УДК 004.04 DOI: 10.31673/2412-4338.2022.028590

Березівський М.Ю., Москаленко Н.В., Гребень В.С.
Державний університет телекомунікацій, Київ

ВИБІР ОПТИМАЛЬНОГО ПІДХОДУ ДО ПОБУДОВИ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ НА ОСНОВІ ДАНИХ ПРО ФІЛЬМИ

Анотація. *Останнім часом системи рекомендацій стають все більш популярними, оскільки вони можуть вирішити проблему перевантаження інформацією, пропонуючи елементи, які цікавлять користувачів. Обсяг інформації з кожним роком збільшується, а можливостей для Інтернет-бізнесу все більше, тому через Інтернет стає легше отримати що завгодно, наприклад товари, книги, фільми, новини. Роботи в області створення рекомендаційних систем можуть носити як комерційний, так і дослідницький характер і вимагати вирішення ряду проблем. У статті досліджено проблему вибору оптимального підходу до побудови системи рекомендацій на основі наявних даних. Неможливо створити систему рекомендацій без даних. Дані зазвичай доступні в явному або неявному вигляді. Дані, які називаються відвертими, можна зібрати шляхом пошуку відгуків і обміну думками користувачів про різні продукти. Однак неявні дані пов'язані з пошуковим журналом і історією даних, доступною в системі. Вибір оптимального підходу до побудови системи рекомендацій дозволить використовувати алгоритм машинного навчання. Змодельована архітектура системи рекомендацій, яка допомагає зрозуміти, як створити визначений користувачем процес рекомендацій. Рекомендаційні системи класифікуються за способом відбору необхідного матеріалу для системи користувача. В основному застосовуються два основні підходи: спільна фільтрація та фільтрація, орієнтована на вміст. Існує також гібридна фільтрація, яка поєднує як спільну, так і орієнтовану на вміст фільтрацію. У роботі проведено аналіз та порівняння рекомендаційних систем за типами. Порівняння типів систем рекомендацій за основними характеристиками: складність впровадження, точність впровадження, швидкість роботи, залежність від користувачів системи. Проведено аналіз основних алгоритмів машинного навчання контент-орієнтованої фільтрації.*

Ключові слова: *система рекомендацій, спільна фільтрація, контент-орієнтована фільтрація, гібридна фільтрація, алгоритм.*

Berezivskyi M.Yu., Moskalenko N.V., Greben V.S.
State University of Telecommunications, Kyiv

CHOOSING THE OPTIMAL APPROACH TO BUILDING A RECOMMENDER SYSTEM BASED ON MOVIE DATA

Abstract. *Recommender systems have become increasingly popular recently because they can address the problem of information overload by suggesting items of interest to the users. The volume of information with each year increases and more opportunities for Internet business, so it becomes easier to get anything through the internet, such as goods, books, movies, news. Work in the field of creating recommendation systems can be both commercial and research in nature, and need to address a number of issues. The paper studies the problem of choosing the optimal approach to building a recommendation system by repulsive from the available data. It is not possible to create a recommendation system without data. The data is usually available in an explicit or implicit way. Data that is called explicit can be collected by finding reviews and sharing of user views about*

different products. However, the implicit data is related to the search magazine and the data history available on the system. Choosing an optimal approach to building a recommendation system will allow the machine learning algorithm to be used. Simulated recommendation system architecture that helps to understand how to build a user-defined recommendation process. Recommender systems are classified in the manner of sampling the necessary material for the user system. Two basic approaches are primarily applied: collaborative filtering and content-oriented filtering. There is also a hybrid filtering that combines both collaborative and content-oriented filtering. The work carried out analysis and comparison of recommendation systems by types. Comparing types of recommendation system based on basic characteristics: implementation complexity, implementation accuracy, job speed, dependency on system users. Analysis of the core content-oriented filtering machine learning algorithms have been conducted.

