RAPORT:

Opracowanie koncepcji systemu rekomendacji dla treści multimedialnych

Autor: Sebastian Ernst, Konrad Kułakowski

Centrum Inteligentnych Systemów Informatycznych Akademia Górniczo-Hutnicza im. Stanisława Staszica al. Mickiewicza 30, 30-059 Kraków
budynek C-2 pokój 426 tel.: 12 617 44 53 www.isi.agh.edu.pl **isi@agh.edu.pl**

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

Opracowanie koncepcji systemu rekomendacji dla treści multimedialnych

dr inż. Sebastian Ernst, dr Konrad Kułakowski
Katedra Informatyki Stosowanej AGH

1 Wprowadzenie

Systemy rekomendacyjne (ang. *recommender systems*) stosowane są od wielu lat [5, 19] w różnych klasach systemów informatycznych. Najczęstszym zastosowaniem systemów rekomendacyjnych jest określanie obiektów (np. produktów), które mogą być atrakcyjne bądź przydatne dla określonego klienta. Ponieważ rekomendacje najłatwiej wyznaczyć w oparciu o historię działań (np. zakupów) innych użytkowników, zastosowanie znajdują tu metody tzw. *filtrowania kolaboratywnego* (ang. *collaborative filtering*), opisane szerzej w sekcji 2.1. Inną metoda dokonywania selekcji rekomendacji opiera się wyłącznie na cechach (atrybutach) rekomendowanych obiektów [17]; jej zastosowanie jest jednak ograniczone ze względu na często spotykany brak wystarczająco dokładnych opisów ocenianych elementów.

Klasyczne systemy rekomendacyjne dzieli się na dwie kategoriie, w zależności od kryteriów wykorzystywanych do określenia rekomendacji:

1. Rekomendacje personalizowane dla użytkownika (ang. *user-based recommendations*), które wyznaczane są w oparciu o określenie grona użytkowników o cechach podobnych do użytkownika dla którego wyznaczane są rekomendacje.

2. Rekomendacje oparte o produkt (ang. *item-based recommendations*), które wyznaczane są wyłącznie w oparciu o cechy produktu – zbiór elementów rekomendowanych nie zależy w tym przypadku od użytkownika, dla którego jest on wyznaczany.

W obu podejściach, cechy te mogą być *jawne* (ang. *explicit*) bądź *niejawne* (ang. *implicit*):

- Cechy jawne to atrybuty, których wartości zostały określone dla danego użytkownika bądź elementu; w przypadku użytkowników może to być płeć, wiek, zawód, miejsce zamieszkania; dla elementów będą to wszelkie parametry użyte do ich opisu (np. gatunek filmu, kolor produktu).

- Cechy niejawne to dane o podobieństwie użytkowników bądź elementów, wywnioskowane z rejestru działań występujących w systemie; w przypadku

użytkowników może to więc być np. historia dokonywanych zakupów lub fakt wystawienia danych ocen określonym produktom; w przypadku elementów może to być np. fakt wystąpienia w jednym zamówieniu.

Należy podkreślić, że ustalenie kryteriów na podstawie których wyznaczane są rekomendacje jest działaniem o charakterze biznesowym. Leży więc ono po stronie operatora danego systemu – rola specjalistów może rozpocząć się dopiero po określeniu przez osoby decyzyjne rodzaju oraz kryteriów dla rekomendacji.

Osobnym zbiorem metod naukowych są tzw. metody *porównywania parami* (ang. *pairwise comparisons*). W ogólnym podejściu, metody te (opisane szerzej w sekcji 2.2) służą do wyznaczenia spójnego rankingu (szeregu ocen) dla elementów, które oceniane są w parach. Oceny te mogą charakteryzować się niespójnością, stąd więc konieczność zastosowania odpowiednich metod oraz algorytmów do uspójnienia tzw. macierzy porównań. Dodatkowym elementem badawczym była więc próba włączenia metod porównywania parami do systemu rekomendacyjnego; w sekcji 4 przedstawiono propozycje modułów integrujących te metody, których funkcjonalność wykracza poza ramy klasycznych systemów rekomendacyjnych.

Prowadzone w ramach projektu prace miały na celu:

- podsumowanie metod stosowanych do budowy systemów rekomendacyjnych, w tym wszystkich elementów składowych systemów filtrowania kolaboratywnego,
- podsumowanie stanu wiedzy dotyczącego metod porównywania parami,
- identyfikację gotowych rozwiązań software'owych wspomagających budowę systemów rekomendacyjnych,
- określenia propozycji funkcjonalności systemu rekomendacyjnego dla danych multimedialnych.

