Content-based vs Collaborative Recommendation Systems: A Brief Comparison

Tsirmpas Dimitris
Athens University of Economics and Business
Department of Informatics

June 28, 2023

Professor: Sarantopoulos Panagiotis
Athens University of Economics and Business
Department of Business Administration
Greece

Abstract

Recommendation systems have become omnipresent in our everyday lives and necessary for most corporations. However, different kinds of recommendation systems exist offering unique advantages, challenges and limitations. In this short report, we describe and compare two of the most popular such systems; Content-based and Collaborative filtering, as well as their shared weaknesses. Understanding the characteristics of the respective systems helps managers make informed decisions for their specific business requirements.

1 Introduction

Recommendation systems have become almost a necessity in the modern era of e-commerce. These systems, used from movies [6], to music [11, 8], news [13], tangible commercial products [12, 2], to even scientific articles [4], have proven extremely sought after in today's era of digitization, information and big data. Besides practical and monetary benefits, they also represent a constantly evolving area for the application of new statistical and machine-learning models, attracting the attention of much of the scientific community.

Traditionally, two of the most successful recommendation systems were **Content-based** (CB) and **Collaborative** filtering (CF). They both attempt to find new items to recommend to a user based on information on the items themselves and the user's recorded purchase (when applied to e-commerce) history. Where they differ, is the kind of information used, where CB systems look into the previously purchased items and CF into the users who have purchased similar items. While these systems were traditionally implemented using standard statistical procedures, the advent of neural networks has greatly impacted the capabilities, implementation and use of these systems.

In this short report we will compare the two types, outlining advantages, disadvantages and similarities. This report does not require any technical depth on the part of the reader, since we intend this to be a short primer for managers and consultants on recommendation systems.

2 Common challenges

Most recommendation systems face a number of generic challenges which are caused by the nature of the common problem they seek to solve. Some of the most important ones include:

- **The cold start problem**. Since recommendation systems are based on utilizing information about the user and the items he interacts with, when information about either the user, or the items themselves, or both, are unavailable (new or inactive users, new or unpopular products), the system struggles to make any recommendation [5].

- **Data sparsity**. A generalization of the cold start problem, where our system is unable to find patterns because of insufficient data, or because of the presence of too many user/item attributes (called *"the curse of dimensionality"* in formal statistic and machine learning circles). In the second case, as we consider more and more features to attach to our users and items, the amount of data we need to back our system increases exponentially. This can be seen in Figure 1.

- **Originality**. Recommendation systems have a tendency to prioritize "safe", popular and highly rated items over items which the user might have not thought of themselves. This inhibits our systems learning patterns from our users and limits their potential. Some authors [10] stress the difference between simply "novel" recommendations, and truly surprising ones.

- **Shilling attacks**. Since recommendation systems are mostly used for monetary benefit, there is a strong incentive for bad actors to attempt to influence them into promoting or hiding particular items.

- **Privacy**. Obtaining detailed information about the user is essential for any recommendation system. However, different legal frameworks restrict the amount of such information that can be obtained.

- **Scalability**. Any algorithm with a linear ($O(N)$) or higher complexity will struggle to perform at all in large datasets. Complexity here means that as the size of a problem increases, the time or resources needed to solve it also increases in a direct and consistent manner. In datasets using potentially tens of millions of users or products, algorithms that don't account for those data sizes, will not be usable.

3 Content-based Filtering

Content-based filtering is a prominent technique employed in information retrieval and recommendation systems to personalize content recommendations for users. It relies on the inherent characteristics and properties of the items themselves, such as textual features or metadata, to establish similarity measures

