Using embeddings for solving cold-start problem

Vicente Ipinza
Pontificia Universidad Católica de Chile
Santiago, Chile

Marta Mariz
Pontificia Universidad Católica de Chile
Santiago, Chile

José Vergara
Pontificia Universidad Católica de Chile
Santiago, Chile

Victor Hernández
Pontificia Universidad Católica de Chile
Santiago, Chile

ABSTRACT

The cold-start problem is a challenge faced by recommender systems. In this research, we are tackling the recommendation of new items that users have not interacted with. We analyze two different datasets; the first one is a dataset that contains reviewed books on Amazon GoodReads and the second is a dataset related to Steam reviews games. For this analysis, we use different embeddings to provide different perspectives in addressing the problem. We found that Tf-Idf is the best embedding type to solve the studied problem.

KEYWORDS

Cold start, Embeddings, Tf-Idf

1 INTRODUCTION

1.1 Context

The cold-start problem in recommender systems refers to the difficulty of providing accurate recommendations when historical information about user preferences is lacking. This research addresses this challenge using a dataset of books evaluated on Amazon and a dataset of games using some data from the database of steam that divides the information between games and users.

1.2 Datasets

Dataset of +3M interactions in +200k books. A subset of +500k interactions on +100k books will be used, so quite a bit of the original set is preserved, and it is a large subset anyway. Amazon Books Reviews [1]. Each book contains multiple attributes but the ones that are used are book Title and book description because the goal was to try to represent an item with textual data. Each interaction has the userID, the bookID, the book Title, and the rating given. The rating is important to later define the user's preferences.

The Steam Database houses a vast collection of 32,135 distinct games, each characterized by 16 informative columns detailing various aspects such as publisher, genres, title, and more. This extensive dataset provides a comprehensive overview of the diverse gaming landscape within the Steam platform, offering valuable insights into the wide array of titles available to users. Within our dataset collection from the Steam Database, there's an additional set comprising 25,389 user IDs, 3,681 game IDs, and corresponding reviews that articulate the experiences and perspectives of users on specific games. This supplementary dataset enriches our understanding by providing valuable user-generated insights, creating a more comprehensive picture of the gaming ecosystem on the Steam platform. Together, these datasets offer a multifaceted view of both the games themselves and the user interactions and sentiments associated with them.

1.3 Contribution

Comparing different embeddings to do content-based recommendation and evaluate their effectiveness on the cold start problem for new items.

2 METHODOLOGY

2.1 Problem Definition

To address the cold start problem, we divide the set of items and interactions into subsets: one representing past interactions and another containing new items and interactions, used for testing. We made sure the users presented in testing were also in the training set so there was a way to represent them and that there were enough reviews to compare with the recommendations. As a baseline, we employ a randomized recommendation approach.

2.2 User Representation

For the Amazon Books reviews, we represented each user by textual descriptions of the books they gave high ratings to. We explored multiple representations, including TFIDF, Bag of Words and BERT embeddings.

2.3 Item Representation

For the Amazon Books reviews similarly to user representation each new item is represented with its description. The representations were the same as described above. In the context of the Steam dataset, we implemented an interaction that enables us to connect entries and establish relationships between the games dataset and user reviews. This linkage, facilitated by certain categories such as Title, price, and genres, allows us to seamlessly integrate information from both datasets. By doing so, we gain a more comprehensive understanding of user interactions, preferences, and sentiments associated with specific games, thereby enriching our analytical capabilities. This interaction was accomplished by utilizing the connection that indicates whether a user has written a review for a game or not. This approach provides valuable insights into user preferences, allowing us to discern what aspects appeal to users and what does not. So each item is

2.4 K-Nearest Neighbors For Recommendation

We employed content-based recommendation, using KNN to compute the similarity between user representation and new items.

This algorithm in the context of item based recommendation operates in the following steps:

(1) Each item (book, movie, product) is first represented as a vector in a high-dimensional embedding space. With this embedding we can capture valuable information like item's properties and latent relationships with other items.
(2) To identify the nearest neighbors, a cosine similarity metric is used to quantify the similarity between item vectors.
(3) A predetermined number k, of nearest neighbors are identified for each item. This k value can affect the balance between precision and recall leading to more diverse recommendations with higher k and more focused suggestions with lower values.
(4) When the k nearest neighbors are identified for an item, their associated items are considered potential recommendations for the user.