2 Charakterystyka stosowanych metod

W niniejszej sekcji podsumowano w zwięzły sposób metody proponowane do wykorzystania w projektowanym systemie.

2.1 Filtrowanie kolaboratywne

W ogólnym pojęciu, metody filtrowania kolaboratywnego mają na celu wybór podzbioru elementów przy wykorzystaniu danych pochodzących z różnych źródeł. Metody te są stosowane do bardzo dużych zbiorów danych, stąd istotną cechą jest ich skalowalność. Oprócz niej, przed systemami filtrowania kolaboratywnego stoi szereg wyzwań:

- rzadkie pokrycie danymi: macierz wiążąca np. użytkowników z produktami często jest wypełniona w bardzo niskim stopniu; może to utrudnić określenie atrakcyjności danego produktu dla danego użytkownika,

- problem „zimnego startu": rekomendacje dla nowych użytkowników lub dotyczące niedawno dodanych produktów mogą być w początkowym okresie niedostępne lub mało miarodajne,
- problem synonimów: w podejściach opartych o cechy jawne (zob. sekcja 1) zbliżone koncepcje mogą być określone różnymi terminami lub etykietami,
- „szare owce": użytkownicy, których działania nie są skorelowane z żadną grupą innych użytkowników; problem może dotyczyć również produktów,
- nierzetelne oceny: w systemach bez ograniczeń dotyczących wystawiania ocen przez użytkowników istnieje ryzyko zaburzenia rzetelności ocen.

Wśród podstawowych metod wykorzystywanych w systemach filtrowania kolaboratywnego znajdują się przede wszystkim metody należące do trzech kategorii:

1. metody znajdowania korelacji,
2. metody redukcji liczby wymiarów,
3. metody klasteryzacji i określania sąsiedztwa.

Metody te scharakteryzowano pokrótce w kolejnych sekcjach na przykładzie algorytmu algorytmu Eigentaste[1], stosowanego w doświadczalnym systemie rekomendacyjnym Jester [7].

2.1.1 Metody określania korelacji

Jednym z najistotniejszych kroków metody filtrowania kolaboratywnego jest określenie podobieństwa pomiędzy parami – zależnie od podejścia – użytkowników lub elementów. Pożądaną formą jest tu (symetryczna) macierz C, określającą podobieństwo pomiędzy określonymi dwoma elementami.

Metodą często stosowaną w systemach rekomendacyjnych jest zbudowanie macierzy z wartości współczynnika korelacji Pearsona (ang. *Pearson product-moment correlation coefficient*, PPMCC) [16] dla poszczególnych par. Jeżeli A jest macierzą zawierającą znormalizowane wartości ocen wystawianych n produktom przez użytkowników, macierz C można zdefiniować jako [7]:

$$C = \frac{1}{n-1} A^T A.$$

Innym sposobem budowy macierzy C jest zapisanie ocen wystawianych przez użytkowników jako n-wymiarowe wektory a następnie obliczenie kąta pomiędzy tymi wektorami [13].

[1] `http://eigentaste.berkeley.edu`

2.1.2 Metody redukcji wymiarowości

Aby określić „podobnych" użytkowników w otoczeniu danego użytkownika (lub uczynić to dla rekomendowanych elementów) użyteczne jest zrzutowanie użytkowników, których preferencje charakteryzują wielowymiarowe wektory ocen, na bardziej intuicyjną przestrzeń dwu- lub trójwymiarową. Jedną z najczęściej stosowanych metod jest analiza głównych składowych (ang. *principal component analysis*; PCA), która pozwala na zredukowanie liczby wymiarów przy zachowaniu najistotniejszych, charakterystycznych cech danych [9].

Dla macierzy C zdefiniowanej jak w sekcji 2.1.1 należy wyznaczyć macierz E (macierz wektorów własnych C) oraz macierz Λ (macierz wartości własnych C), takie że [7]:

$$C = E^T \Lambda E,$$

oraz

$$ECE^T = \Lambda.$$

2.1.3 Metody określania sąsiedztwa

Gdy użytkownicy bądź elementy są już zrzutowane na intuicyjną (często dwuwymiarową) przestrzeń, można wyznaczyć klasy ich podobieństwa. Typowo wykorzystywane są do tego algorytmy klasteryzacji, m.in.:

- rekurencyjny podział na prostokąty [7],
- algorytm centroidów [14],
- próbkowanie Gibbsa [**?**].

Granulacja i liczba klastrów powinna zostać dobrana tak, aby uzyskać równowagę pomiędzy dostępnością rekomendacji dla danego użytkownika/elementu a trafnością samych rekomendacji.

