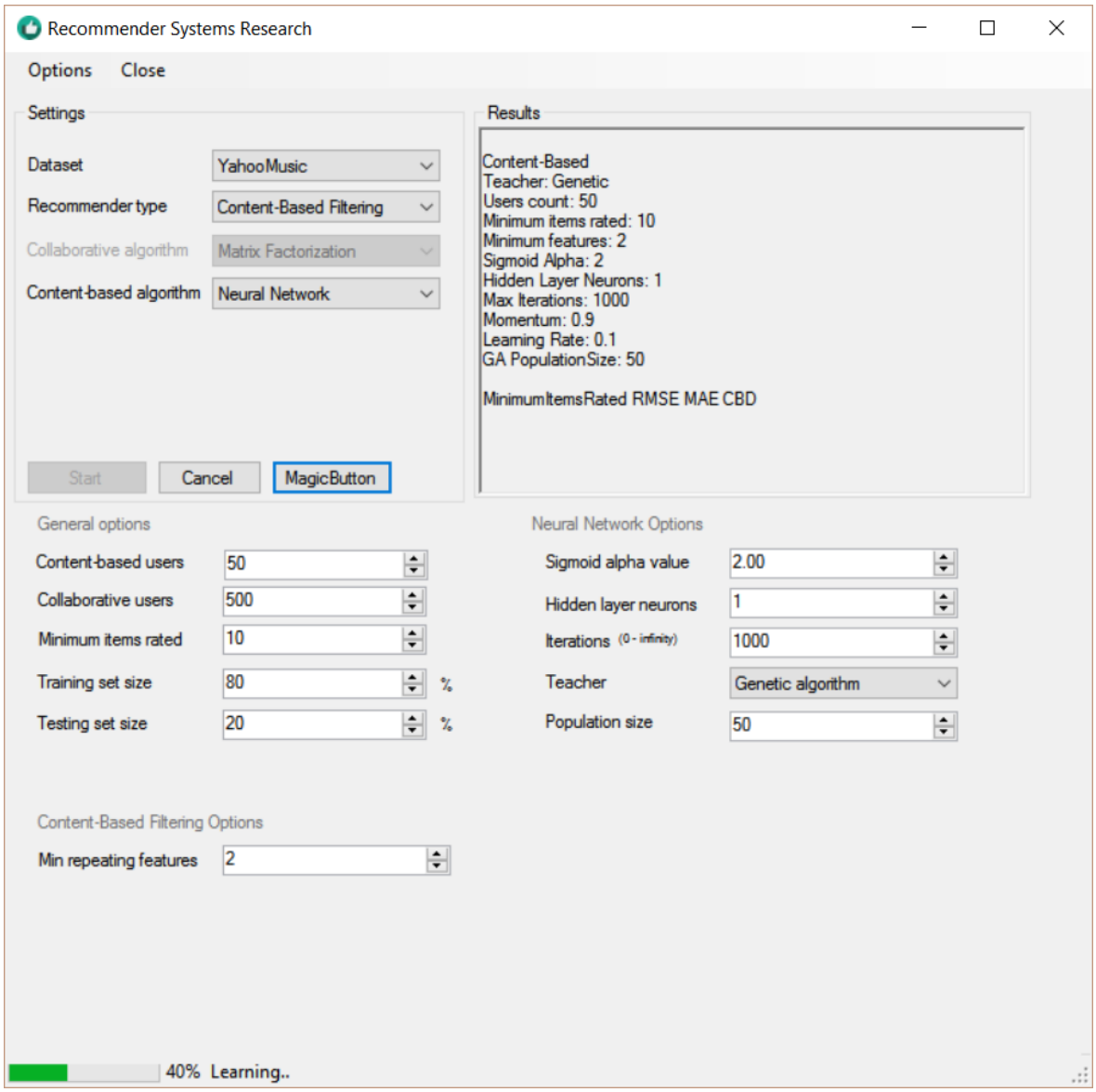
Amazon Meta

Opisać Amazon Meta

[21]

6.2. Środowisko symulacyjne

Rys. 6.2 przedstawia zrzut ekranu środowiska symulacyjnego stworzonego przez autorkę pracy na potrzeby przeprowadzenia badań algorytmów. Główny interfejs programu składa się z trzech części: ustawienia podstawowe, okno z wynikiem oraz panel sterujący ustawieniami zaawansowanymi.