Key words: *recommendation system, collaborative filtering, content-oriented filtering, hybrid filtering, algorithm.*

1. Постановка проблеми.

Обсяги інформації з кожним роком збільшується та з'являються більше можливостей для інтернет-бізнесу, отже стає легше отримати будь-що через інтернет, наприклад, товари, книги, фільми, новини. Широкий вибір варіантів може бути дуже спокусливим для користувачів, але доведено, що вони стають менш мотивованими купувати продукт пізніше. Рекомендаційні системи служать вирішенням цієї проблеми, допомагаючи користувачам знаходити предмети, які їх зацікавлять, використовуючи поведінку користувачів, наприклад, лайків чи перегляду інших предметів. Зараз, майже всі найбільші компанії світу, такі як, Google, Amazon, Facebook і Netflix, використовують рекомендаційні системи у широкому спектрі програм, такі як реклама, рекомендації продуктів і новин**.** Проблема оптимального вибору фільму виникала у кожного адже всім користувачам хочеться менше витрачати часу та обрати найбільше цікавий фільм.

2. Аналіз останніх досліджень і публікацій.

Сьогодні доступна велика кількість рекомендаційних систем, побудованих відповідно до певного підходу. Формування системи рекомендацій починається з інформаційного аналізу елементів і користувачів, за яким виконується створення моделі користувача. У цієї моделі зберігається інформація, оброблена шляхом аналізу інформації, після чого модель використовується для генерації рекомендацій [1].

3. Мета і задачі дослідження

Метою дослідження є виявлення оптимальних підходів для побудови рекомендаційної системи фільмів.

Для досягнення мети було вирішено наступні завдання, які розбито на етапи:

1. Архітектура рекомендаційної системи.
2. Порівняння рекомендаційних систем за типами.
3. Аналіз алгоритмів машинного навчання контент-орієнтованої фільтрації.

4. Результати дослідження.

Рекомендаційної системи складається з алгоритму і набору даних, як показано на рис. 1. Вибір типу рекомендаційної систем залежить від даних, які доступні розробнику такої системи. Для рекомендаційної системи фільмів можна виділити два типи даних: дані від користувачів системи та дані про конкретні фільми. Алгоритм рекомендації використовує набір даних для обчислення моделі. Після цього модель використовується для обчислення рекомендацій відповідним користувачам системи.

Рис. 1. Модель рекомендаційної системи

Рекомендаційні системи класифікують за способом відбору необхідного користувачеві матеріалу. В основному застосовується два базових підходи: колаборативна фільтрація і контентна фільтрація. Також існує гібридна фільтрація, яка поєднує в собі як колаборативну, так і контентну фільтрацію.

Система рекомендацій за допомогою колаборативної фільтрації.

Основна ідея алгоритмів колаборативної фільтрації [2] полягає в пропозиції нових фільмів для конкретного користувача на основі попередніх оцінок користувача. Ці алгоритми ґрунтуються на статистичних методах, щоб знайти групу користувачів близьких до цільового користувача. Цей підхід ще називають метод найближчих сусідів: використання попередніх оцінок, зроблених клієнтом, і аналіз оцінок інших користувачів, які мають подібні переваги. Тоді рекомендації (прогноз) для цільового користувача формуються на підставі обчислення якоїсь міри схожості за всіма накопиченими даними. Колаборативна фільтрація на основі подібності користувачів (User-based) має високу точність [3]. Однак, недоліком є ресурсомісткість (вимога до пам'яті) і складність (кількість обчислень, необхідну для отримання рекомендацій). До того ж обчислення ступеня близькості може проводитися тільки в реальному часі, тому даний метод може застосовуватися тільки до відносно невеликих баз даних.

Система рекомендацій за допомогою контентної фільтрації.