2.2 Porównywanie parami

Pierwszy przypadek wykorzystania formalnie zdefiniowanej metody porównania parami zwykło wiązać się z z osobą *Ramona Llulla* [2], żyjącego w XIII wieku katalońskiego zakonnika, misjonarza, alchemika i błogosławionego *Kościoła Katolickiego*. Wykorzystał on metodę porównywania parami do ulepszenia sposobu w jaki wyłaniano (głosowano na) przełożonych zakonnych. W późniejszych czasach metoda porównywania parami (ang. pairwise comparisons method, w skrócie *PC* method) była wielokrotnie wykorzystywana i de facto na nowo odkrywana. W szczególności markiz *de Condorcet* (XVIII wiek) [3], ponownie zaproponował wykorzystanie metody porównywania parami w procedurze wyborczej. W XIX wieku *Fechner* [6] dostrzegł znaczenie porównywania w parach dla wyznaczania i definiowania relacji pomiędzy zjawiskami, bodźcami, artefaktami

postrzeganymi przez ludzi. Dwudziestowiecznym kontynuatorem myśli *Fechnera* był *Thurstone* [20]. Rozwinął on metodę porównania parami nadając jej znaczenie ilościowe w kontekście skali pomiarowej. W chwili obecnej najbardziej wpływowym i rozpowszechnionym nurtem teoretyczno-aplikacyjnym w obrębie metody porównywania parami jest *AHP (Analytic Hierarchy Process)* [18] - zaproponowana przez Thomasa Saaty hierarchiczna, wielokryterialna metoda decyzyjna oparta o ilościową metodę porównywania parami.

U podstaw metody porównywania parami leży przekonanie o tym, że niektóre obiekty, pojęcia lub zjawiska prościej jest ocenić, i tym samym stworzyć ranking tych obiektów, porównując je najpierw parami, a dopiero potem zbiór porównań parowych uogólnić do całkowitego (totalnego) porządku w zbiorze wszystkich porównywanych obiektów. O słuszności tego przekonania łatwo się przekonać wchodząc choćby do sklepu AGD w poszukiwaniu telewizora. Od razu, my klienci, zostaniemy „zaatakowani" kilkudziesięcioma modelami różnych telewizorów, z pośród których wyłonić ten najbardziej dla nas odpowiedni jest wręcz niemożliwe. Nawet po zawężeniu ilości modeli do tych o najbardziej nam odpowiadającej przekątnej, marce, zakresie cenowym etc., pozostanie nam jeszcze całkiem spora grupa odbiorników z których wybór tego najlepszego wcale nie musi być rzeczą prostą. Zwykle jednak jeśli postawiono by przed nami wybór pomiędzy tylko dwoma konkretnymi egzemplarzami telewizorów, nie mieli byśmy problemu z wyborem tego z pary, który w naszym przekonaniu jest lepszy, bardziej nam pasuje. Zgodnie z metodą porównywania parami, najlepiej jest porównać wszystkie dostępne pary alternatyw określając który obiekt w każdym z porównań jest lepszy (i o ile lepszy) od drugiego, a następnie zsyntetyzować wyniki określając tym samym linearny porządek w zbiorze alternatyw. Pierwsza, najlepsza alternatywa jest zwykle tą na którą się zdecydujemy.

2.2.1 Ustalanie rankingu metodą porównywania parami

Danymi wejściowymi do metody porównywania parami jest macierz $M = [m_{ij}]$ porównań parowych obejmujących n obiektów, gdzie $m_{ij} \in \mathbb{R}_+$ oraz $i, j = 1, \ldots, n$. Elementy m_{ij} oraz m_{ji} oznaczają względną wartość preferencji oceniającego (eksperta) dla pary dwóch ocenianych pojęć c_i oraz c_j. Jeśli zatem zdaniem eksperta c_i jest dwukrotnie lepsze (dwukrotnie bardziej preferowane) niż c_j to wartość m_{ij} powinna wynieść 2, a wartość $m_{ji} = \frac{1}{2}$. Macierze $M = [m_{ij}]$ dla których zachodzi równość $m_{ij} = \frac{1}{m_{ji}}$ nazywane są (ang. reciprocal matrices) a sama własność określana jest jako (ang. reciprocity). Przyjmuje się, że macierze porównań parowych są (ang. reciprocal).

Syntezy zbioru porównań parowych danego macierzą M można dokonać na co najmniej kilka sposobów. Najbardziej popularne podejście EVM (ang. eigenvalue method), proponowane przez Saaty [18], oparte jest o wyznaczenie wektora wartości własnych związanego z promieniem spektralnym macierzy M, tu zwanego tu po prostu największą wartością własną (ang. principial eigenvalue).