Rys. 6.2: Zrzut ekranu środowiska symulacyjnego

W sekcji z ustawieniami podstawowymi możliwy jest wybór:

- bazy danych, która zostanie wykorzystana do pomiarów;
- typu algorytmu rekomendującego (content-based, collaborative bądź hybrydowy)
- w przypadku wyboru kolaboratywnego filtrowania lub filtrowania hybrydowego możliwy jest wybór typu algorytmu: Matrix Factorization, Biased Matrix Factorization lub SVD++.

- w przypadku wyboru filtrowania z analizą zawartości lub filtrowania hybrydowego automatycznie wybierany jest algorytm oparty na sieci neuronowej.

Widok ustawień zaawansowanych zmienia się w zależności od wybranego filtrowania. W przypadku wyboru content-based istnieje możliwość regulowania parametrów sieci neuronowej. Zawsze istnieje możliwość regulowania kryteriów doboru zestawu danych.

Kryteria doboru zestawu danych są następujące:

- liczba użytkowników do pobrania do algorytmu content-based;
- liczba użytkowników do pobrania do algorytmu collaborative;
- minimum elementów, jakie zostały ocenione przez każdego pobranego użytkownika;
- stosunek wielkości zbioru treningowego do zbioru testowego (domyślnie 80%-20%).

Parametry sieci neuronowej podlegające regulacji to:

- Sigmoidalna wartość alfa;
- liczba neuronów w warstwie ukrytej;
- maksymalna liczba iteracji uczenia sieci neuronowej;
- nauczyciel sieci neuronowej: propagacja wsteczna, rprop (resilient backpropagation) lub algorytm genetyczny;
- w przypadku wyboru algorytmu genetycznego – rozmiar każdej kolejnej populacji;
- minimalna ilość powtórzeń danej cechy aby była brana pod uwagę w trakcie budowania rekomendacji.

6.3. Metodologia

Metodologia

6.4. Przeprowadzone eksperymenty

Przeprowadzone eksperymenty

Rozdział 7

Wnioski

Rozdział 8

CHAPTER 1

8.1. SECTION

Algorytm 4 Alghoritm 4

$T \leftarrow$ text under analysis
for each word $w \in T$ **do**
  $S_w \leftarrow FIND\_SENTIMENT(w)$
  if $S_w = POSITIVE$ **then**
    $Sentiment[POSITIVE] + +$
  else if $S_w = NEGATIVE$ **then**
    $Sentiment[NEGATIVE] + +$
  else
    $Sentiment[NEUTRAL] + +$
  end if
end for
return $\arg\max_x Sentiment[x]$

Rys. 8.1: Schema 1

¡GRAPHIC¿

8.2. Section 2

8.2.1. Subsection 1

Subsubsection 1
Definicja 1
Definicja - *pierwsza*

Dodatek A

Appendix 1

Spis rysunków

Spis wzorów

Spis algorytmów

Bibliografia

[1] About the Music Genome Project. `http://www.pandora.com/about/mgp`. Data dostępu: 2016-07-19.

[2] Allegro – korzystanie z systemu rekomendacji. `http://faq.allegro.pl/artykul/27613/korzystanie-z-systemu-rekomendacji`. Data dostępu: 2016-07-19.

[3] Celma O. *The Long Tail in Recommender Systems*, pages 87–107. Springer-Verlag Berlin Heidelberg, 2010.

[4] Cheng J., Liu Y., Zhang H., Wu X., Chen F. A new recommendation algorithm based on user's dynamic information in complex social network. *Mathematical Problems in Engineering*, 2015, 2015.

[5] Davidson J., Liebald B., Liu J., Nandy P., Van Vleet T., Gargi U., Gupta S., He Y., Lambert M., Livingston B. et al. The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296. ACM, 2010.

[6] Desrosiers C., Karypis G. *Recommender Systems Handbook*, chapter A Comprehensive Survey of Neighborhood-based Recommendation Methods, pages 107–144. Springer, New York Dordrecht Heidelberg London, 2010.

[7] Filmweb – najczęściej zadawane pytania. `http://www.filmweb.pl/help`. Data dostępu: 2016-07-19.

[8] Gantner Z., Rendle S., Drumond L., Freudenthaler C. Mymedialite recommender system library. `http://www.mymedialite.net/`. Data dostępu: 2016-09-05.

[9] Gantner Z., Rendle S., Freudenthaler C., Schmidt-Thieme L. Mymedialite: a free recommender system library. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 305–308. ACM, 2011.

