Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies

Ákos Lévárdy

Odporúčacie systémy založené na AI

Bachelor thesis

October 14, 2024

Degree course: Informatics
Field of study: 9.2.1 Informatics
Place: FIIT STU, Bratislava
Supervisor: PaedDr. Pavol Baťalík

ANNOTATION

500 words

ANOTÁCIA

500 slov

DECLARATION OF OATH

I hereby declare upon my honour that I wrote this thesis single-handed with usage of quoted literature and based on my knowledge and professional supervision of my supervisor.

ACKNOWLEDGMENT

First and foremost, I would like to thank my supervisor for their invaluable guidance and support throughout the duration of this project.

Contents

1 Introduction **1**

2 Analysis **2**
2.1 Role of Recommendation Systems 5
2.2 Collaborative Filtering 6
2.3 Content-Based Filtering 9
2.4 Knowledge Graphs 12
2.5 Hybrid approach 12
2.6 Search Engines 13
2.7 Chosen Algorithms - Comparison 14
2.8 Difficulties related to recommendation systems 15
2.9 Performance - measures - Metrics 16

3 Specification of requirements **17**

4 Implementation **18**
4.1 Dataset 18

5 Conclusion **19**

List of Figures

List of Tables

List of Abbreviations

1 Introduction

As Internet and Web technologies continue to evolve rapidly, the amount of information available online has expanded excessively across sections such as e-commerce, e-government or e-learning. To help users navigate this vast sea of content, Recommender Systems (RS) have become fundamental. They are very effective tools for filtering out the most appropriate information any user would like to find. The primary focus of these recommendations is to predict if a specific user will be interested in the distinct items.

"Item" is the general term used to denote what the system recommends to users. A RS normally focuses on a specific type of item (e.g., CDs, or news) and accordingly its design, its graphical user interface, and the core recommendation technique used to generate the recommendations are all customized to provide useful and effective suggestions for that specific type of item. [1]

The main target of this project is to create a recommendation system that uses ..... (text, materials).

MORE TEXT - 1.5 - 2 pages together for INTRODUCTION

2 Analysis

Making decisions is not always easy. People are frequently presented with an overwhelming number of options when picking a product, a movie, or a destination to travel to, and each option comes with different levels of information and trustworthiness.
While there are many situations in which users know exactly what they are looking for and would like immediate answers, in other cases they are willing to explore and extend their knowledge [2].
The main purpose of recommendation systems is to predict useful items, select some of them and after comparing them, the system recommends the most accurate ones.
These Personalized recommendation systems are emerging as appropriate tools to aid and speed up the process of information seeking, considering the dramatic increase in big data [3]. They need to handle a large amount of textual data in order to accurately understand users' reading preferences and generate corresponding recommendations [4].
Because of this amount of detail from all of the items, recommendation systems are becoming increasingly important. They help reduce options and offer better suggestions for the user so that they will have a personalized list to select their favourite. Fast and efficient access to information is essential in any field of study. Information systems often deal with changing data over time. The term called Concept drift describes when sometimes the patterns or behaviors in the data change unexpectedly which affects how the system makes predictions [5].
The task to provide users with well chosen options for

products that fit their requirements and interests is very important in todays consumer society. The products are mostly supplied by inputs [6] , sometimes even matching the users distinct tastes.

When someone is trying to find a movie to watch, it would be hard for them to start searching without any starting options. After all a blank page and no suggestions to choose from might even make the user decide not to pick anything.

Recommending items can be done in a variety of ways. Several types of recommendation systems exist, and their methods of operation differ. These recommendation types can be divided into 3 main categories, which are Content-Based Filtering approaches (CB), Collaborative Filtering approaches (CF) and Hybrid approaches which are the combinations of the two.
Other categories also include Knowledge-Based , Context-Aware, Popularity-Based and Deep Learning-Based Recommendation.

