June 11, 2019

Recommending content based on anticipated knowledge gain

Victor Cărbune

Alexandru Damian

Recommending content based on anticipated knowledge gain

ABSTRACT

This disclosure describes machine learning techniques to automatically remove, reorder, or prioritize segments of content such that users can quickly access information that is new to them. With user permission, data such as user profile, historical content consumption patterns and its relevance, etc. is obtained. This data is used to chunk and rank historical content segments based on relevance to the user. A machine learning model is trained to identify new content (or segments thereof) based on the knowledge gain that the content is likely to provide to the user.

KEYWORDS

- Content retrieval
- Content recommendation
- Recommender system
- Knowledge gain
- Information gain
- User embeddings

BACKGROUND

Users often spend time viewing content only to realize part way through that they already have substantial knowledge of the content being presented. For example, consider a user that already has prior exposure to a topic and then watches a technical presentation to learn more about the topic. It is possible that the user views nearly the entire content only to realize near the end that only a small portion of the content is new to them. Such a situation can arise with any type of content, e.g., text, online tutorials, podcasts and other audio, video, etc. Currently, topic-based partitioning of content is done by content creators with techniques such as table of contents, etc. However, the division, categorization, and recommendation of content to a user are personal choices that are difficult for a content creator to make on a per user basis.

DESCRIPTION

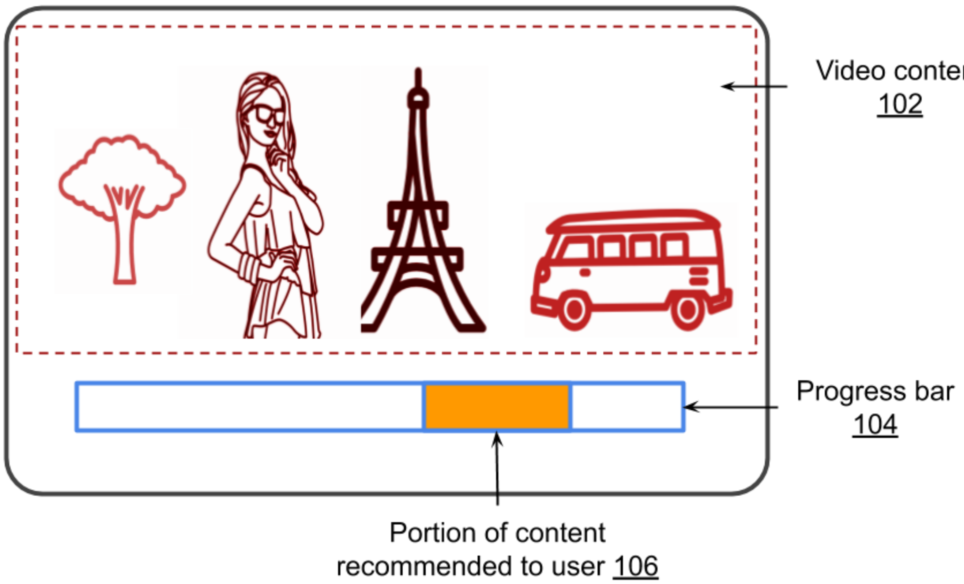

Fig. 1: Content recommendation based on anticipated knowledge gain

This disclosure describes machine learning techniques to automatically identify, remove, reorder, or prioritize segments of content such that users can quickly access information that is new to them. The techniques are implemented with user permission to access available user data to make the determination of content that is new to a user.

Fig. 1 illustrates an example of prioritizing content for a given user. As illustrated in Fig. 1, the user is viewing a video (102) that relates to vacationing in a foreign country. From available user data, e.g., prior content viewed by the user, portions of the video that offer new information are highlighted, e.g., a segment of the progress bar (104) of the video is indicated as most likely of interest (106).