Niech macierz M:

$$M = \begin{bmatrix} 1 & m_{12} & \cdots & \cdots & m_{1n} \\ \vdots & 1 & \cdots & \cdots & \vdots \\ \vdots & \cdots & \ddots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ m_{\mathrm{n1}} & \cdots & \cdots & m_{n,n-1} & 1 \end{bmatrix}$$

będzie wynikiem pracy ekspertów oceniających n alternatyw $c_1, \ldots, c_n$ względem wybranego przez siebie kryterium. Zgodnie z podejściem Saaty'ego [18], $w^{(max)} = \left[w_1^{(max)}, \ldots, w_n^{(max)}\right]$ niech to wektor własny macierzy M stojący przy największej wartości własnej tej macierzy, tj. spełniający równanie:

$$Mw^{(max)} = \lambda_{max} w^{(max)}$$

gdzie λ_{max} to największa wartość własna macierzy M (promień spektralny macierzy M).

Wektorem wyznaczającym ranking pojęć będziemy nazywać wektor $w^{(ev)} = [w_1^{(ev)}, \ldots, w_n^{(ev)}]$ przeskalowany tak by suma jego składowych wynosiła 1. Innymi słowy każda składowa $w_i^{(ev)}$ wektora $w^{(ev)}$ będzie dana wzorem:

$$w_i^{(ev)} \stackrel{df}{=} \frac{w_i^{(max)}}{\sum_{j=1}^{n} w_j^{(max)}}$$

Takie przeskalowanie umożliwia efektywną ocenę pojęć nie tylko w obrębie jednego rankingu, ale także rozważanie popularności (istotności) danego pojęcia w kontekście różnych rankingów.

Warto zauważyć, że zgodnie z twierdzeniem *Frobeniusa-Perrona* dla macierzy $M = [m_{ij}]$ gdzie $m_{ij} \in \mathbb{R}_+$ istnieje taki wektor $w^{(ev)}$ który jest rzeczywisty i dodatni [15, rozdz. 8].

Inną popularną metodą syntezy wyników porównań parowych jest podejście GMM (ang. *geometric mean method*). Polega ono na wykorzystaniu średnich geometrycznych wierszy macierzy M [4] jako wartości rankingowych dla poszczególnych pojęć będących przedmiotem ewaluacji. W tym podejściu składowe wektor wag $w^{(gm)} = \left[w_1^{(gm)}, \ldots, w_n^{(gm)}\right]$ zadane są następująco:

$$w_i^{(gm)} \stackrel{df}{=} \frac{p_i}{\sum_k^n p_k} \quad \text{gdzie} \quad p_i \stackrel{df}{=} \left(\prod_{j=1}^{n} m_{ij}\right)^{\frac{1}{n}}$$

2.2.2 Problem niespójności (niezgodności)

Wynikiem porównania dwóch pojęć c_i oraz c_j jest liczba rzeczywista m_{ij} mająca za zadanie odzwierciedlić proporcję opisującą stosunek wartości preferencji

przypisanych do c_i oraz c_j. Oznaczmy wartość preferencji dla pojęcia c_i przez $w(c_i)$ i odpowiednio wartość preferencji przypisanej do pojęcia c_j jako $w(c_j)$. Liczba m_{ij} stara się zatem wyrażać proporcje

$$m_{ij} \approx \frac{w(c_i)}{w(c_j)}$$

Wielką zaletą, ale równocześnie i słabością metody porównywania parami jest niezależność poszczególnych porównań. Oceniając jedną parę, ekspert skupia się tylko na tym porównaniu, nie oceniając otrzymanego wyniku w kontekście innych porównań. Może się zatem zdarzyć (i najczęściej w praktyce tak właśnie się zdaża), że

$$m_{ij} \cdot m_{jk} \neq m_{ik}$$

W takiej sytuacji mówimy, że macierz porównań parowych M jest niespójna (niezgodna). Oczywiście istnieje spora różnica pomiędzy sytuacją w której różnica pomiędzy $m_{ij} \cdot m_{jk}$ i m_{ik} jest niewielka, a sytuacją w której ta różnica jest spora. W pierwszym przypadku skłonni byli byśmy ew. błąd przypisać ludzkiej omylności i nadal ufać w osąd eksperta (bądź ekspertów) którzy przygotowali zbiór ocen dany macierzą M. W drugim przypadku zaczęli byśmy się zastanawiać, czy w przypadku tak dużej różnicy w ocenie porównywanych pojęć nadal można tym ocenom ufać.