Content-Based Filtering works in a way that it creates user profiles and suggests the individual items or products based on the users past choices with similar items. The items have various features and characteristics which connect them.
Collaborative Filtering relies more on preferences of other users and their behaviour. The point is that users who had similar interests before will have them again in the future for new items.
Knowledge-Graphs use a network of data where items are linked through their features. Showing how items

relate to one another and connecting them with more information and detail.
Hybrid methods try to combine the useful characteristics of both collaborative filtering and content-based filtering methods. They take into account both the users past preferences and the preferences of other people who might share the users taste.

2.1 Role of Recommendation Systems

asd

2.2 Collaborative Filtering

One of the most popular methods used for personalized recommendations is collaborative filtering. This method filters information from users, which means it compares users behaviour, interactions with items and data, item correlation and ratings from users.

It can perform in domains where there is not much content associated with items, or where the content is difficult for a computer to analyze - ideas, opinions etc. [7] Collaborative filtering can be divided into 2 methods which are "Memory-based" and "Model-Based" collaborative filtering. The first one relies on historical preferences, whereas the second method is based on machine learning models to predict the best options.

Memory-based CF
Recommender systems based on memory automate the common principle that similar users prefer similar items, and similar items are preferred by similar users [8].
Memory-based collaborative filtering, which can also be called Neighborhood-based is further divided into 2 basic types, which are:

- User-Based Collaborative Filtering
  - The main idea is that 2 completely distinct users who have an interest in a specific item and they rate this item similarly will probably be drawn to a new item the same way.
- Item-Based Collaborative Filtering
  - Calculates similarity between items, rather than users. The user will probably like a new item

which is similar to another item they were interested in before.

When trying to implement this type of recommendation system it is important to consider the key components, which are:

- Rating Normalization - adjusts individual user ratings to a standard scale by addressing personal rating habits. Using for example Mean-Centering or Z-Score Normalization.
- Similarity Weight Computation - helps to select reliable neighbors for prediction and deciding how much impact each neighbor's rating has. A lot of Similarity measures can be used, such as Correlation-Based Similarity, Mean Squared Difference or Spearman Rank Correlation.
- Neighborhood Selection - selects the most appropriate candidates for making predictions based on each unique scenario, eliminating the least likely ones to leave only the best options.

Model-based CF

Recommender systems based on models, also known as Learning-based methods, try to develop a parametric model of the relationships between items and users. These models can capture patterns in the data, which can not be seen in the previous recommendation type.

Model-based algorithms do not suffer from memory-based drawbacks and can create prediction over a shorter period of time compared to memory-based algorithms because these algorithms perform off-line computation for training. The well-known machine learning techniques

for this approach are matrix factorization, clustering, probabilistic Latent Semantic Analysis (pLSA) and machine learning on the graph [9].

Matrix Factorization
In its basic form, matrix factorization characterizes both items and users by vectors of factors inferred from item rating patterns. High correspondence between item and user factors leads to recommendations [10].
People prefer to rate just a small percentage of items, therefore the user-item rating matrix, that tracks the ratings people assign to various items, is frequently sparse. In order to deal with this sparsity, matrix factorization (MF) algorithms split the matrix into two lower-rank matrices: one that shows the latent properties of the items and another that reflects the underlying user preferences. These latent representations can be used to predict future ratings or complete the matrix's missing ratings after factorization [11].

Advantages and Disadvantages of CF
MORE TEXT - 0.5 page

It is important to mention that the effectiveness depends on the ratio of users and items. For example when trying to recommend songs, there are usually way more users than songs and generally, many users listened to the same songs or same genres. Which means like-minded users are found easily and the recommendations will be effective. On the other hand, in a different field, when it comes to recommending books or articles the systems deals with millions of articles but a lot less users. This

leads to less ratings on papers or no ratings at all, so it is harder to find people with shared interests [12].

