Search and Recommendation Systems with Metadata Extensions

Woo-Hyeon Kim*, Joo-Chang Kim**

* Division of AI Computer Science and Computer Engineering, Kyonggi University, South Korea

** Contents Convergence Software Research Institute, Kyonggi University, South Korea

whkim712@kyonggi.ac.kr, kjc2232@naver.com

Abstract— **This paper proposes an AI-based video metadata extension model to overcome the limitations of video search and recommendation systems in the multimedia industry. Current video searches and recommendations utilize pre-added metadata. Metadata includes filenames, keywords, tags, genres, etc. This makes it impossible to make direct predictions about the content of a video without pre-added metadata. These platforms also analyze your previous search history, viewing history, etc. to understand your interests in order to serve you personalized videos. This may not reflect the actual content and may raise privacy concerns. In addition, recommendation systems suffer from a cold start problem, which is the lack of an initial target, as well as a bubble effect. Therefore, this study proposes a search and recommendation system by expanding metadata in videos using techniques such as shot boundary detection, speech recognition, and text mining. The proposed method selects the main objects required by the recommendation system based on the object frequency and extracts the corresponding objects from the video frame by frame. In addition, we extract the speech from the video separately, convert the speech to text to extract the script and apply text mining techniques to the extracted script to quantify it. Then, we synchronize the object frequency and the transcript to create a single contextual data. After that, we group videos and clips based on the contextual data and index them. Finally, we utilize Shot Boundary Detection to segment videos based on their content. To ensure that the generated contextual data is appropriate for the video, the proposed model compares the extracted script with the video's subtitle data to check and calibrate its accuracy. The model can then be fine-tuned by tuning and cross-validating the hyperparameter to improve its performance. These models can be incorporated into a variety of content discovery and recommendation platforms. By using expanded metadata to provide results close to a search query and recommend videos with similar content based on the video, it solves problems with traditional search, recommendation, and censorship schemes, allowing users to explore more similar videos and clips.**

Keywords— **Multimedia, Recommendation, Speech Recognition, Contextualized Data, Metadata**

I. INTRODUCTION

The multimedia industry is currently experiencing the proliferation of personal internet broadcasting and OTT platforms, creating a huge amount of new videos. Currently, general video search and recommendation uses metadata such as filenames, keywords, tags, and genres pre-added by the creator or distributor as an index [1]. Therefore, access to the video content will take a long time if there is no separate internal index. OTT platforms and video platforms that have a large amount of content may analyze a user's previous browsing history, viewing history, subscribed channels, etc. to understand their interests in order to provide them with customized videos. They compare this with metadata to recommend relevant content or analyze what other users with similar interests have watched, what's popular, etc. Based on this analysis, it selects recommended videos and serves them to you in order of priority. While most search and recommendation algorithms work well, they rely heavily on your sensitive information, such as your viewing history, search history, and internet browser cookies. This can lead to inaccurate recommendations if a user deletes their information or, depending on their level of privacy, lacks the necessary information to make a recommendation, a cold start problem that occurs in most recommendation systems. In addition, since metadata is added by the content creator or distributor, there is a possibility that information about the actual content is excluded, reducing the accuracy of search or recommendation, or being abused. Currently, similar content recommendation systems on OTT platforms often exclude key content that could be spoilers, even if the plot is included in the recommendation process. This can cause users to lose trust in the recommendation system if the actual content is different, even if it is recommended as similar simply because of a similar genre or visual atmosphere. Additionally, recommendation algorithms based on personal history can suffer from the bubble effect. The bubble effect is when only content relevant to your history appears in the recommendation algorithm, exposing users to biased content [2]. Recently, crimes have been committed to exploit this to expose children to harmful videos, raising concerns about recommendation algorithms and video censorship schemes [3]. To solve these problems, it is necessary to strengthen the censorship work and introduce stricter regulations to block and remove harmful content, which requires a lot of time and manpower.

In this study, we use AI-based content analysis to extract contextual data from real-world content. By indexing the extracted contextual data into videos and clips, we propose a model to extend metadata and improve search and recommendation systems.

