December 2022

Generating Personalized Recommendations via Large Language Models (LLMs)

Hakim Sidahmed

Sameer Ahuja

Manoj Kumar Tiwari

Generating Personalized Recommendations via Large Language Models (LLMs)

ABSTRACT

Personalized recommendations used in many applications and websites are generated using techniques such as collaborative filtering, content-based filtering, reinforcement learning, etc. These are task-specific approaches. Large language models (LLMs) can generate predictions based on priming with specific input without the need for task-specific model tuning. However, LLMs have not been applied for making personalized recommendations because their maximum input size is smaller than the typical size of user histories used to personalize recommendations. This disclosure describes techniques to obtain personalized recommendations via LLMs by automatically augmenting a user command or query with relevant text phrases about the user. The set of relevant phrases that fit within the input limits of the LLM are extracted from a collection of phrases obtained from relevant historical and contextual information sources based on the embeddings generated based on the user command or query. Implementation of the techniques can improve the relevance and utility of personalized recommendations and can lead to increased user engagement with the recommended content.

KEYWORDS

- Recommender system
- Collaborative filtering
- Content-based filtering
- Reinforcement learning
- Large language model (LLM)
- Task-specific model tuning
- Priming input

BACKGROUND

In many online contexts, users are offered personalized recommendations for a variety of items, such as movies, songs, queries, consumer products, services, etc. Sometimes, the recommendations can also be in the form of personalized advertisements, when the user has provided permission for such personalization. To generate personalized recommendations, user permission is obtained to access user attributes and/or data related to their explicit or implicit ratings of a set of items. The personalized recommendations are usually generated using one or more of the following underlying approaches:

1. **Collaborative Filtering**: In this technique, users are grouped by similarity based on their explicit or implicit ratings for a large number of items. The user grouping is used to make predictions for the likelihood of a particular user being interested in individual items that they have not yet rated. The techniques can be implemented using algorithms such as matrix factorization, neighborhood-based selection, etc.
2. **Content-based Filtering**: Such approaches are used to obtain item recommendations based on a model of the user’s preferences about various item features (e.g., name, brand, size, price, description, weight, genre, etc.). The applicable set of features are dependent on the characteristics of the item being recommended.
3. **Reinforcement Learning**: In this approach, recommendations are derived by maximizing a specific reward function applicable to the items to be recommended based on user interaction with previously shown items. The user is treated as the environment, and the recommender system is the agent. For instance, the recommended personalized advertisements can be selected to maximize click-through rate (CTR).

In a somewhat similar vein, large language models (LLMs) are applied to process a given input sequence of words to predict the next set of words that are likely to follow. LLMs are large neural networks that typically employ transformer-based architectures with billions of parameters. LLMs are trained for language understanding and generation tasks. Although LLMs have been used for few-shot learning to achieve impressive results without the need for task-specific model tuning, LLMs have not been applied for making recommendations. Using LLMs to generate suitable personalized recommendations via a primed input sequence is challenging because LLMs typically operate on a limited number of input tokens (e.g., a few hundred). The maximum input size is smaller than the typical size of user histories that are used to personalize recommendations. Moreover, fine tuning LLMs is costly because of the large number of parameters.

DESCRIPTION

This disclosure describes techniques that use LLMs for making personalized recommendations. With user permission, the recommendations can be generated by prepending a user's command or query with relevant text phrases about the user. For instance, phrases relevant for personalizing recommendations for songs can be "I like classic rock," "I frequently attend live concerts," etc. In contrast, phrases such as "I enjoy playing beach volleyball" can be excluded because these are likely to be irrelevant for the purposes of recommending songs.

A collection of phrases describing a user's characteristics, context, history, etc. can be obtained with user permission from relevant sources such as the user's conversations, preference settings, interaction history, etc. Alternatively, or in addition, users can explicitly provide such descriptive phrases themselves. The user is provided options to enable/disable use of such phrases, the context in which they are used for generation of recommendations, and to exclude certain phrases from use.

