Extracting and Restructuring of Multimedia Contents based on Metadata Analysis

Extracting and Restructuring of Multimedia Contents based on Metadata Analysis

Honoka KAKIMOTO

Abstract:

Our research aims to support viewers of multimedia contents, including video, to raise their understanding and interest in the contents, and to obtain the information they want. There are three approaches in this research: video supplementation, learning support, and personalization. In chapter 2, we propose and evaluate a system that recommends related and detailed information according to the scene, using a Web page extraction method that uses location names and topics included in the closed caption data of the travel program. In chapter 3, we define and extracts the editing features included in the movie trailers and evaluate the bias of editing features to generate a personalized movie trailer. chapter 4 proposes and evaluates a method to extract learner's interest from the interaction log of multimedia devices of a museum to recommend video contents suitable for post-learning. In the conclusion section, we present discussions and conclusions based on the evaluation results shown in this research and discuss future issues.

1. Introduction

In recent years, information technology has become more sophisticated, and Web services have become an environment where information is automatically personalized and recommended. On the other hand, to differentiate services and systems, a more accurate and effective information recommendation system such as user preference prediction based on machine learning has been required. However, personalized information based on the analysis of the user's behavior does not always indicate the user's preference, and the recommended information may not be appropriate. This means that in any service, the amount of information that can be automatically obtained is large, and the user needs to select information. Therefore, we focused on the use of multimedia metadata to solve a problem that makes it difficult for users to select information. For example, information recommendation for video on TV and the video streaming service of the web has been actively researched. However, since the video contains various information, the supplementary information about the video required by the viewer is also various. Also, the recommended information is often of the same type of media, such as a video similar to the viewed videos, and it is difficult for a user to search a plurality of pieces of information immediately when he wants to search for something. To meet such diverse demands of users, it is necessary to provide cross-sectional information with Multimedia information. On the other hand, there are fields other than the Web that require support using multimedia metadata. In recent education, interactive learning using devices in classes has been actively conducted. In such learning as well, the information that learners can obtain has increased, and their interests and desired information have diversified.

Based on these backgrounds, our research aims to support viewers of multimedia contents, including video, to raise their understanding and interest in the contents, and to obtain the information they want. There are three approaches in this research shown as figure 1.

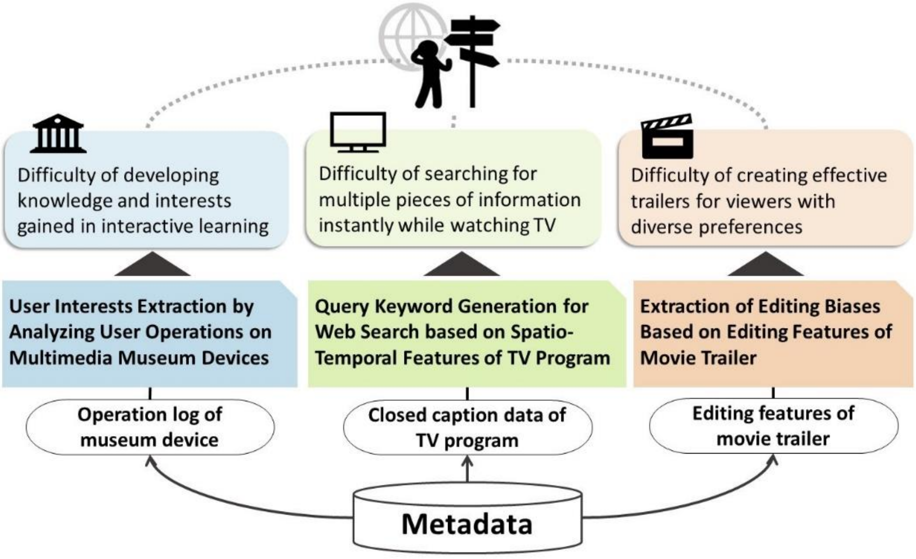

Difficulty of developing knowledge and interests gained in interactive learning
Difficulty of searching for multiple pieces of information instantly while watching TV
Difficulty of creating effective trailers for viewers with diverse preferences
User Interests Extraction by Analyzing User Operations on Multimedia Museum Devices
Query Keyword Generation for Web Search based on Spatio-Temporal Features of TV Program
Extraction of Editing Biases Based on Editing Features of Movie Trailer
Operation log of museum device
Closed caption data of TV program
Editing features of movie trailer
Metadata