In the Amazon Books reviews dataset, the user is represented as the items they gave high ratings to and then KNN is used to find similar representations among the new items.

2.5 Evaluation

Recommendations were compared with user-rated items, sorted by rating. To evaluate the quality of the recommendations, we used Precision@K and NDCG@K. To evaluate the diversity of recommendations, we used the Coverage metric.

(1) Precision at k: proportion of relevant items in the top k recommended items
(2) NDCG at k: metric that evaluates the ranking quality of the recommendations. Considers relevance of the recommendation and it's position.
(3) Item Coverage: Proportion of all existing items that are recommended

For the baselines we made random recommendations from our testing subset and then compare it with the actual user interactions and evaluate with the same metrics mentioned before

3 EXPERIMENTS

Within the experiments we obtained the following results:

nDCG Table Dataset	Models k	Baseline	Tf-idf	Bag-of-words	sBERT
Amazon Books Reviews	@3	.0007	.1388	.0094	.0098
	@5	.0017	.1389	.0088	.0103
	@10	.0033	**.1413**	.0093	.0119

Table 1: nDCG in Amazon Books Reviews

Precision Table Dataset	Models k	Baseline	Tf-idf	Bag-of-words	sBERT
Amazon Books Reviews	@3	.0019	.136	.009	.010
	@5	.0018	.137	.008	.011
	@10	.0019	**.141**	.009	.013

Table 2: Precision in Amazon Books Reviews

Coverage Table Dataset	Models k	Baseline	Tf-idf	Bag-of-words	sBERT
Amazon Books Reviews	@3	.7240	-	-	-
	@5	.8813	-	-	-
	@10	.9859	**.7883**	.0985	.2903

Table 3: Coverage in Amazon Books Reviews

Precision Table Dataset	Models -	Baseline	Basic KNN
Steam Games Reviews	All	.00597	.00007

Table 4: Precision in Steam Games Reviews

Recall Table Dataset	Models -	Baseline	Basic KNN
Steam Games Reviews	All	.00252	.00007

Table 5: Recall in Steam Games Reviews

4 CONCLUSIONS

While the content-based recommendation approach outperformed the random baseline in the experiments with the Amazon Book reviews dataset, the overall quality of recommendations remained deficient due to the inherent challenges of the cold-start problem. Surprisingly, Bert's more complex feature representation yielded somewhat underwhelming results. Interestingly, TF-IDF emerged as the most effective method, demonstrating superior performance in both recommendation quality and variety within the chosen dataset. These findings suggest further exploration of TF-IDF's efficacy in cold-start settings.

Regarding the Steam dataset, addressing the content-based approach presented significant challenges due to the inherent characteristics of our datasets. The lack of detailed game descriptions limited our ability to employ embeddings, a common technique in this type of analysis. Opting to use the KNN algorithm to cluster the games, which totaled 32,000, and considering that only 4,000 games provided meaningful information for model learning, the results proved unsatisfactory. The shortage of relevant data and the simplicity of the model contributed to poor performance. This issue was exacerbated when adding new games, as the model lacked the capability to make effective recommendations overall. This was partly due to the diversity of player preferences, as they are not restricted to seeking exclusively one genre of game. Consequently, the complexity of user preferences posed an additional challenge for the model, emphasizing the need for more sophisticated approaches in future developments.

Overall given the nature of the problem the possible results depend a lot on the characteristics of the items and how users interact with different kinds of content. Content-based recommendation seems to be the way to approach the problem and the ways to better represent items remain a challenge.

5 FUTURE WORK

Include a more thorough analysis of the metrics to be used and evaluate the metrics by type of the items to obtain more information

about the domain. Include new embeddings for the study, such as GPT-2 and explore specialized embedding for the used domains Conduct experiments with datasets that include items with more complete and descriptive characteristics.

6 REFERENCES

[1] Dataset Amazon Books Reviews. Kaggle Dataset

[2] Steam Video Game and Bundle Data. Steam Dataset

[3] Introduction To Recommender Systems- 1: Content-Based Filtering And Collaborative Filtering. Content-Based Filtering And Collaborative Filtering

ACKNOWLEDGMENTS

This work was carried out by Vicente Ipinza, Marta Mariz, José Vergara and Víctor Hernández, under the course "Recommender Systems" of the Pontificia Universidad Católica de Chile.