The number of phrases in such a collection can grow to be much larger than the limited priming input size expected by typical LLMs. Therefore, it is necessary to select phrases that are the most relevant for the type of personalized recommendation the user wishes to obtain. The set of relevant phrases can be extracted by using the appropriate embeddings based on the user's command or query. If the size of the relevant phrases is larger than the number of permissible input tokens for the LLM, the top N phrases that fit within the input limitations can be selected by ranking them based on relevance and/or other appropriate criteria. The selected set of relevant phrases are then automatically added to the user's command or query in any suitable manner, such as adding them before the user input.

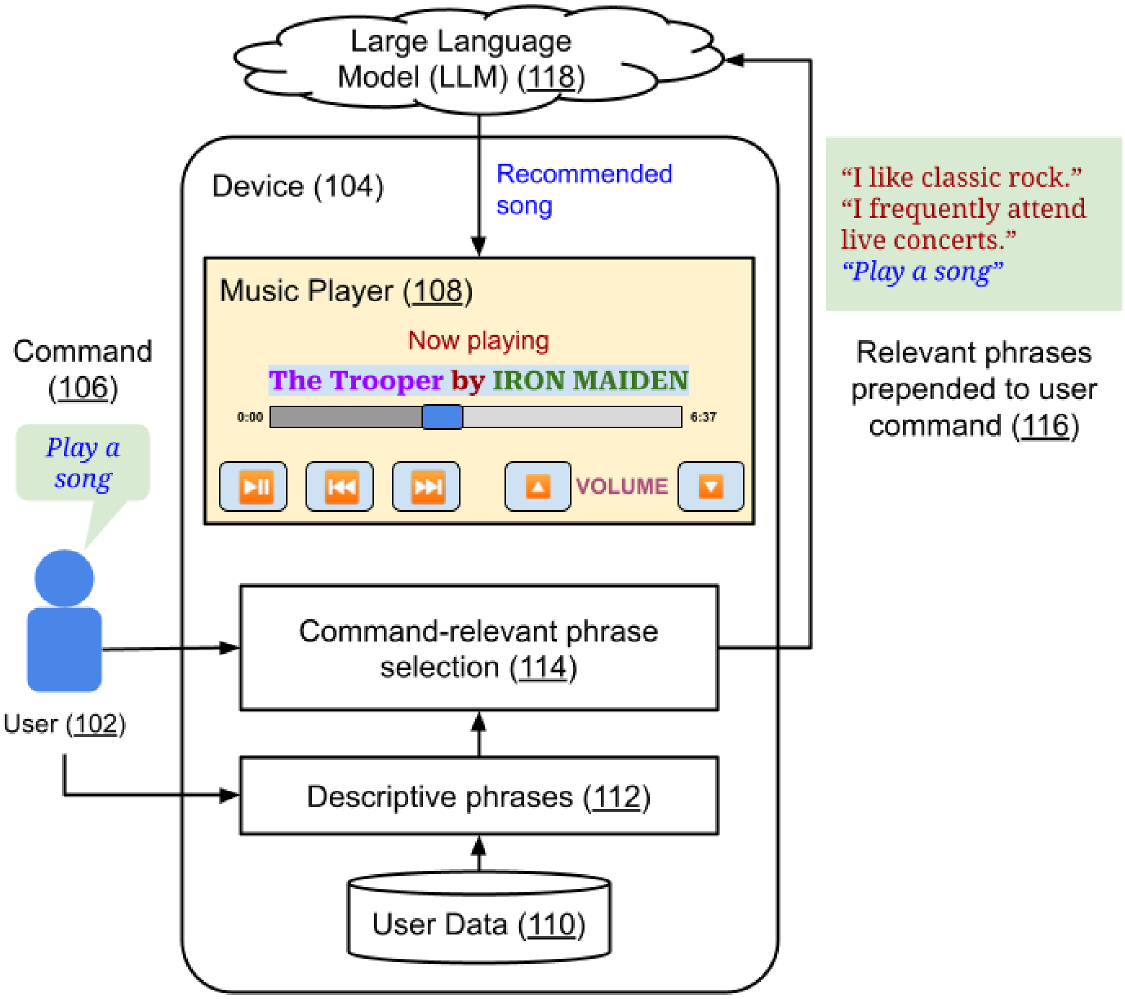
Large Language Model (LLM) (118)
Device (104)
Recommended song
Music Player (108)
Now playing
The Trooper by IRON MAIDEN
0:00
6:37
VOLUME
Command (106)
Play a song
User (102)
Command-relevant phrase selection (114)
Descriptive phrases (112)
User Data (110)
"I like classic rock."
"I frequently attend live concerts."
"Play a song"
Relevant phrases prepended to user command (116)