Figure 1 Our approaches

2. Query Generation for Web Search based on Spatio-Temporal Features of TV Program

2.1 System overview

In recent years, various TV programs are broadcast, and it is common to search for information of interest on the Web. Therefore, the information that the viewer additionally desires for the topic occurs while they are watching the program. In particular, the travel program introduces a wide variety of information such as culture and history. Therefore, the information that the viewer wants to search occurs simultaneously or continuously, but it is difficult for viewers to enter a query suitable for searching or select media that contains the information that they wanted while they are watching the video. Our system aims to provide supplementary information and related information along with the video. we aim to achieve the following three goals.

- ●Extraction of regional information based on keyword appearance intervals and appearance frequencies in caption data of travel programs.
- ●Query generation for searching Web pages as detailed information and related information on topics of the scene
- ●Evaluation of detailed information and related information recommendation systems for video scenes

Figure 2 shows a system flow. First, nouns are extracted from caption data by morphological analysis. From these nouns, location names and topic keywords are extracted for each scene based on the spatio-temporal features of keywords in caption data. To recommend Web pages, location names and other topic keywords are extracted, and keywords to be used for query generation are determined based on frequency and interval of the appearance. Then a detailed search query and a related search query are generated with keywords. After that, a detailed page that supports the viewer's understanding of the content of the scene and a related page that supports the generation of interest is extracted and recommended.

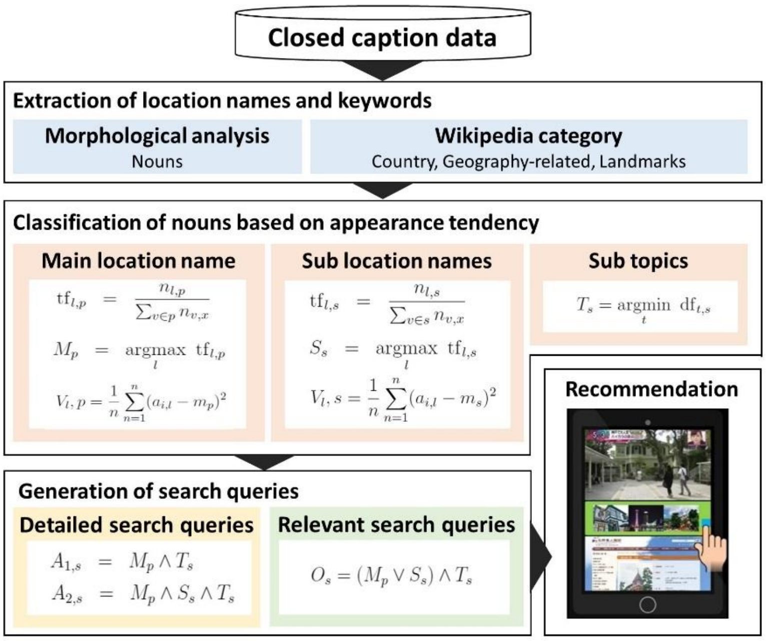

Closed caption data
Extraction of location names and keywords
Morphological analysis
Nouns
Wikipedia category
Country, Geography-related, Landmarks
Classification of nouns based on appearance tendency
Main location name
$\mathrm{tf}_{l,p} = \frac{n_{l,p}}{\sum_{v \in p} n_{v,x}}$
$M_p = \operatorname*{argmax}_l \mathrm{tf}_{l,p}$
$V_l, p = \frac{1}{n} \sum_{n=1}^{n} (a_{i,l} - m_p)^2$
Sub location names
$\mathrm{tf}_{l,s} = \frac{n_{l,s}}{\sum_{v \in s} n_{v,x}}$
$S_s = \operatorname*{argmax}_l \mathrm{tf}_{l,s}$
$V_l, s = \frac{1}{n} \sum_{n=1}^{n} (a_{i,l} - m_s)^2$
Sub topics
$T_s = \operatorname*{argmin}_t \mathrm{df}_{t,s}$
Recommendation
Generation of search queries
Detailed search queries
$A_{1,s} = M_p \wedge T_s$
$A_{2,s} = M_p \wedge S_s \wedge T_s$
Relevant search queries
$O_s = (M_p \vee S_s) \wedge T_s$

Figure 2 System flow 1

2.2 Preliminary Experiment

In preliminary experiments, we investigated what kind of search query was created when a TV program viewer searched while watching a video. The results showed that viewers tended to search using location names while watching travel programs. Based on the experimental results, this paper proposes two types of recommendation methods, related information and detailed information, to support viewing of TV programs.