В рекомендаційних системах, що використовують контенту фільтрацію, користувачі не залежать від інших користувачів системи [4, 5]. Перевага контентної фільтрації полягає в тому, що для початку надання рекомендацій не потрібна велика кількість зареєстрованих користувачів, тобто рекомендації не залежать від інших користувачів системи. Також об’єкти

рекомендації можуть бути запропоновані одразу, оскільки всі дані від рекомендацій доступні відразу, отже швидкість досить висока. До недоліків можна віднести, що для побудови такої системи потрібно багато даних про об’єкт рекомендації, тому модель дуже залежна від повних даних, як будівля від правильно побудованого фундаменту.

Для даної моделі потрібно створювати профіль об’єкту рекомендації, який визначається параметрами, наприклад, для фільму актори і актриси, режисер, рік випуску, жанр, рейтинги IMDb тощо.

Гібридна система рекомендацій.

Гібридні системи поєднують колаборативну фільтрацію і контентну фільтрацію. Це дає змогу вирішити ряд проблем, які виникають при використанні цих методів окремо. В гібридній системі інформація про вподобання користувачів представлена, як перелік властивостей об'єкта та його оцінка користувачем [6]. Гібридна система значно покращує продуктивність роботи системи та дозволяє рекомендувати користувачеві не тільки об'єкти, які оцінили як позитивні, інші користувачі, а й ті об'єкти, які можуть сподобатись виходячи з його особистих переваг.

Розглянувши основні типи рекомендаційних систем можна зазначити, що для додатку для перегляду фільмів, можна використовувати всі, але потрібно відштовхуватися від даних, які ви можете використати, складності реалізації, точність реалізації, швидкості роботи, залежність від користувачів системи. У таблиці 1 представлено порівняння типів рекомендаційних систем за основними для них характеристиками.

Таблиця 1

Порівняння рекомендаційних систем за типами

	Контент-орієнтована фільтрація	Колаборативна фільтрація	Гібридна фільтрація
Проблема холодного старту	Відсутня	Присутня	Відсутня
Складність реалізації	Низька	Низька	Висока
Точність рекомендацій	Середня	Середня	Висока
Швидкість роботи	Висока	Середня	Висока
Залежність від користувачів системи	Відсутня	Присутня	Залежить від реалізіції
Специфіка роботи	Музика, кіно, магазини, інтернет-портали	Музика, кіно, магазини, інтернет-портали	Будь-яка область

Якщо постає питання створення додатку для перегляду та рекомендації фільмів, потрібно визначити, які дані маємо у системі, або які дані можна знайти на просторах

інтернету. Коли не достатньо інформації від користувачів системи, це означає, що для моделі колаборативної фільтрації не буде достатньо даних. Тому можна виділити контент-орієнтовану фільтрацію, яка дасть достатньо гарні рекомендації та швидкість роботи, а коли вже достатньо користувачів та достатньо даних від них, то слід поміркувати про створення гібридної чи колаборативної фільтрації для рекомендацій.

Як правило для фільтрації на основі змісту використовуються два популярні алгоритми: відстань векторів та класифікація.

Найбільше використовується алгоритм під назвою Cosine Similarity/Distance (відстань векторів), який використовується для підрахунку математичної подібності між двома векторами, у нашому випадку це подібність двох фільмів. Cosine Similarity [7] математично рахується за допомогою формули:

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

де A и B вказують на розташування об’єкта(точки) на відповідних векторах, ||A|| та ||B|| довжина векторів. На рис. 2 представлено графічне представлення подібності для двох векторів.