[10] Gupta P., Goel A., Lin J., Sharma A., Wang D., Zadeh R. Wtf: The who to follow service at twitter. In *Proceedings of the 22nd international conference on World Wide Web*, pages 505–514. ACM, 2013.

[11] Harper F. M., Konstan J. A. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):19, 2016.

[12] Hertz J. A., Krogh A. S., Palmer R. G., Jankowski S. *Wstęp do teorii obliczeń neuronowych*. Wydawnictwa Naukowo-Techniczne, 1993.

[13] Huttner J. From Tapestry to SVD: A survey of the algorithms that power recommender system. Master's thesis, Haverford College Department of Computer Science, 05 2009.

[14] Huynh T., Hoang K. Modeling collaborative knowledge of publishing activities for research recommendation. In *International Conference on Computational Collective Intelligence*, pages 41–50. Springer, 2012.

[15] Hyndman R. J., Koehler A. B. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

[16] IMDb database statistics. `http://www.imdb.com/stats`. Data dostępu: 2016-07-19.

[17] Ji K., Sun R., Shu W., Li X. Next-song recommendation with temporal dynamics. *Knowledge-Based Systems*, 88:134–143, 2015.

[18] Kirillov A. AForge.NET framework. http://www.aforgenet.com/framework/. Data dostępu: 2016-09-05.

[19] Koren Y., Bell R. *Recommender Systems Handbook*, chapter Advances in Collaborative Filtering, pages 145–186. Springer, New York Dordrecht Heidelberg London, 2010.

[20] Koren Y., Bell R., Volinsky C. et al. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[21] Leskovec J., Adamic L. A., Huberman B. A. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5, 2007.

[22] Linden G., Smith B., York J. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1):76–80, 2003.

[23] Lops P., de Gemmis M., Semeraro G. *Recommender Systems Handbook*, chapter Content-based Recommender Systems: State of the Art and Trends, pages 73–100. Springer, New York Dordrecht Heidelberg London, 2010.

[24] Maleszka M., Mianowska B., Nguyen N. T. A method for collaborative recommendation using knowledge integration tools and hierarchical structure of user profiles. *Knowledge-Based Systems*, 47:1–13, 2013.

[25] Mymedialite: Example experiments. `http://www.mymedialite.net/examples/datasets.html`. Data dostępu: 2016-07-24.

[26] Netflix Prize (I tried to resist, but...). `https://www.snellman.net/blog/archive/2006-10-15-netflix-prize.html`. Data dostępu: 2016-07-08.

[27] Netflix Prize: forum. `http://www.netflixprize.com/community/viewtopic.php?id=1537`. Data dostępu: 2016-07-08.

[28] Netflix Prize Rankings. `http://www.hackingnetflix.com/2006/10/netflix_prize_r.html`. Data dostępu: 2016-07-08.

[29] Netflix Prize Rules. `http://www.netflixprize.com//rules`. Data dostępu: 2016-07-08.

[30] Pogue D. A Stream of Movies, Sort of Free. *The New York Times*, 2007.

[31] Ricci F., Rokach L., Shapira B. *Recommender Systems Handbook*, chapter Introduc-

tion to Recommender Systems Handbook, pages 1–35. Springer, New York Dordrecht Heidelberg London, 2010.

[32] Rohrmann T. Computing recommendations at extreme scale with apache flink. *http://data-artisans.com/computing-recommendations-at-extreme-scale-with-apache-flink*, 2015.

[33] Rubens N., Kaplan D., Sugiyama M. Active learning in recommender systems. In Kantor P., Ricci F., Rokach L., Shapira B., editors, *Recommender Systems Handbook*, pages 735–767. Springer, 2011.

[34] Sarwar B., Karypis G., Konstan J., Riedl J. Application of dimensionality reduction in recommender system-a case study. Technical report, DTIC Document, 2000.

[35] Schafer J., Frankowski D., Herlocker J., Sen S. *The Adaptive Web*, chapter Collaborative filtering recommender systems, page 291–324. Springer Berlin / Heidelberg, 2007.

[36] Sharma R., Singh R. Evolution of Recommender Systems from Ancient Times to Modern Era: A Survey. *Indian Journal of Science and Technology*, 9(20), 2016.

[37] Timothy M. Sieci neuronowe w praktyce. *WNT, Warszawa*, 1996.

[38] Willmott C. J., Matsuura K. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79–82, 2005.

[39] Zhang H.-R., Min F., He X., Xu Y.-Y. A hybrid recommender system based on user-recommender interaction. *Mathematical Problems in Engineering*, 2015, 2015.