2.3 Content-Based Filtering

Recommender Systems which are using content-based filtering, review a variety of items, documents and their details. Each product has their own description which is collected to make a model for each item. The model of an item is composed by a set of features representing its content. The main benefit of content-based recommendation methods is that they use obvious item features, making it easy to quickly describe why a particular item is being recommended. [13]
This also allows for the possibility of providing explanations that list content features that caused an item to be recommended, potentially giving readers confidence in the system's recommendations and insight into their own preferences [14].
These profiles for items are different representations of information and users interest about the specific item.
The recommendation process basically consists in matching up the attributes of the user profile against the attributes of a content object. [13]
There can also be side information about items, where this side information predominantly contains additional knowledge about the recommendable items, e.g., in terms of their features, metadata, category assignments, relations to other items, user-provided tags and comments, or related textual content. [15]

The process for recommending items using content-based filtering has 3 different phases:

- Content Analyzer - Turns the unstructured information (text) into structured, organized information using pre-processing steps which are basic methods in Information Retrieval, such as feature extraction.
- Profile Learner - Collects data of the users preference (feedback) that can be either positive information reffering to features which the active user likes or negative ones which the user does not like. After generalization it tries to construct user profiles for later use.
- Filtering Component - Matches the items for the user, based on the similarities between item representations and user profiles, meaning it compares the features of new items with features in user preferences that are stored in the users profile. [16]

The user modeling process has the goal to identify what are the users needs and this can be done 2 ways. Either the system calculates them from the interactions between the user and items through feedback or the user can specify these needs directly by giving keywords to the system, providing search queries [12].

Feedback

When trying to acquire helpful information or criticism that is given by the user there are 2 separate ways.
The first one is called Explicit Feedback where it is necessary for the user to give item evaluation or actively rate products. Most popular options are gathering like/dislike ratings on items or the ratings can be on a scale either from 1 to 5 or 1 to 10. After the ratings the user can also give comments on separate items.

The other way is called Implicit Feedback where the information is collected passively from analyzing the users activities. Some alternatives can be clicks on products, time spent on sites or even transaction history [16].

Advantages and Disadvantages of CB Filtering
MORE TEXT - 0.5 page

Semantics ?
- connect with CB and/or knowledge graph
Ontology ?
- describe it
- connect with semantics

2.4 Knowledge Graphs

Knowledge graph is a knowledge base that uses a graph-structured data model. It is a graphical databases which contains a large amount of relationship information between entities and can be used as a convenient way to enrich users and items information. [17]

MORE TEXT - DETAILS - 1.5 - 2 pages

2.5 Hybrid approach

- uses both the CF and the CB filtering, for more accuracy

MORE TEXT - 0.5 page

2.6 Search Engines

Search Engines have become crucial for navigating the vast amount of information available online. They make it possible for people to quickly look up solutions, learn new things, and browse the wide variety of resources available on the internet. Search engine optimization is now necessary to guarantee that search engines deliver relevant results, quick search times, and a top-notch user experience given the explosive growth of online information.

A search engine is essentially a software that finds the information the user needs using keywords or phrases. It delivers results rapidly, even with millions of websites available online. The importance of speed in online searches is highlighted by how even minor delays in retrieval can negatively affect users' perception of result quality. [18]

2.7 Chosen Algorithms - Comparison

When trying to choose which recommendation approach is the best, first it is important to know the use case for the specific system.
In the domain of scientific publications, where users are relatively few with respect to the available documents, information needs and interests easily change in an unpredictable way over time due to evolving professional needs, there is no advertising pushing new items, and the long tail of infrequently read articles may contain the so-called sleeping beauties, that are documents containing extremely relevant results, but that remain unknown to most researchers for a very long time. The Content-based approach does not require particular assumptions over the size and the activity of the user base. It does not penalize items that have less ratings or are less frequently consumed by many users as long as enough metadata are available, which even allows detailed explanations. These advantages over Collaborative Filtering techniques make this approach particularly attractive to the purpose of providing recommendation in the domain of scientific publications [19].
A study shows that more than half of the recommendation approaches applied Content-based filtering, when making recommendations for research papers and articles in libraries [12].

2.8 Difficulties related to recommendation systems

- cold-start problem
- data sparsity
- scalability
- bias and diversity
- privacy
- serendipity
- over-specialization problem can occur - CBF
- ...