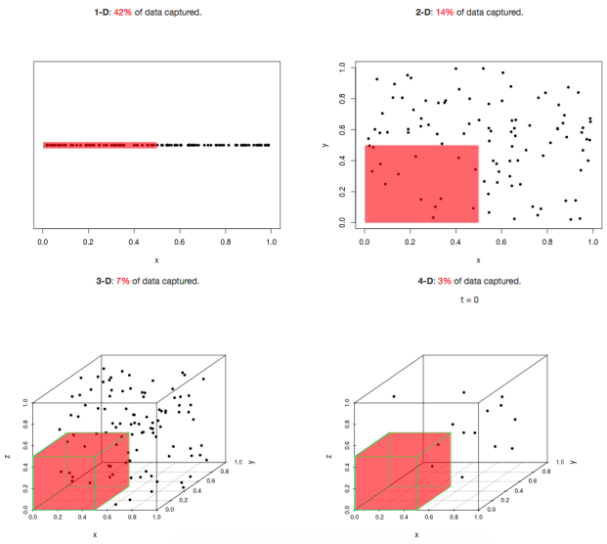

Figure 1: A visualization of a dataset suffering from the curse of dimensionality. Note how the more dimensions we view our data, the less data there are to observe. Image credits: (Eran Raviv).

between items. By analyzing the content of items that a user has already interacted with, content-based filtering algorithms identify other items with comparable attributes to recommend.

CB systems can be categorized in two main categories:

- **Memory-Based** systems, which query an extensive dataset (usually in a database), and attempt to find patterns by utilizing similarity metrics such as Pearson's coefficient or cosine similarity.
- **Model-based** systems, which "encode" the dataset in an ideal, internal representation which is used to make predictions by a model. The model in this case usually is a machine learning algorithm.

The main advantages of a CB system are [3, 10]:

- Their results can be more personalized, since they are independent of other users.
- They are easy to interpret, making explaining the results to a user easy.
- They can recommend products not picked by other users (e.g brand-new or unpopular products).

The disadvantages of a CB system are [3, 10]:

- Modelling and selecting the attributes of an item in order for the system to properly ”understand” it, is difficult.
- The system is prone to overfitting to its own predictions, since its predictions are used to drive future ones.
- Verifying the performance of a CB system is difficult. Accuracy metrics exist, but only in an offline environment [9, 7].

Additionally, Glauber and Loula [7] seem to indicate that CB system tend to select different items when in small recommendation lists than CF algorithms, although their results tend to converge the bigger these lists become. Their results also show that CB algorithms tend to be more robust, as the CF algorithms tended to struggle in datasets with many items but low user collaboration.

Furthermore we believe that the issue of feature selection and representation may be solved by modern neural network architectures, which have the ability to learn their own representations from raw data. For instance, while traditional CB models relied on ”manual” representations for text, advancements in Natural Language Processing (NLP) could drastically improve the quality of the data that can be automatically extracted from product titles and descriptions.

4 Collaborative Filtering

This approach relies on the collective ”wisdom” of a user community by analyzing their past behaviors and preferences,as opposed to the items themselves. Through the identification of similar users or items, collaborative filtering algorithms seek to generate recommendations for a target user based on the experiences and choices of other like-minded individuals. By leveraging user-item interaction data, such as ratings or purchase history, collaborative filtering algorithms establish patterns of user behavior and establish relationships between users or items. Similarly to CB systems, CF systems are divided to **memory-based** and **model-based** in an identical fashion.

The main advantages of a CF system are [3, 10]:

- Easier implementation (especially for memory-based systems).
- Memory-based systems are much easier to incrementally enrich with new data.
- Can discover more complex patterns than CB systems, since it relies on the behavioral patterns of other humans.

The disadvantages of a CF system are [3, 10]:

- Especially vulnerable to the *cold start problem.*
- The **grey sheep** problem, where some users do not fit in any of the categories discovered by our system.
- Due to the enormous amounts of users compared to products, CF algorithms are much less scalable.
- Ratings are only attributed sparsely and in a small subset of all products, leading to data sparsity issues.

Nallamala et al. [9] claim that hybrid recommendation systems can alleviate most of the issues of CF systems. Hybrid systems are systems using a combination of CB and CF algorithms (although other recommendation systems can be mixed as well), either by using both of their recommendations for predictions, using one's recommendations to train the other or by combining their inputs. Additionally, they have become an intense field of study for the application of neural networks [1], leveraging the expressive power of neural network architectures, allowing them to combine knowledge of traditionally CB and CF exclusive information.

5 Conclusions

Recommendation systems play a vital role in various domains, and understanding the different types of systems is crucial. Content-based filtering focuses on the attributes of items and recommends similar items based on user preferences, while collaborative filtering relies on user behavior to find like-minded individuals and generate recommendations. Both approaches have shared weaknesses, such as the cold start problem, data sparsity, and scalability issues. There is no ideal system, nor is either of the two systems more effective than the other, meaning that the prospective manager must thus make an informed decision, based on the nature of their business and data.