Obserwacja ta skłoniła wielu badaczy do przedstawienia tzw. indeksów spójności/niespójności macierzy M pozwalających oszacować stopień jej niespójności . Najbardziej rozpowszechnionym indeksem niespójności macierzy PC jest CI (od ang. consistency index) indeks *Saaty'ego* [18]:

$$CI \stackrel{df}{=} \frac{\lambda_{max} - n}{n-1}$$

Przykładem innego indeksu określającego spójność macierzy PC jest indeks *Koczkodaja* [10]:

$$\mathscr{K} = \max_{i,j,k \in \{1,\ldots,n\}} \left\{ \min \left\{ \left| 1 - \frac{m_{ij}}{m_{ik} m_{kj}} \right|, \left| 1 - \frac{m_{ik} m_{kj}}{m_{ij}} \right| \right\} \right\}$$

Przegląd innych indeksów niespójności wraz z analizą ich własności można znaleźć w [1]. Wspólną cechą indeksów niespójności jest osiąganie przez nie wartości 0 dla macierzy spójnych tj. takich dla których $m_{ij} \cdot m_{jk} = m_{ik}$, dla $i, j, k = 1, \ldots, n$.

2.2.3 Ustalanie rankingu z wykorzystaniem wartości referencyjnych

Nie zawsze trzeba cały ranking tworzyć od początku. Czasem wartości preferencji dla niektórych ocenianych pojęć mogą być znane wcześniej (mogą być na przykład wynikiem innego rankingu niekoniecznie sporządzonego w oparciu o metodę porównywania parami). W takiej sytuacji wygodnie jest wykorzystać podejście *HRE* (and. *Heuristic Ranking Estimation*) pozwalające w ramach metody porównywania parami skorzystać z wcześniej istniejących danych [11, 12].

Oznaczmy przez $C_U = \{c_1, \ldots, c_k\}$ zbiór pojęć (ang. concepts), których ranking musimy stworzyć a przez $C_K = \{c_{k+1}, \ldots, c_n\}$ zbiór pojęć referencyjnych, dla których wartość funkcji preferencji w jest znana. W celu wyznaczenia wartości $w(c_1), \ldots, w(c_k)$ potrzeba będzie rozwiązać układ równań:

$$Aw = b \tag{1}$$

gdzie macierz A przybiera postać:

$$A = \begin{bmatrix} 1 & -\frac{1}{n-1}m_{1,2} & \cdots & -\frac{1}{n-1}m_{1,k} \\ -\frac{1}{n-1}m_{2,1} & 1 & \cdots & -\frac{1}{n-1}m_{2,k} \\ \vdots & \vdots & \vdots & \vdots \\ -\frac{1}{n-1}m_{k-1,1} & \cdots & \ddots & -\frac{1}{n-1}m_{k-1,k} \\ -\frac{1}{n-1}m_{k,1} & \cdots & -\frac{1}{n-1}m_{k,k-1} & 1 \end{bmatrix}$$

a wektor b wyrazów wolnych jest następujący:

$$b = \begin{bmatrix} \frac{1}{n-1}m_{1,k+1}w(c_{k+1}) + \ldots + \frac{1}{n-1}m_{1,n}w(c_n) \\ \frac{1}{n-1}m_{2,k+1}w(c_{k+1}) + \ldots + \frac{1}{n-1}m_{2,n}w(c_n) \\ \vdots \\ \frac{1}{n-1}m_{k,k+1}w(c_{k+1}) + \ldots + \frac{1}{n-1}m_{k,n}w(c_n) \end{bmatrix}$$

Na ostateczny n-elementowy wektor wartość funkcji preferencji w składają się wartości: $w(c_1), \ldots, w(c_k)$ będące rezultatem obliczenia (1), oraz wcześniej znane wartości referencyjne $w(c_{k+1}), \ldots, w(c_n)$. Stad też po uzupełnieniu o wartości referencyjne wektor w (1) wygląda następująco:

$$w = [w(c_1), \ldots, w(c_k), w(c_{k+1}), \ldots, w(c_n)]^T$$

Czasami wygodnie jest przeskalować wektor w tak by suma jego kolejnych wartości wynosiła 1.

$$\widetilde{w} = \left[\frac{w(c_1)}{\sum_{i=1}^{n} w(c_i)}, \ldots, \frac{w(c_n)}{\sum_{i=1}^{n} w(c_i)}\right]^T$$