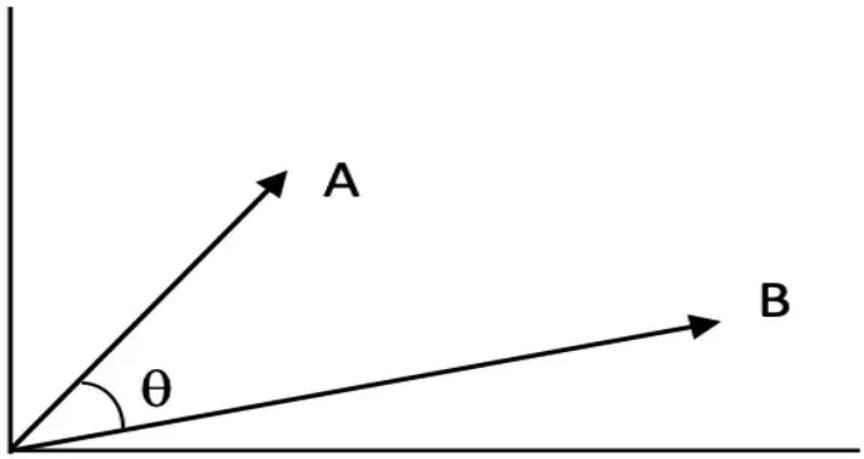

Рис. 2. Графічне представлення подібності

Але так як за допомогою цієї формули не можна прорахувати подібності ключових слів, імена знімальної команди, акторів та жанри фільму, тоді виникає потреба в перетворені слів у числове представлення. Тоді потрібно використовувати TF-IDF [8]. TF-IDF – це статистичний показник, що використовується для оцінки важливості слів у контексті документа, який є частиною колекції документів.

Другий метод за яким можна вирішити задачу фільтрації контенту для машинного навчання – Classification. Алгоритми класифікації, такі як баєсові класифікатори або моделі дерева рішень, можуть бути використані для того, щоб зробити рекомендації. Наприклад, кожен рівень дерева рішень може бути використаний, щоб відфільтрувати різні уподобання користувача, щоб зробити більш уточнений вибір, цей приклад можна представити на рис. 3.

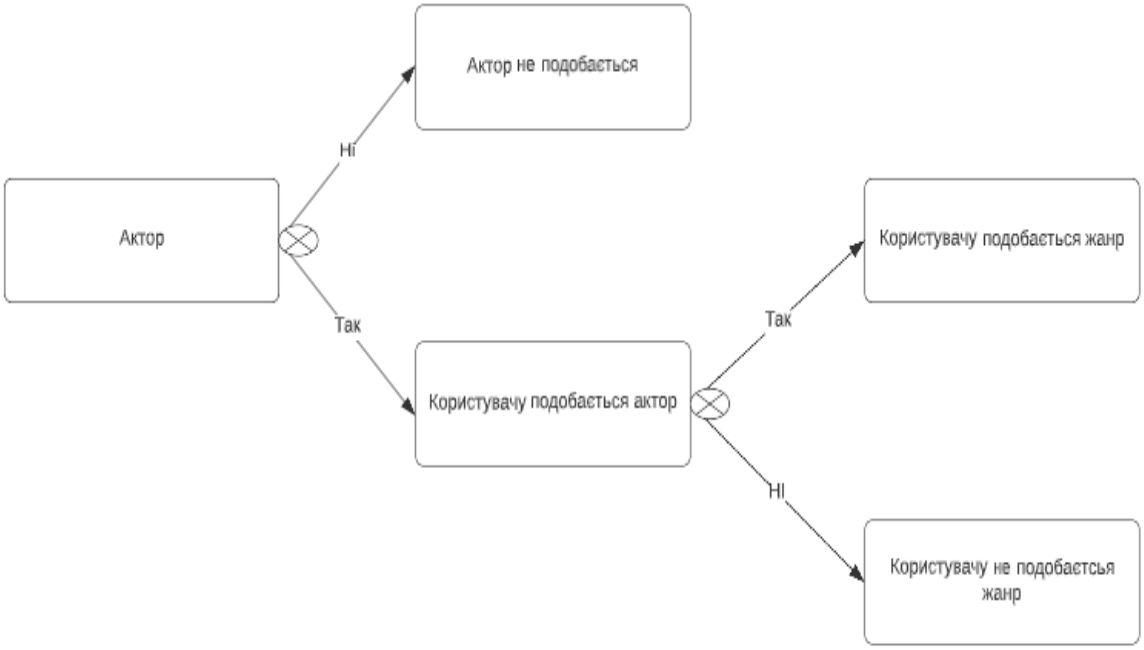

Рис. 3. Дерево рішення

Для задачі побудови рекомендаційної системи контент-орієнтованої фільтрації можна використовувати один з наведених алгоритмів машинного навчання . При виборі алгоритму слід проводити порівняльний аналіз на доступних даних.