2.3 Query Keyword Extraction based on Spatio-Temporal Features

In this section, we explain a method to extract three types of the main location name, sub place names, and subtopics from the nouns extracted from the caption data of a travel program. First, nouns are classified based on the frequency and frequency of appearance. A place name that appears most frequently in the closed caption data of the entire video is the main location name. Location names that appear most frequently in each scene of the program are sub location names, and keywords that appear most frequently in each scene of the program are extracted as sub topics. In the proposed method, a query is generated using three types of keywords extracted from caption data. The purpose of AND search is to extract a detailed page that assist viewers in understanding the content of the program. On the other hand, AND-OR search aims to extract related pages that evokes viewers' interests.

2.4 Evaluation

We evaluated Web pages recommended by the proposed method. In the evaluation, three videos of a travel program of 5 minutes each were used as target data, and for each video, nine Web pages searched and recommended based on a query generation method were presented together with the video. Evaluation results show that viewing support is effective when the viewer's interest is raised by the details pages of the location introduced in the scene and the related page containing information of other locations. To support content understanding more effectively, it is necessary to indicate the relevance of recommended web pages and scenes by displaying a part of the search query with the web page. Also, it is necessary to extract pages that contain supplementary information, excluding reservation sites, etc.

3. Extraction of Editing Biases Based on Editing Features of Movie Trailer

3.1 System overview

Since trailer is edited for a certain target audience, the scenes and effects that can be included in the trailer are limited. Therefore, it is difficult to edit a trailer that caters to the various users in the target audience. However, the viewer could be lost if the trailer is not attractive enough to them. Figure 3 shows flow of this research. To solve the problem, we define seven editing features that occur when movies are summarized and edited into trailers based on definitions in related work and editing features of a movie. In addition, we evaluate the editing features in movie trailers through two evaluation: user study and evaluation based on the result of genre prediction model.

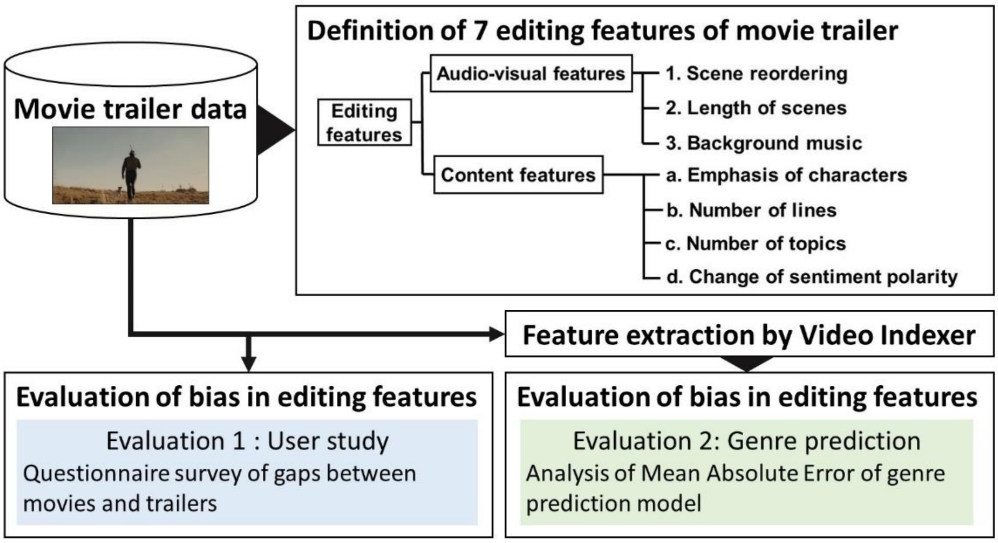

Movie trailer data
Definition of 7 editing features of movie trailer
Editing features
Audio-visual features
1. Scene reordering
2. Length of scenes
3. Background music
Content features
a. Emphasis of characters
b. Number of lines
c. Number of topics
d. Change of sentiment polarity
Feature extraction by Video Indexer
Evaluation of bias in editing features
Evaluation 1 : User study
Questionnaire survey of gaps between movies and trailers
Evaluation of bias in editing features
Evaluation 2: Genre prediction
Analysis of Mean Absolute Error of genre prediction model