Inny sposób obliczenia wartości funkcji preferencji w przypadku gdy dla niektórych pojęć c_i jest ona zanana dostarcza podejście *GHRE* (ang. *geometric HRE*) [12]. Metoda ta wymaga rozwiązania liniowego układu równań:

$$\widehat{A}\widehat{w} = \widehat{b} \tag{2}$$

w którym macierz $\widehat{A}$ zadana jest następująco

$$\widehat{A} = \begin{bmatrix} (n-1) & -1 & \cdots & -1 \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ -1 & -1 & \cdots & (n-1) \end{bmatrix}$$

oraz

$$\widehat{b} = \begin{bmatrix} \sum_{j=1,j\neq 1}^{k} \widehat{m}_{1,j} + \widehat{g}_1 \\ \sum_{j=1,j\neq 2}^{k} \widehat{m}_{1,j} + \widehat{g}_2 \\ \vdots \\ \sum_{j=1,j\neq k}^{k} \widehat{m}_{1,j} + \widehat{g}_k \end{bmatrix}$$

przy oznaczeniach $\log_\xi w(c_j) \stackrel{df}{=} \widehat{w}(c_j)$, $\widehat{m}_{ij} \stackrel{df}{=} \log_\xi m_{ij}$ oraz $\widehat{g}_j \stackrel{df}{=} \log_\xi g_j$ dla pewnego $\xi \in \mathbb{R}_+$. Tym samym wyznaczenie wektora $\widehat{w}$ (2) jest równoznaczne obliczeniu wektora $w = \left[\xi^{\widehat{w}(c_1)}, \ldots, \xi^{\widehat{w}(c_k)}\right]^T$. Podobnie jak poprzednio wygodnie jest posługiwać się odpowiednio przeskalowanym wektorem wartości funkcji preferencji.

2.2.4 Podejście wielokryterialne

Istnieje wiele metod wielokryterialnego podejmowania decyzji [8]. Część z nich, np. *ELECTRE* [8] wykorzystuje metody porównywania parami. Przedstawione powyżej podejścia obliczania rankingu na podstawie porównań parowych charakterystyczne są dla metody *AHP* (ang. *Analytic Hierarchy Process*) zaproponowanej i rozwijanej przez *Thomasa Saaty'ego* [18].

W *AHP* najpierw wyodrębnia się cechy względem których będą porównywane obiekty, a po stworzeniu rankingów obiektów z uwagi na każdą z cech (rankingów częściowych), obliczana jest sumaryczna wartość funkcji preferencji dla każdego z ocenianych obiektów. Finalne obliczenie sumarycznej wartość funkcji preferencji dokonuje w oparciu o ranking kryteriów. Ranking kryteriów podobnie jak rankingi częściowe obliczane są z wykorzystaniem podejścia *EVM*.

3 Stan wiedzy: istniejące narzędzia

W niniejszej sekcji opisano gotowe narzędzia software'owe implementujące metody opisane w sekcji 2.

3.1 Filtrowanie kolaboratywne

W pionierskich czasach systemów rekomendacyjnych algorytmy implementowane były najczęściej od podstaw, jako logika aplikacji bądź w postaci procedur wbudowanych w systemach zarządzania bazami danych. Z czasem powstał szereg narzędzi do filtrowania kolaboratywnego, które różnią się zakresem oraz architekturą. Do najistotniejszych należą:

- Apache Mahout[2],
- LensKit[3],
- easyrec[4].

[2] `http://mahout.apache.org`
[3] `http://lenskit.org`
[4] `http://www.easyrec.org`

W kolejnych sekcjach przedstawiono krótką charakterystykę poszczególnych narzędzi.

3.1.1 Apache Mahout

Apache Mahout jest skalowalną biblioteką, stosowaną przede wszystkim do uczenia maszynowego. Dostępne w pakiecie algorytmy obejmują m.in.:

- filtrowanie kolaboratywne,
- klasyfikację (m.in. sieci Bayesa, ukryte modele Markova),
- klasteryzację,
- redukcję wymiarowości.

Algorytmy te mogą być wykorzystane w postaci:

- bezpośrednio, jako metody Java,
- procedur MapReduce pakietu Hadoop[5],
- zadań Apache Spark[6],
- procedur H2O[7],
- procedur Apache Flink (alternatywnego do MapReduce środowiska uruchomieniowego)[8].

3.1.2 LensKit

LensKit jest stworzoną w języku Java biblioteką do tworzenia systemów rekomendacyjnych. Zawiera on API do budowania systemu rekomendacyjnego, jak również narzędzia do ewaluacji trafności rekomendacji oraz modułowe implementacje algorytmów wykorzystywanych w systemach rekomendacyjnych.