- Exploration VS Exploitation

Both CB and CF approaches encounter significant challenges such as the Cold-Start Problem, Data Sparsity or Scalability. The Cold-Start Problem arises when making recommendations to new users and/or items for which the available information is limited. As a result, the recommendations offered in such cases tend to be of poor quality and lack usefulness. [20]

2.9 Performance - measures - Metrics

- recall rate
- root mean square error
- precision
- cumulative gain ?
- accuracy
- overall efficacy
- f1 - measure
- Normalized Discounted Cumulative Gain (NDCG)
- ...

3 Specification of requirements

4 Implementation

4.1 Dataset

5 Conclusion

References

[1] *Introduction to Recommender Systems Handbook*, pages 1–35. 2010. doi:10.1007/978-0-387-85820-3_1.

[2] Roi Blanco, Berkant Barla Cambazoglu, Peter Mika, and Nicolas Torzec. Entity recommendations in web search. 8219 LNCS(PART 2):33 – 48, 2013. doi:10.1007/978-3-642-41338-4_3.

[3] Khalid Haruna, Maizatul Akmar Ismail, Suhendroyono Suhendroyono, Damiasih Damiasih, Adi Cilik Pierewan, Haruna Chiroma, and Tutut Herawan. Context-aware recommender system: A review of recent developmental process and future research direction. 7(12), 2017. doi:10.3390/app7121211.

[4] Ke Yan. Optimizing an english text reading recommendation model by integrating collaborative filtering algorithm and fasttext classification method. 10(9), 2024. doi:10.1016/j.heliyon.2024.e30413.

[5] Yingying Sun, Jusheng Mi, and Chenxia Jin. Entropy-based concept drift detection in information systems. 290, 2024. doi:10.1016/j.knosys.2024.111596.

[6] Simon Philip, P.B. Shola, and Abari Ovye John. Application of content-based approach in research paper recommendation system for a digital library. *International Journal of Advanced Computer Science and Applications*, 5(10), 2014. doi:10.14569/IJACSA.2014.051006.

[7] Prem Melville, Raymond J. Mooney, and Ramadass Nagarajan. Content-boosted collaborative filtering for improved recommendations. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI-02)*, pages 187–192, Edmonton, Alberta, 2002.

[8] X. Ning, C. Desrosiers, and G. Karypis. *A comprehensive survey of neighborhood-based recommendation methods*. 2015. doi:10.1007/978-1-4899-7637-6_2.

[9] Mehrbakhsh Nilashi, Othman Ibrahim, and Karamollah Bagherifard. A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques. 92:507 – 520, 2018. doi:10.1016/j.eswa.2017.09.058.

[10] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. 42(8):30 – 37, 2009. doi:10.1109/MC.2009.263.

[11] Srilatha Tokala, Murali Krishna Enduri, T. Jaya Lakshmi, and Hemlata Sharma. Community-based matrix factorization (cbmf) approach for enhancing quality of recommendations. 25(9), 2023. doi:10.3390/e25091360.

[12] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. Research-paper recommender systems: a literature survey. 17(4):305 – 338, 2016. doi:10.1007/s00799-015-0156-0.

[13] *Content-based Recommender Systems: State of the Art and Trends*, pages 73–105. 2010. doi:10.1007/978-0-387-85820-3_3.

[14] Raymond J. Mooney and Loriene Roy. Content-based book recommending using learning for text categorization. page 195 – 204, 2000. doi:10.1145/336597.336662.

[15] Pasquale Lops, Dietmar Jannach, Cataldo Musto, Toine Bogers, and Marijn Koolen. Trends in content-based recommendation: Preface to the special issue on recommender systems based on rich item descriptions. 29(2):239 – 249, 2019. doi:10.1007/s11257-019-09231-w.

[16] M. De Gemmis, P. Lops, C. Musto, F. Narducci, and G. Semeraro. *Semantics-aware content-based recommender systems.* 2015. doi:10.1007/978-1-4899-7637-6_4.

