Hotel Recommendations Based on Pre-Trained Language Models: Evaluation of Effectiveness and Comparison with Traditional Methods

Sistemas Recomendadores 2023-2

Jairo Navarro
PUC Chile
jznavarro@uc.cl

Begoña Pendas
PUC Chile
mbpendas@uc.cl

Fabián Riveros
PUC Chile
fcriveros@uc.cl

Abstract

This study evaluates the effectiveness of hotel recommendations using pre-trained language models compared to traditional models. An approach using language models is proposed and compared to conventional methods such as matrix factorization. The methodology encompasses data collection and processing, model application, and evaluation using metrics such as MAP and NDCG. Initial results show a promising outlook, although additional iteration is required to improve the accuracy of the proposed model.

Keywords

Next-Item Recommendation, Large Language Models, Prompting.

1 Introduction

This paper focuses on the relevance of recommendations in the hotel sector, applying methodologies based on recommender systems. This recommendation approach allows enriching the user experience by facilitating the search for hotels according to their profiles, based on their preferences and requirements. By offering personalized recommendations, customers are empowered to make decisions that enhance their satisfaction with their choice.

Furthermore, the adoption of techniques based on pre-trained language models could improve the ability of hotels to generate recommendations more efficiently and more conveniently for users. To this end, the study seeks to address the central issue of validating the usefulness of recommendations based on pre-trained language models for users. This would allow the generation of recommendations based on similarity to their historical interactions.

The study proposes to demonstrate the effectiveness of a prompting-based strategy for hotel recommendation by means of language models. The objective of this research project is to test the feasibility of hotel recommendations made by pre-trained language models. For this purpose, we propose to implement a Next Item Recommendation to a language model (GPT-3). Then, to make recommendations to representative users, by means of the pre-trained language model. Finally, a comparison of relevant metrics, such as MAP@k and NDCG@k, is performed.

2 State of art

This work is based on the work done by Wang & Ee-Peng (2023), in which they evaluate the implementation of a Next Item Recommendation (NIR) to a language model. This study explores the possibility of using Large Language Models (LLM) as recommenders, due to their good performance in Natural Language Processing (NLP) tasks. For this, the authors propose the use of a 3-step strategy. First, a prompting that captures user preferences. Then, a prompting that delivers a selection of representative movies and finally a prompting for the language model to deliver the recommendations.

In the context of recommendations based on Language Models, it is a novel area because it has not been fully developed. One of the most recent researches is the one presented by Cui et al. (2022), in which they develop the problem of the large amount of data in the hotel area and the linguistic difficulty of online recommendations, proposing probabilistic linguistic term sets to face these problems. They conclude that language models have potential in the field and the difficulty of the data can be dealt with probabilistically.

3 Dataset

For the construction of the dataset, hotel reviews provided by Datafinity (2023), composed of data from 1000 hotels with user reviews, were used. Among the characteristics provided by this list

are name, location, rating, among others. The attributes of the data are presented in the following table.

	Type
id	string
categories	string
city	string
country	string
name	string
province	string
reviews date	datatimestamp
reviews rating	float
reviews text	string
reviews title	string
reviews user city	string
reviews province	string
reviews username	string

Table 1: Attributes categorization

4 Methodology

4.1 Tools

For the study, the programming language used was Python with the Jupyter and Surprice libraries for the implementation of recommendation models by matrix factorization and for prefiltering by means of KNN. With this code, the implementation of the aforementioned metrics, MAP@K and NDCG@K, was also carried out.

4.2 Algorithms

This study was based on the realization of a prefiltering of the data by means of a data clustering considering the K Nearest Neighbours (KNN) method. This, so that the language model can focus on more relevant elements to generate more accurate recommendations.
On the other hand, the iteration strategy for the queries made to GPT-3 consisted of giving it a pre-filtered dataset. Thus, it was given as a prompt with the following question: "What features are most important to me when selecting hotels? (Summrize my preferences briefly)". Then, based on the answer received, the following question was asked: "Select the hotels (at most 5 hotels) that appeal to me the most from de list of hotels I have visited, based on my personal preferences. The selected hotels will be presented in descending order of preference (Format: no. A visited hotel)". Once the answer is obtained, where the top 5 hotels visited by the user in the previous step are presented. Once the response is obtained, which contains the top 5 hotels visited by the user according to the inference regarding the user's preferences in the previous step, the model is asked to deliver a recommendation of hotels from the set of candidates defined in the first query to the model.

4.3 Metrics

Among the metrics used for the evaluation of the study were MAP@k and NDCG@K. With these, traditional recommendation models were evaluated, such as matrix factorization by means of Alternate Least Squares (ALS) and Bayesian Personalized Ranking (BPR).
The Mean Avarege Precision (MAP) metric is a metric that evaluates the average precision of the recommender system considering the relevance of the recommended items. For each query, the precision is calculated at the position layer where a relevant result is presented. Then, the average of these precisions is taken for all queries or users.

$$AP = \frac{\sum_{k=1}^{N} P@\imath * rel(k)}{|Relevants|}$$

$$MAP = \frac{1}{N} \sum_{u=1}^{N} AP(u)$$

On the other hand, the Normalized Discounted Cumulative Gain (NDCG) metric measures the usefulness or relevance of an ordered list of items based on the position and relevance of each item. It is obtained by normalizing the DCG (Discounted Cumulative Gain) by dividing it by the ideal DCG, where the ideal DCG is obtained by ordering all the elements by their relevance.