II. Related Research

OTT is an acronym for Over The Top, which refers to media content such as video content, audio, and text that is typically delivered over the Internet. OTT services work by delivering content directly to consumers, bypassing traditional media distribution channels such as traditional broadcast or cable television. The main characteristics of OTT services include internet-based delivery, diverse content, subscription models, personalized recommendations, content diversity, and creator opportunities. Internet-based delivery means that OTT services deliver content to users via an internet connection, so users can watch or listen to content anytime, anywhere on a variety of devices, including smartphones, tablets, and smart TVs. Next, by diverse content, we mean that OTT services offer many different types of content, including movies, dramas, TV shows, sports broadcasts, audio podcasts, web series, and more. This allows users to choose and watch content that suits their different interests and needs. Subscription is a big feature of OTT services. Many OTT services adopt a subscription model, where users pay a set amount of money each month or year to access the service. Subscribers can use these services to watch the content they want without any ads. Next is personalized recommendations. Many OTT platforms offer personalized content recommendations by analyzing users' viewing habits. This makes it easier for users to find content that matches their tastes and interests [4]. Finally, there is content diversity and creator opportunities. Unlike traditional broadcast networks, OTT services have relatively low barriers to entry. This gives many creators the opportunity to produce and distribute their own content. Major OTT services include Netflix, Amazon Prime Video, Disney+, Hulu, Apple TV+, and Youtube Premium, which have millions of users worldwide. OTT is disrupting the media industry, and research is still ongoing.

Shot Boundary Detection is a technique for detecting scene transitions in video content. A video consists of a sequence of consecutive frames, and it detects changes in color, lighting, objects, etc. in a scene. Based on this, Shout Boundary Detection is the process of automatically segmenting video into basic units called shots. A shot is a group of consecutive video frames in a video. A shot usually consists of consecutive video frames from a single camera. There are several types of Shot Boundaries: Cut, Dissolve, Wipe, and Fade Out/In. Cut is the most common type of Shot Boundary and represents a transition from one frame to the next. Dissolve is a transition type where one scene gradually disappears and the next appears. Wipe is a transition from one scene to another, such as a horizontal or vertical wipe pattern across the screen. Fade Out/In is a type of transition where one scene gradually fades out or brightens and then disappears. Over the years, the design of the Shot Boundary Detection algorithm has evolved from simple feature comparisons to the use of rigorous probabilistic and complex models. In addition, to detect transitions with higher accuracy, orthogonal polynomials are applied to derive features in the orthogonal transform domain to detect hard transitions in video sequences [5].

Speech Recognition is a technology that recognizes human speech and converts it into text by a computer. This allows users to interact with computers through voice commands or voice input. This speech recognition technology is also known as Speech to Text (STT). STT involves voice input, digital signal processing (DSP), feature extraction, acoustic modeling, and language modeling. The user enters information by voice through a microphone, and then the voice signal is converted from analog to digital. This is followed by digital signal processing, which involves preprocessing to remove background noise, cancel echoes, etc. Then, important features are extracted from the digitized speech. This is usually done using an algorithm such as MFCC (Mel Frequency Cepstral Coefficients) [6]. The extracted features are then applied to a pre-trained model such as an HMM or DNN for acoustic modeling to recognize each word or pronunciation. Finally, linguistic modeling is performed by selecting the words that are most naturally connected into a sentence from several word candidates. STT technology is used in personal assistant services on smartphones such as Siri, Google Assistant, and Bixby, automatic translation services, and assistive devices. Other applications include customer service centers and education. STT technology is also evolving with advances in AI and machine learning. Deep learning methodologies such as LSTM (Long Short-Term Memory) and Transformer have contributed to the improvement of STT's performance. However, it is still difficult to achieve 100% accuracy due to unclear pronunciation, various accents and dialects, and noise.

Text Mining refers to the process of extracting useful information from large amounts of text data by combining Natural Language Processing and Data Mining techniques. It is used to discover and analyze patterns, trends, information, and statistical characteristics from text data [7]. The main purposes for which it is used are information retrieval, sentiment analysis, topic modeling, document classification, and information extraction. Information retrieval is concerned with finding and ranking relevant documents for a user's search query. It is one of the core technologies used by web search engines. Next, sentiment analysis is used to analyze the sentiment or opinion of a particular text to determine if it is positive, negative, or neutral. Subject modeling is a technique for identifying and grouping key subjects in large amounts of text data. This allows you to discover hidden patterns in text data. Document classification is the categorization of text into predefined categories or categories, which can be used for spam filtering, news article categorization, legal document categorization, and more. Information extraction is the process of extracting important information from text. Text Mining does this by applying machine learning and statistical analysis techniques to text data.