[17] Saidi Imène, Klouche Badia, and Mahammed Nadir. Knowledge graph-based approaches for related entities recommendation. 361 LNNS:488 – 496, 2022. doi:10.1007/978-3-030-92038-8_49.

[18] Serge Stephane AMAN, Behou Gerard N'GUESSAN, Djama Djoman Alfred AGBO, and KONE Tiemoman. Search engine performance optimization: methods and techniques. 12, 2024. doi:10.12688/f1000research.140393.3.

[19] D. De Nart and C. Tasso. A personalized concept-driven recommender system for scientific libraries. volume 38, page 84 – 91, 2014. doi:10.1016/j.procs.2014.10.015.

[20] Malak Al-Hassan, Bilal Abu-Salih, Esra'a Alshdaifat, Ahmad Aloqaily, and Ali Rodan. An improved fusion-based semantic similarity measure for effective collaborative filtering recommendations. 17(1), 2024. doi:10.1007/s44196-024-00429-4.

[21] Hongbo Wang, Yizhe Wang, and Yu Liu. A sequential recommendation model for balancing long- and short-term benefits. 17(1), 2024. doi:10.1007/s44196-024-00460-5.

[22] Shilpa S. Laddha and Pradip M. Jawandhiya. Semantic search engine. 10(21):1–6, 2017. doi:10.17485/ijst/2017/v10i23/115568.

[23] Dirk Lewandowski. Understanding search engines. page 1 – 296, 2023.

[24] T. R. Mahesh, V. Vinoth Kumar, and Se-Jung Lim. Uscotc: Improved collaborative filtering (cfl) recommendation methodology using user confidence, time context with impact factors for performance enhancement. 18(3):e0282904, 2023. doi:10.1371/journal.pone.0282904.

[25] Tasnim M. A. Zayet, Maizatul Akmar Ismail, Sara H. S. Almadi, Jamallah Mohammed Hussein Zawia, and Azmawaty Mohamad Nor. What is needed to build a personalized recommender system for k-12 students' e-learning? recommendations for future systems and a conceptual framework. 28(6):7487 – 7508, 2023. doi:10.1007/s10639-022-11489-4.

[26] Ali Taleb Mohammed Aymen and Saidi Imène. Scientific paper recommender systems: A review. 361 LNNS:896 – 906, 2022. doi:10.1007/978-3-030-92038-8_92.

[27] Akhil M. Nair, Oshin Benny, and Jossy George. Content based scientific article recommendation system using deep learning technique. 204 LNNS:965 – 977, 2021. doi:10.1007/978-981-16-1395-1_70.

[28] Cataldo Musto, Pierpaolo Basile, Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. Introducing linked open data in graph-based recommender systems. 53(2):405 – 435, 2017. doi:10.1016/j.ipm.2016.12.003.

[29] Enrico Palumbo, Giuseppe Rizzo, and Raphaël Troncy. Entity2rec: Learning user-item relatedness from knowledge graphs for top-n item recommendation. page 32 – 36, 2017. doi:10.1145/3109859.3109889.

[30] Peng Yang, Chengming Ai, Yu Yao, and Bing Li. Ekpn: enhanced knowledge-aware path network for recommendation. 52(8):9308 – 9319, 2022. doi:10.1007/s10489-021-02758-9.

[31] Zhiyu Li, Yanfang Chen, Xuan Zhang, and Xun Liang. Bookgpt: A general framework for book recommendation empowered by large language model. 12(22), 2023. doi:10.3390/electronics12224654.

[32] Gediminas Adomavicius, Konstantin Bauman, Alexander Tuzhilin, and Moshe Unger. *Context-Aware Recommender Systems: From Foundations to Recent Developments*. 2022. doi:10.1007/978-1-0716-2197-4_6.

[33] Cataldo Musto, Marco de Gemmis, Pasquale Lops, Fedelucio Narducci, and Giovanni Semeraro. *Semantics and Content-Based Recommendations*. 2022. doi:10.1007/978-1-0716-2197-4_7.