3.1.3 easyrec

W odróżnieniu od omawianych wcześniej narzędzi, easyrec jest stworzoną w języku Java „wolnostojącą" aplikacją webową, wykorzystującą relacyjny system zarządzania bazami danych MySQL, służący do rozszerzenia funkcjonalności istniejących aplikacji o moduł do generowania spersonalizowanych rekomendacji. Komunikacja z aplikacją „główną" następuje za pośrednictwem usług sieciowych opartych o REST[9]

[5] `http://hadoop.apache.org`
[6] `https://spark.apache.org`
[7] `http://0xdata.com`
[8] `http://flink.apache.org`
[9] Representational State Transfer – metoda integracji modułów oprogramowania poprzez „lekką" komunikację opartą o protokół HTTP, wykorzystującą zgodnie z semantyką jego metody i używającą formatu JSON do serializacji wymienianych danych; więcej informacji: `http://en.wikipedia.org/wiki/Representational_state_transfer`.

3.2 Porównywanie parami – podejście AHP

Rynek oprogramowania wspomagającego decyzje w oparciu o porównywania parami można podzielić na oprogramowanie komercyjne i darmowe. To drugie często ma charakter naukowo-badawczy. Do pierwszej kategorii zaliczyć trzeba:

- MakeItRational AHP Software (http://makeitrational.com)
- AHP Project (http://www.ahpnet.com/home.aspx)
- Transparent Choice (http://www.transparentchoice.com)

Przykłady narzędzi niekomercyjnych to:

- Pairwise Comparisons Mathematica Package (https://code.google.com/p/pairwise-comparisons/)
- AHP Solver (http://sourceforge.net/projects/ahpsolver/)
- Analytic Hierarchical Process.NET (http://www.kniaz.net/software/ahp.aspx)
- Priority Estimation Tool (AHP) (http://sourceforge.net/projects/priority/)

Narzędzia komercyjne zapewniają zwykle dość podstawową funkcjonalność zapewniającą jednakże:

- graficzny interfejs użytkownika pozwalający na stosunkowo łatwe definiowanie hierarchii kryteriów
- weryfikacje spójności (zgodności) zbioru porównań
- obsługę wielu użytkowników/ekspertów/decydentów w ramach jednego procesu decyzyjnego
- analizę wrażliwości otrzymanego wyniku (ang. sensitivity analysis)

Możliwość nie wypełniania całej macierzy porównań parowych a tylko jej części również wymienia się jako pożądaną funkcjonalność dobrego oprogramowania tej klasy[10].

Oprogramowanie niekomercyjne często jest tworzone w ramach prowadzonych prac badawczych. Wtedy jego celem jest weryfikacja nowych hipotez dotyczących sposobu obliczania priorytetów ewaluowanych pojęć, agregacji wyników pochodzących od różnych ekspertów etc. Stąd też jego wykorzystanie przez użytkownika nieznającego dobrze metody porównywania parami może być utrudnione a nawet niebezpieczne.

[10] http://blog.transparentchoice.com/analytic-hierarchy-process/5-must-have-features-for-effective-and-intuitive-ahp-software

4 Koncepcja proponowanego rozwiązania

Opracowana koncepcja ma na celu opracowanie architektury systemu rekomendacyjno-rankingowego dla portalu internetowego służącego do dystrybucji treści multimedialnych, przede wszystkim filmów.

Ze względu na dużą liczbę dostępnych w systemie obiektów oraz wysoką aktywność użytkowników, kluczowym elementem będzie *skalowalność* systemu. Z tego względu pożądanym jest wykorzystanie systemu o rozproszonej architekturze, skalowalnego horyzontalnie.

Spośród omówionych w sekcji 3 narzędzi wspomagających filtorwanie kolaboratywne, najbardziej obiecującym wydaje się Apache Mahout, ze wzgędu na:

- wsparcie dla systemu Hadoop, który jest wykorzystywany m.in. jako repozytorium i narzędzie do przetwarzania plików multimedialnych,
- dostępność zarówno gotowych narzędzi do budowy systemów rekomendacyjnych, jak i implementacji składowych metod wykorzystywanych do filtrowania kolaboratywnego,
- dużą aktywność grupy rozwijającej to rozwiązanie.