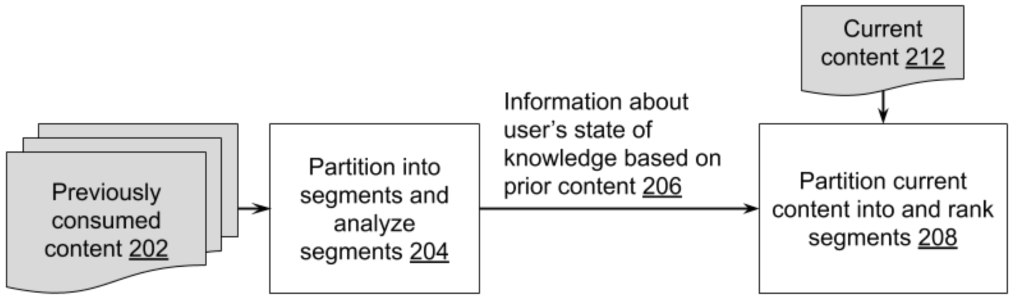

Fig. 2: Delineation of historically consumed content into segments of varying interest

With user permission, permitted user data such as user profile information (e.g., demographic information, location, etc.), previously viewed content, etc. are obtained and analyzed to identify segments of interest from a current piece of content. Fig. 2 illustrates the partitioning (204) of previously consumed content (202) into segments. The segments are analyzed to determine topics that the user has viewed and are therefore indicative of the user's state of knowledge (206). For example, for audio and video content, segments of content consumed at a normal pace indicate interest and skipped or fast-forwarded segments indicate low interest. For text content, user interest can be measured by the interactivity of the user with the content. For example, copying-and-pasting text, reading a segment for a long time, etc. are indications of high interest.

A current piece of content (212) data is partitioned into segments (208). The segments are ranked based on user's state of knowledge determined based on prior content. Segments that are determined as likely of interest from the current piece of content are marked by relevance to the user, e.g., whether there is substantial information gain to the user in light of previously consumed content.

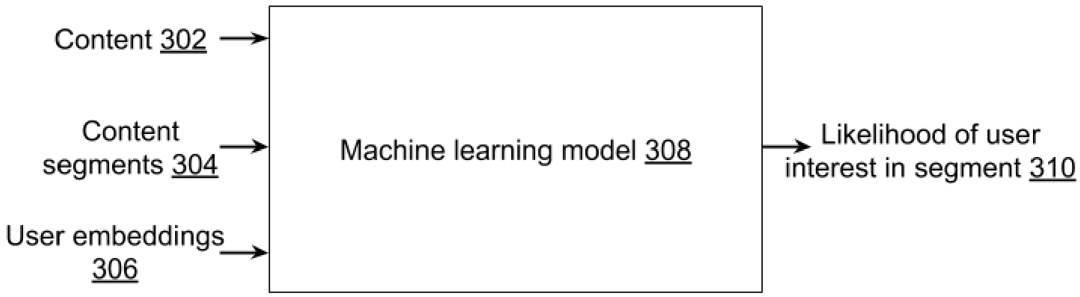

Fig. 3: Machine learning model to predict likelihood of user interest in a content segment

Fig. 3 illustrates the use of a machine learning model (308) to determine the likelihood of user interest in various content segments of a particular piece of content. The machine learning model utilizes user embeddings (306) that are representative of the user, content (302), content segment (304), etc. to assign a score for the likelihood of the user being interested in the content segment (310), e.g., whether there is substantial information gain by viewing this content based on content previously consumed by the user. The content can be pre-segmented, e.g., by using heuristic or other models that learn meaning and structure.

The machine learning model effectively trains on the previously consumed content and the segmented parts of the content to predict a score. A higher score indicates that the chunk represents content with useful knowledge to gain, which can be recommended to the user. The model works directly on content such as video, images, audio, text, etc. The machine learning model can be, e.g., convolutional towers, recurrent networks, transformer blocks, etc. Previously collected datasets can be used in supervised learning fashion. Partial-information friendly algorithms such as contextual bandits or reinforcement learning can also be used.

The described techniques can be implemented on platforms such as search engines, virtual assistants, video consumption platforms, etc. The described techniques are implemented with specific user permission to access user data to determine content likely of interest to the user.

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 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

This disclosure describes machine learning techniques to automatically remove, reorder, or prioritize segments of content such that users can quickly access information that is new to them.