Figure 3 System flow 2

3.2 Evaluations

The main purpose of evaluation 1 is to analyze which of the seven editing features have a large influence on the impressions of viewers. In addition, we determined which editing feature impedes a viewer's understanding of the content of a movie. To analyze the influence of the seven editing features, we presented a questionnaire using the seven movies. The results of evaluation 1 show that scene reordering, length of scenes and number of topics have the largest influence on the differences between a viewer's impressions of the trailer and film. In addition to the preliminary evaluation, we evaluated the strength of bias of editing features based on the result of the genre prediction model. The bias of each feature is determined based on the Mean Absolute Error of genre prediction model. We determined that editing features with low error have weak bias and editing features with high error have strong bias based on the results. In evaluation 2, it was found that a strong bias appeared in length of scenes, emphasis of characters and change in sentiment polarity. Length of scenes appeared as a strong bias in both evaluations. The bias of other editing features was different between evaluation 1 and evaluation 2. As a result, it became clear that length of scenes appears most strongly as bias among the seven editing features defined in this chapter.

4. User Interests Extraction by Analyzing User Operations on Multimedia Museum Devices

4.1 System overview

In recent years, information on museum exhibits has been digitized, and the information that visitors can obtain about museum exhibits has also diversified. For this reason, on-site learning using multimedia devices in museums and online learning platforms using video and VR are also provided. There are three types of learning: pre-learning, on-site learning, and post-learning. Pre-learning and on-site learning are active now, but there is not enough support for the learner to deepen the interest gained from on-site

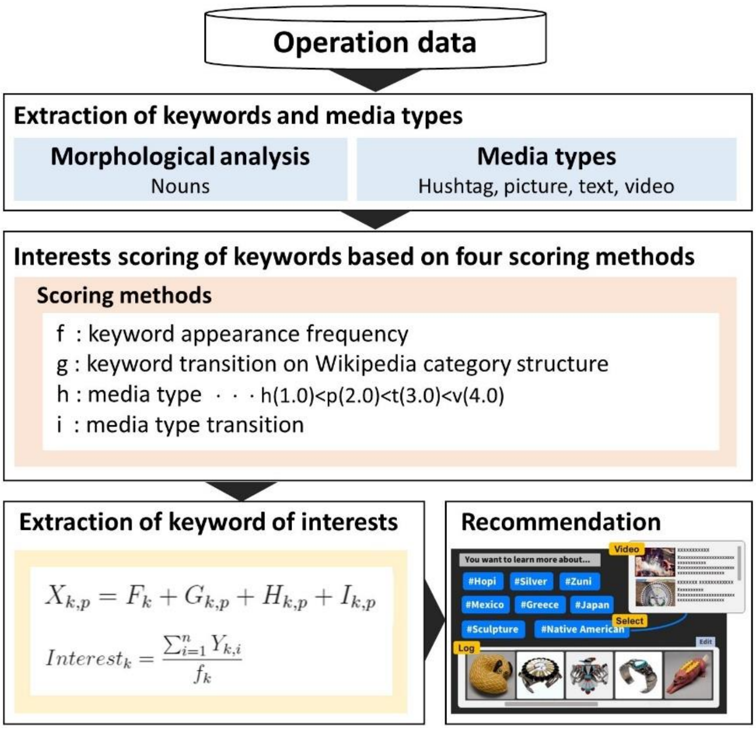
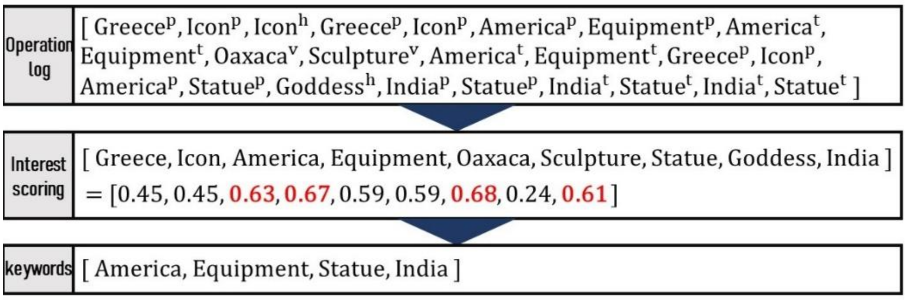