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log_2(1 + \imath)}$$

$$nDCG_p = \frac{DCG_p}{\imath DCG_p}$$

In the context of leveraging large language models for recommendation systems, the employed metrics exhibit some variance compared to conventional ones. The compilation of the candidate set was facilitated through collaborative filtering techniques, specifically employing user-based and item-based methodologies. Following the acquisition of predictions for n users, a subset comprising 10 representative predictions was chosen. This subset was characterized by diverse

rankings to facilitate a comprehensive comparative analysis of the recommendations. Consequently, an acceptable recommendation was defined as one that, within the pool of candidates, identified hotels with higher predicted ratings.

Results

Tables 2, 3, 4 and 5 show the results obtained from the metrics evaluated for the classical recommendation methods, such as matrix factorization. This table shows ALS and BPR evaluated for MAP and NDCG @10 and 20. It is worth noting that, for both recommendation models using matrix factorization, the results are similar with low performance.

	MAP@10	**MAP@20**
50	0.012	0.019
100	0.005	0.019
200	0.003	0.011
500	0.003	0.015
1000	0.005	0.012

Table 2: MAP metric for ALS.

	NDCG@10	**NDCG@20**
50	0.012	0.019
100	0.005	0.019
200	0.003	0.011
500	0.003	0.015
1000	0.005	0.012

Table 3: NDCG metric for ALS.

	MAP@10	**MAP@20**
50	0.008	0.009
100	0.007	0.011
200	0.007	0.018
500	0.015	0.022
1000	0.009	0.024

Table 4: MAP metric for BPR.

As for the proposed Language Model, in order to assess the efficacy of the proposed methodology, a sample of 15 users was randomly chosen from the database, and prompts were formulated for each user in accordance with the previously delineated procedure. Subsequently, each user was solicited for five recommendations, utilizing the metric explicated in the antecedent section. Specifically, if the recommendation proffered by the language model attained a high ranking in comparison to the previous prediction, it was assigned a value of 1; otherwise, it received a value of 0. This process resulted in the calculation of an average rating for each of the 15 users based on the five recommendations. The resultant mean value was determined to be 0.71. This figure signifies that approximately 70% of the time, the imparted recommendations are deemed satisfactory, underscoring a highly promising outcome.

	NDCG@10	**NDCG@20**
50	0.008	0.009
100	0.007	0.011
200	0.007	0.018
500	0.015	0.022
1000	0.009	0.024

Table 5: NDCG metric for BPR.

Discussion

In general, it can be observed that the results delivered by the recommendation made with matrix factorization deliver results that can be considered low for this type of recommendations. Likewise, it is not possible to be certain about a comparison with the proposed prompting model in Language Models, since, due to the lack of tools, it was not possible to generate a contrast between both approaches.

For the study of recommendations based on Language Model promptings, browsing was used directly for the generation of recommendations, not an API because of the complications in accessing it. This influenced the way the model was calibrated, because there was no access to the code. Also, the generalization of recommendations was complicated because users were arbitrarily chosen for recommendations and the number of users to whom recommendations were given was limited to the programmer's ability to perform GPT-3 queries.

Given the above, there may be several problems with the recommendations obtained. First, and as mentioned above, the loss of generality of the results. This is due to the fact that the sample size used in the GPT prompting is not representative of the population. Secondly, the experiment could be categorized as empirical, because the samples taken may not be representative of the population, causing a bias within the results obtained. Finally, the lack of a common line between the recommendation methods due to the lack of tools present may make it difficult to define a comparative baseline between the different recommendation models.

Conclusions

Language Models have shown a good performance in conventional tasks, so their use as recommendation methods is an area that is being explored, with different approaches that show their potential within these. This study sought to see their feasibility as recommenders within the hotel industry. However, due to the lack of tools for the construction of a robust model, their usefulness cannot be generalized for all recommendations made.

However, it was also contrasted with the recommendations made by means of matrix factorization models. These recommendations showed relatively low comparison metrics for the expected results. Nevertheless, from the empirical perspective of the recommendations made by the GPT-3 Language Model, it was possible to obtain relevant recommendations for the features described by the user. This shows that LLMs do have the potential to make personalized recommendations to users, in this case, hotels. As future work, the use of LLM as Llama, instead of GPT-3, could be explored.

The proposed future work includes the implementation of the prompt strategy to the Language Model from its API, which would allow a better management of the parameters provided and the ability to generalize to users. Also, the implementation of metrics that allow a comparison between the classical models with this proposal.

References

Cui, Chunsheng, Meng Wei, Libin Che, Shouwen Wu & Erwei Wang. 2022. Hotel recommendation algorithms based on online reviews and probabilistic linguistic term sets .

Datafinity. 2023. Hotel reviews dataset .

Wang, Lei & Lim Ee-Peng. 2023. Zero-shot next-item recommendation using large pretrained language models .