Automatic metadata expansion and recommendation systems have been studied before. For example, T. Tsunoda et

ISBN 979-11-88428-12-0 ICACT2024 February 04 ~ 07, 2024

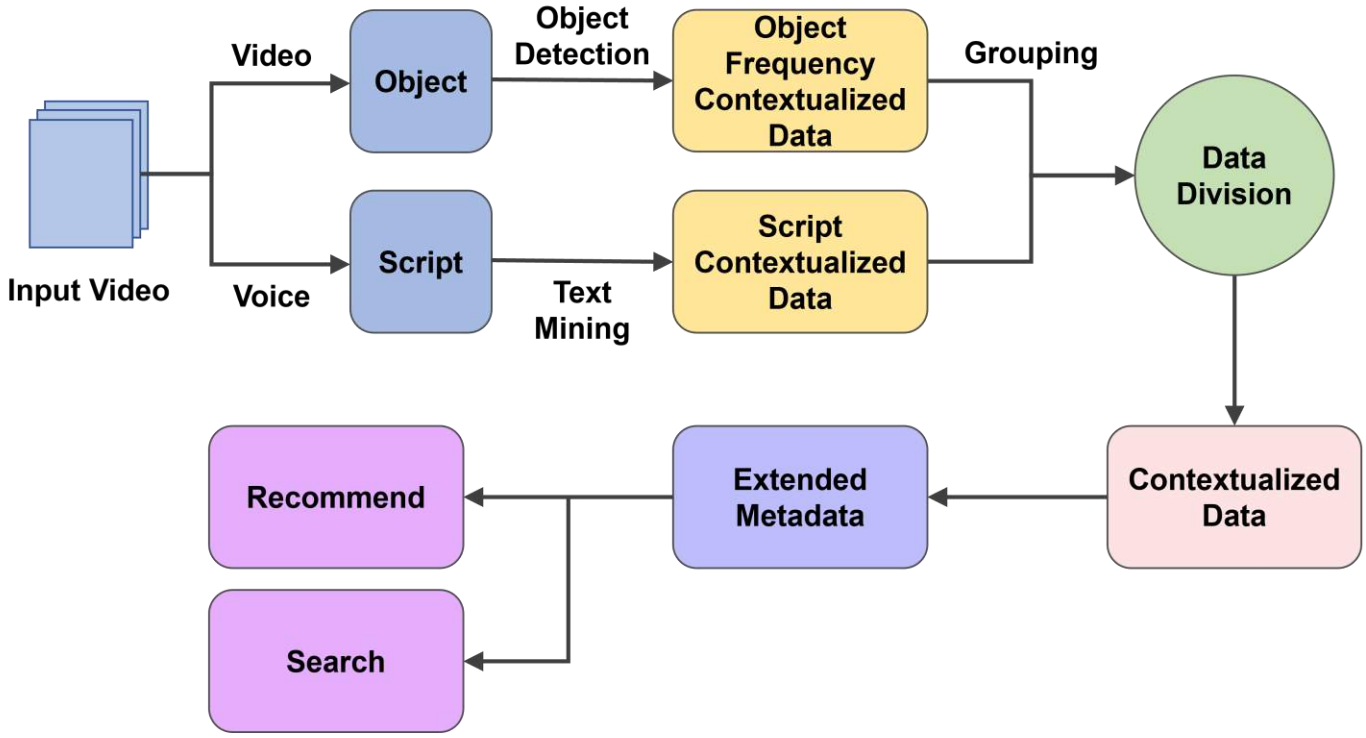

Figure 1. Search and recommendation Process based on Metadata Extensions

al. proposed "Automatic metadata expansion and indirect collaborative filtering for TV program recommendation system" [8]. They proposed two methods to improve the accuracy of TV program recommendations. The first proposed automatic metadata expansion (AME) and enrichment of TV program metadata from electronic program guide (EPG). The second is an indirect collaborative filter (ICF) to recommend non-persistent items such as TV shows based on the preferences of other members of the community. The proposed methods can generate rich data about target items. In addition, the recommendation accuracy can be improved not only by the user's preferences but also by the preferences of other users in the community. However, there are some challenges. First, AMEs can generate common profiles, such as lifestyle. This creates a dependency on users for sensitive information, which needs to be addressed. The second is that there is a lot of computation on attributes as parameters are set for each user to improve accuracy. Therefore, a way to reduce this computation is needed.