Fig. 1: Obtaining personalized recommendations via LLMs by appending relevant descriptive phrases to the user command

Fig. 1 shows an example operational implementation of the techniques described in this disclosure. A user (102) issues a command (106) "play a song" to a music playback application

(108) on a device (102). With user permission, descriptive phrases about the user (112) obtained based on the user’s interaction history (110) and/or input by the user are analyzed (114) to identify phrases that are relevant to the command.

In the example illustrated in Fig. 1, two such phrases are identified - “I like classic rock” and “I frequently attend live concerts.” The command is prepended with the relevant phrases (116) and the combination is input to an LLM (118). The output of the LLM is the personalized recommendation for the song to be played.

The techniques described in this disclosure can be implemented on any device, application, or platform to support the use of any LLMs to obtain personalized recommendation for any content, product, service, etc. Alternatively, or in addition, the techniques can be made available via standard mechanisms such as an application programming interface (API). The number of the relevant phrases to stay within permissible input limits for the LLM can be set by the developers and/or determined dynamically at runtime. Implementation of the techniques enables the use of LLMs to obtain personalized recommendations, without the need for explicitly specifying a large amount of background or contextual information about the user. Such an interactive experience can be provided via any suitable user interface (UI), such as chat, dialog boxes, notifications, text boxes for entering queries, virtual assistants, etc. Alternatively, or in addition, the functionality can also be provided at the backend for various applications to access programmatically, without requiring users to interact directly with the LLM.

The techniques described in this disclosure makes the powerful text generation capabilities of LLMs available for generating personalized recommendations without the need for task-specific model tuning. The techniques add to the repertoire of techniques available for generating personalized recommendations that are relevant for a user’s needs. Implementation of the techniques can improve the relevance and utility of personalized recommendations and make it possible to obtain such recommendations without spending resources on task-specific model tuning. These improvements can serve to enhance the user experience of the underlying application or service within which the personalized recommendations are provided. Moreover, the enhancements in personalized recommendations that result from the implementation of the techniques can lead to increased user engagement with the recommended content which can help increase revenue for application and service providers.

Further to the descriptions above, a user may be provided with controls allowing the user to make an election as to both if and when systems, programs or features described herein may enable collection of user information (e.g., information about a user’s interaction history, a user’s context, social network, social actions or activities, profession, a user’s preferences, or a user’s current location), and if the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user’s identity may be treated so that no personally identifiable information can be determined for the user, or a user’s geographic location may be generalized where location information is obtained (such as to a city, ZIP code, or state level), so that a particular location of a user cannot be determined. Thus, the user may have control over what information is collected about the user, how that information is used, and what information is provided to the user.

CONCLUSION

Personalized recommendations used in many applications and websites are generated using techniques such as collaborative filtering, content-based filtering, reinforcement learning, etc. These are task-specific approaches. Large language models (LLMs) can generate predictions based on priming with specific input without the need for task-specific model tuning. However, LLMs have not been applied for making personalized recommendations because their maximum input size is smaller than the typical size of user histories used to personalize recommendations. This disclosure describes techniques to obtain personalized recommendations via LLMs by automatically augmenting a user command or query with relevant text phrases about the user. The set of relevant phrases that fit within the input limits of the LLM are extracted from a collection of phrases obtained from relevant historical and contextual information sources based on the embeddings generated based on the user command or query. Implementation of the techniques can improve the relevance and utility of personalized recommendations and can lead to increased user engagement with the recommended content.

REFERENCES

1. Borgeaud, Sebastian, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche et al. "Improving language models by retrieving from trillions of tokens." In International conference on machine learning, pp. 2206-2240. PMLR, 2022.
2. Weston, Jason, Emily Dinan, and Alexander H. Miller. "Retrieve and refine: Improved sequence generation models for dialogue." arXiv preprint arXiv:1808.04776 (2018).