Таким чином обрати потрібний алгоритм можливо за допомогою порівняння результатів роботи відповідної моделі, але дуже часто не має потреби створювати всі можливі варіанти моделей машинного навчання, а вибрати алгоритм, який показує гарні результати на подібних даних. Найчастіше для побудови рекомендаційної системи фільмів контент-орієнтованої фільтрації обирають Cosine Similarity [9].

5. Висновки і перспективи подальших досліджень.

В роботі досліджено оптимальний підхід до побудови рекомендаційної системи відштовхуючись від доступних даних. В роботі було проведено аналіз та порівняння рекомендаційних систем за типами. В результаті обрано контент-орієнтовану фільтрацію та досліджено алгоритми машинного навчання для цього типу. Застосування такої системи дозволить створити рекомендаційну систему використовуючи лише дані від об’єктів(фільмів), не втрачаючи якості.

Список використаних джерел

1. Yadav N., Kumar R., Singh A., Pal S. Diversity in Recommendation System: Cluster Based Approach. Hybrid Intelligent Systems. 2020. 113–122c.

2. John S. Breese, David Heckerman, and Carl Myers Kadie. Empirical analysis of predictive algorithms for collaborative filtering. CoRR, abs/1301.7363, 2013. 69c.

3. Liu Na, Ming-Xia Li, Qiu Hai-yang, and Hao-Long Su. A hybrid user-based collaborative filtering algorithm with topic model. Appl. Intell., 51(11):7946–7959, 2021.

4. Soumen Chakrabarti. Mining the web - discovering knowledge from hypertext data. Morgan Kaufmann, 2003.

5. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. Cambridge University Press, 2008.

6. Robin D. Burke. Hybrid recommender systems: Survey and experiments. User Model. User Adapt. Interact., 12(4):331–370, 2002.

7. Cosine similarity [Електроний ресурс] – Режим доступу : https://www.wikiwand.com/en/Cosine_similarity

8. Tf–idf[Електроний ресурс] – Режим доступу : https://en.wikipedia.org/wiki/Tf–idf

9. Roshan Bharti and Deepak Gupta. Recommending Top N Movies Using Content-Based Filtering and Collaborative Filtering with Hadoop and Hive Framework,2019.102-112c.

References:

1. Yadav N., Kumar R., Singh A., Pal S. Diversity in Recommendation System: Cluster Based Approach. Hybrid Intelligent Systems. 2020. 113–122c.

2. John S. Breese, David Heckerman, and Carl Myers Kadie. Empirical analysis of predictive algorithms for collaborative filtering. CoRR, abs/1301.7363, 2013. 69c.

3. Liu Na, Ming-Xia Li, Qiu Hai-yang, and Hao-Long Su. A hybrid user-based collaborative filtering algorithm with topic model. Appl. Intell., 51(11):7946–7959, 2021.

4. Soumen Chakrabarti. Mining the web - discovering knowledge from hypertext data. Morgan Kaufmann, 2003.

5. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. Cambridge University Press, 2008.

6. Robin D. Burke. Hybrid recommender systems: Survey and experiments. User Model. User Adapt. Interact., 12(4):331–370, 2002.

7. Cosine similarity [Electronic resource] – Online: https://www.wikiwand.com/en/Cosine_similarity

8. Tf–idf [Electronic resource] – Online: https://en.wikipedia.org/wiki/Tf–idf

9. Roshan Bharti and Deepak Gupta. Recommending Top N Movies Using Content-Based Filtering and Collaborative Filtering with Hadoop and Hive Framework,2019.102-112c.