W drodze analizy wypracowano następujące funkcjonalności, przekładające się na moduły docelowego systemu:

1. Moduł rekomendacji bezpośrednich. Moduł wykorzystuje „klasyczne" podejście do filtrowania kolaboratywnego, w następujących scenariuszach:
   (a) rekomendacje obiektów dla użytkowników, w oparciu o ich cechy określone w sposób jawny (wiek, lokalizacja, płeć, itd.),
   (b) rekomendacje obiektów dla użytkowników, w oparciu o historię ich zachowań,
   (c) rekomendacje obiektów niezależną od użytkownika, w oparciu o częstotliwość występowania par produktów w tym samym projekcie.
2. Moduł ocen porównawczych, zbierający od użytkowników oceny polegające na porównaniu dwóch obiektów, udostępniony użytkownikowi na etapie decydowania o zakupie jednego z prooponowanych obiektów.
3. Moduł do generowania rankingów, pozwalający na zbudowanie obiektywnego rankingu na podstawie subiektywnych ocen wystawionych przez różnych użytkowników. Jako wejście do modułu mają być oceny wystawiane przez użytkowników (np. w postaci „gwiazdek") poszczególnym obiektom.
4. Moduł do wyznaczania rekomendacji na podstawie porównań, stosujący metody filtrowania kolaboratywnego dla danych zebranych przez moduł ocen porównawczych.

5 Podsumowanie

W ramach projektu przeprowadzono analizę stanu wiedzy pozwalającej na implementację systemów rekomendacyjnych, ze szczególnym uwzględnieniem dwóch metod:

1. filtrowanie kolaboratywne,
2. porównywanie parami.

Przeprowadzono analizę sposobu działania ww. metod, a także narzędzi pozwalających na implementację tego typu systemów. W oparciu o jej wyniki zbudowano wizję i szkic architektury systemu rekomendującego treści multimedialne, wraz z określeniem modułów funkcjonalnych przydatnych w tym zastosowaniu.

Dodatkowym efektem przygotowania koncepcji jest nawiązanie współpracy z operatorem portalu Vpuzzler.com, który wykazał zainteresowanie wdrożeniem systemu rekomendacyjnego w tym serwisie. Powinno to umożliwić pozyskanie rzeczywistych danych, co stanowi podstawę odpowiedniego „dostrojenia" systemu i pozwoli na praktyczną ewaluację proponowanych metod.

References

[1] M. Brunelli, L. Canal, and M. Fedrizzi. Inconsistency indices for pairwise comparison matrices: a numerical study. *Annals of Operations Research*, 211:493–509, February 2013.

[2] J. M. Colomer. Ramon Llull: from 'Ars electionis' to social choice theory. *Social Choice and Welfare*, 40(2):317–328, October 2011.

[3] M. Condercet. Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. Paris: Imprimerie Royale, 1785.

[4] G. B. Crawford. The geometric mean procedure for estimating the scale of a judgement matrix. *Mathematical Modelling*, 9(3–5):327 – 334, 1987.

[5] Sebastian Ernst, Dominik Pacewicz, and Radoslaw Klimek. Recommendation Systems: Prediction of Web Site User Preferences. In *CMS'05 - Computer Methods and Systems*, pages 533–538, 2005.

[6] G. T. Fechner. *Elemente der Psychophysik*. Breitkopf und Härtel, Leipzig, 1860.

[7] Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A Constant Time Collaborative Filtering Algorithm. *Information Retrieval*, 4(2):133–151, 2001.

[8] S. Greco, editor. *Multiple Criteria Decision Analysis: State of the Art Surveys*. Springer, 2005.

[9] H. Hotelling. Analysis of a complex of statistical variables into principal components., 1933.

[10] W. W. Koczkodaj. A new definition of consistency of pairwise comparisons. *Math. Comput. Model.*, 18(7):79–84, October 1993.

[11] K. Kułakowski. Heuristic Rating Estimation Approach to The Pairwise Comparisons Method. *Fundamenta Informaticae*, 133:367–386, 2014.

[12] K. Kułakowski, K. Grobler-Dębska, and J. Wąs. Heuristic rating estimation: geometric approach. *Journal of Global Optimization*, 2014.

[13] Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[14] J Macqueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:281–297, 1967.

[15] C. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM: Society for Industrial and Applied Mathematics, April 2000.

[16] Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2:559–572, 1901.

[17] Anand Rajaraman and Jeffrey D Ullman. Mining of Massive Datasets. *Lecture Notes for Stanford CS345A Web Mining*, 67:328, 2011.

[18] T. L. Saaty. A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology*, 15(3):234 – 281, 1977.

[19] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Analysis of recommendation algorithms for e-commerce. *Organization*, 5(1/2):158–167, 2000.

[20] L. L. Thurstone. The Method of Paired Comparisons for Social Values. *Journal of Abnormal and Social Psychology*, pages 384–400, 1927.