References

[1] Yassine Afoudi, Mohamed LAZAAR, and Mohammed Al Achhab. "Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network". In: *Simulation Modelling Practice and Theory* 113 (July 2021), p. 102375. DOI: 10.1016/j.simpat.2021.102375.

[2] Pegah Malekpour Alamdari et al. “A systematic study on the recommender systems in the E-commerce”. In: *Ieee Access* 8 (2020), pp. 115694–115716.

[3] Poonam B.Thorat, R. Goudar, and Sunita Barve. “Survey on Collaborative Filtering, Content-based Filtering and Hybrid Recommendation System”. In: *International Journal of Computer Applications* 110 (Jan. 2015), pp. 31–36. DOI: 10.5120/19308-0760.

[4] Xiaomei Bai et al. “Scientific paper recommendation: A survey”. In: *Ieee Access* 7 (2019), pp. 9324–9339.

[5] Jesus Bobadilla Sancho et al. “A collaborative filtering approach to mitigate the new user cold start problem.” In: *Knowledge-Based Systems* 26 (Feb. 2012), pp. 225–238. ISSN: 0950-7051. URL: http://www.journals.elsevier.com/knowledge-based-systems/.

[6] Qiming Diao et al. “Jointly Modeling Aspects, Ratings and Sentiments for Movie Recommendation (JMARS)”. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’14. New York, New York, USA: Association for Computing Machinery, 2014, pp. 193–202. ISBN: 9781450329569. DOI: 10.1145/2623330.2623758. URL: https://doi.org/10.1145/2623330.2623758.

[7] Rafael Glauber and Angelo Loula. “Collaborative filtering vs. content-based filtering: differences and similarities”. In: *arXiv preprint arXiv:1912.08932* (2019).

[8] Marius Kaminskas, Francesco Ricci, and Markus Schedl. “Location-Aware Music Recommendation Using Auto-Tagging and Hybrid Matching”. In: *Proceedings of the 7th ACM Conference on Recommender Systems*. RecSys ’13. Hong Kong, China: Association for Computing Machinery, 2013, pp. 17–24. ISBN: 9781450324090. DOI: 10.1145/2507157.2507180. URL: https://doi.org/10.1145/2507157.2507180.

[9] Sri Hari Nallamala et al. “A Brief Analysis of Collaborative and Content Based Filtering Algorithms used in Recommender Systems”. In: *IOP Conference Series: Materials Science and Engineering* 981 (Dec. 2020), p. 022008. DOI: 10.1088/1757-899X/981/2/022008.

[10] Sandeep K. Raghuwanshi and R. K. Pateriya. “Collaborative Filtering Techniques in Recommendation Systems”. In: *Data, Engineering and Applications: Volume 1*. Ed. by Rajesh Kumar Shukla et al. Singapore: Springer Singapore, 2019, pp. 11–21. ISBN: 978-981-13-6347-4. DOI: 10.1007/978-981-13-6347-42. URL: https://doi.org/10.1007/978-981-13-6347-42.

[11] Markus Schedl and David Hauger. “Tailoring Music Recommendations to Users by Considering Diversity, Mainstreaminess, and Novelty”. In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’15. Santiago, Chile: Association for Computing Machinery, 2015, pp. 947–950. ISBN: 9781450336215. DOI: 10.1145/2766462.2767763. URL: https://doi.org/10.1145/2766462.2767763.

[12] Sanjeevan Sivapalan et al. “Recommender systems in e-commerce”. In: *2014 World Automation Congress (WAC)*. IEEE. 2014, pp. 179–184.

[13] Jason Turcotte et al. “News Recommendations from Social Media Opinion Leaders: Effects on Media Trust and Information Seeking”. In: *Journal of Computer-Mediated Communication* 20.5 (June 2015), pp. 520–535. ISSN: 1083-6101. DOI: 10.1111/jcc4.12127. eprint: https://academic.oup.com/jcmc/article-pdf/20/5/520/19492447/jjcmcom0520.pdf. URL: https://doi.org/10.1111/jcc4.12127.