III. Search and Recommendation Systems with Metadata Extensions

In this study, we use a YOLO deep learning model to extract objects from videos frame by frame to develop a metadata expansion-based search and recommendation system. At the same time, we extract speech from the video separately and convert it into a script using STT technology. The extracted script data is then segmented into scenes. Then, we synchronize the object and text data to create a single contextual data. We check whether the context data is appropriate for the video and fine-tune it by making adjustments. Index the generated contextual data by grouping it with the video. Expand the metadata based on the indexed information. Include the expanded metadata in search or recommendation to improve the performance of the search and recommendation system. Figure 1. illustrates the process of generating metadata from videos.

A. Data Preprocessing

Select key objects required in a video recommendation system based on object frequency. We use the YOLO deep learning model to train a deep learning-based object recognition model on the labeled image dataset to detect and extract objects based on frames. Since the YOLO deep learning model is trained using the COCO Dataset, it can detect and classify a total of 91 objects, from 1 to 91. In addition, we extract the voice from the video separately and then use STT technology to extract the script according to the frame. To apply machine learning techniques to the extracted script in natural language form, we apply text mining techniques to quantify it.

Text mining libraries that can be used include Python's Natural Language Toolkit (NLTK), spaCy, gensim, scikit-learn, TensorFlow, PyTorch, which utilize deep learning techniques. The process of quantifying through text mining involves many different steps. For our proposed method, the text is tokenized by breaking it into small chunks and going through lemmatization to remove unimportant words, root extraction to find the basic form of words, and part-of-speech tagging to understand sentence structure and meaning.

B. Extending Metadata with Contextual Data

Object frequencies and scripts extracted from videos have the characteristics of time series data. Object frequencies are expressed as the number and type of objects observed in each frame. Scripts appear across multiple frames and are generated as contextual data from the frame where the dialog starts. The object frequencies and scripts are then synchronized. When dividing the context data into scenes, we do so based on scripts. We use iterative experiments to divide the scripts into appropriately sized sentences or paragraphs, as too small a division would clutter the data, too large a division would distort the features.

C. Grouping Video – Clips and Indexing

Since video content has a time-series nature over time, changing with events or scene transitions, Shot Boundary Detection segments videos based on content, and then analyzes and processes videos as content units.

D. Fine-tuning

In this paper, we propose a video-clip indexing model using contextual data based on intelligent content analysis. This model enables fine-tuning using contextual data extracted from videos. First, we use scripts extracted using STT technology. By comparing the extracted script with the subtitle data embedded in the video, the accuracy of the extracted script is checked and calibrated.

Adjust the hyper-parameters used to train the model and improve the model performance through cross-validation. For cross-validation, compare the extracted script with the object frequency context data obtained using the YOLO deep learning model. We then evaluate and test the developed model. This can be evaluated by putting some videos into the test and checking if the correct metadata is extracted. Figure 2. shows the fine-tuning process of the proposed model.

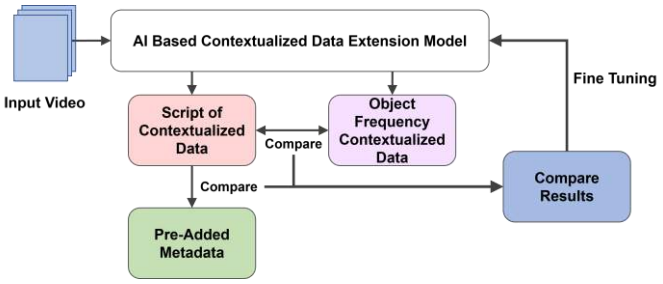

Figure 2. Process of Model Fine-Tuning

Use it to develop a proof-of-concept video-clip search program and integrate it into your environment. Continuously evolve the model through user feedback and performance monitoring, and then compare it to traditional search and recommendation methods to improve performance.

IV. CONCLUSIONS

In this paper, we proposed a model to enhance search and recommendation systems by extending metadata with contextual data extracted from actual content content through AI-based content analysis.

If this model is applied to various content search and recommendation platforms, the extended metadata can be used to provide results that are close to the search query when the user enters the search query. It can also be used to help users discover videos that contain similar content based on the video. This will provide a way to discover more similar videos and clips by allowing metadata to reflect the actual content of content that was previously excluded. This will compensate for problems with existing search, recommendation, and censorship systems, enabling more in-depth recommendations. It can also be extended to various video platforms, such as managing videos with long running times, such as CCTV, by splitting them. Therefore, it can play a role in saving time and manpower in parts of the social safety surveillance system such as CCTV control centers. However, there is a limitation in that human conversations can be transcribed by applying STT technology, but background sounds such as car sounds, or other sounds such as animal cries cannot be transcribed. Therefore, in future research, we will solve the problem of analyzing other sounds based on the proposed model.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(No: RS-2023-00248899).

REFERENCES.

[1] J. C. Kim and K. Y. Chung, “Knowledge expansion of metadata using script mining analysis in multimedia recommendation,” *Multimedia Tools and Applications*, vol. 80, pp. 34679-34695, Mar. 2020.

[2] M. Ekstrand and J. Riedl, “When recommenders fail: predicting recommender failure for algorithm selection and combination,” in *Proc. RecSys'12*, 2012, p. 233-236.

[3] A. Ishikawa, E. Bollis and S. Avila, "Combating the elsagate phenomenon: Deep learning architectures for disturbing cartoons," in *Proc.IWBF'19*, 2019, p. 1-6.

[4] A. Yousaf., A. Mishra., B. Taheri and M. Kesgin, “A cross-country analysis of the determinants of customer recommendation intentions for over-the-top (OTT) platforms,” *Information & Management*, vol. 58, no. 8, pp. 103543, Oct. 2021.

[5] S. H. Abdulhussain, A. R. Ramli, B. M. Mahmmod, M. I. Saripan, S. A. R. Al-Haddad, and W. A. Jassim, “Shot boundary detection based on orthogonal polynomial,” *Multimedia Tools and Applications*, 78, pp. 20361-20382, Feb. 2019.

[6] C. Ittichaichareon., S. Suksri. and T. Yingthawornsuk, “Speech recognition using MFCC,” In *Proc. ICGSM'12*, 2012, p. 135-138.

[7] A. H. Tan, “Text mining: The state of the art and the challenges,” In Proc. KDAD'99, 1999, p. 65-70.

[8] T. Tsunoda. and M. Hoshino, “Automatic metadata expansion and indirect collaborative filtering for TV program recommendation system. Multimedia Tools and applications,” Multimedia Tools and applications, vol. 36, pp. 37-54, Jan. 2008.

Woo-Hyeon Kim (M'23) was born in Geoje-si, Gyeongsangnam-do, South Korea, July 12, 2003.
In 2022, he enrolled in the Division of AI Computer Science and Computer Engineering at Kyonggi University, majoring in Computer science.
His educational background includes:
- Bachelor's Degree:Computer Science, Kyonggi University, South Korea, 2022
His Research interests encompass Data Mining, Big Data, and Anomaly Detection

Joo-Chang Kim has received B.S. and M.S. degrees from the School of Computer Information Engineering, Sangji University, South Korea in 2014 and 2016, respectively. He has received Ph.D. from Department of Computer Science, Kyonggi University, South Korea in 2021. He is currently a research professor in Contents Convergence Software Research Institute, Kyonggi University. Since 2021, he is currently a lecturer in the Department of Software Convergence Engineering, Inha University, South Korea. His research interests include data mining, data management, knowledge systems, machine learning, deep learning, big data, healthcare, and recommendation systems.
.

ISBN 979-11-88428-12-0 ICACT2024 February 04 ~ 07, 2024