Operation data
Extraction of keywords and media types
Morphological analysis
Nouns
Media types
Hushtag, picture, text, video
Interests scoring of keywords based on four scoring methods
Scoring methods
f : keyword appearance frequency
g : keyword transition on Wikipedia category structure
h : media type ・・・h(1.0)<p(2.0)<t(3.0)<v(4.0)
i : media type transition
Extraction of keyword of interests
$X_{k,p} = F_k + G_{k,p} + H_{k,p} + I_{k,p}$
$Interest_k = \frac{\sum_{i=1}^{n} Y_{k,i}}{f_k}$
Recommendation
You want to learn more about...
#Hopi
#Silver
#Zuni
#Mexico
#Greece
#Japan
#Sculpture
#Native American
Video
Select
Log
Edit

Figure 4 System flow 3

learning and start more advanced learning. Therefore, the interests gained in the museum need to be linked to the next interests and more effective and scalable learning. In our research, we aim to extract the interests of learners from their interaction on museum multimedia devices and provide post-learning content. Figure 4 shows system flow of our proposed method. In this chapter, to extract learners' interests, we first analyze learners' interactions for exhibits on museum devices. Then, we propose the scoring method based on four features of interaction data: keyword appearance frequency, keyword transition, media type, and media transition.

4.2 Extraction of Learners' Interests

In this section, we extract learner's interactions on a multimedia device of a museum as their interests: tablet-type devices with direct access to videos, pictures, related links, etc. The top part of Figure 5 presents an example of the extraction of keywords from operation log data. All nouns are extracted from words in hashtags, text, caption, title in operation log data of the museum device based on the morphological analysis. Then the keyword string is generated with nouns and media types. In this research, we extract four media types: hashtag, picture, text, and video. To determine keywords as learner's interests, we score on nouns in operation log data based on the following four scoring methods: keyword appearance frequency, keyword transitions on Wikipedia category structure, media type, and media type transitions. These scoring methods were set based on the following three conditions: nouns selected repeatedly by the learner, nouns selected consecutively by the learner, and nouns selected in multiple media by the learner.



Operation log	[Greece^{p}, Icon^{p}, Icon^{h}, Greece^{p}, Icon^{p}, America^{p}, Equipment^{p}, America^{t}, Equipment^{t}, Oaxaca^{v}, Sculpture^{v}, America^{t}, Equipment^{t}, Greece^{p}, Icon^{p}, America^{p}, Statue^{p}, Goddess^{h}, India^{p}, Statue^{p}, India^{t}, Statue^{t}, India^{t}, Statue^{t}]
Interest scoring	[Greece, Icon, America, Equipment, Oaxaca, Sculpture, Statue, Goddess, India] = [0.45, 0.45, 0.63, 0.67, 0.59, 0.59, 0.68, 0.24, 0.61]
keywords	[America, Equipment, Statue, India]

Figure 5 Example of interest scoring

4.3 Evaluation

To evaluate the effectiveness of the proposed method, we compared the evaluation results of keywords extracted by the proposed method (A) with the evaluation results based only on the keyword appearance frequency (B). Evaluation data is the operation log data in which five respondents operated a multimedia device for three minutes and keywords extracted from the log data based on the proposed method. As a result, this evaluation revealed that the proposed method can extract interests that reflect the learner's intention to obtain detailed information on specific topics contained in the exhibits. Therefore, it is possible to extract a keyword of interest suitable for the post-learning for the learner to deepen the interest obtained in the on-site learning. Also, it was found that more effective interest extraction can be expected by incorporating the scoring method media browsing time in the future.

4.4 User Interface

Interface in Figure 3 is designed to visualize a learner's interests in exhibits and recommend video scenes for post-learning. The learner's operation log is displayed at the bottom of the interface, and keywords extracted by interest extraction from the log are displayed as tags. Then, scene(s) related to the tag is recommended by selecting the tag(s). To make the post-learning effective, it is necessary to extract a scene suitable for the post-learning, not the whole video. Also, the learner should be able to dynamically select the extracted scenes. Huh et al. [1] focused on hyperlinks in a system that recommends videos based on learner interest. They proposed a method of extracting the target of the video viewer's interest based on the ER (entity-relationship) model and creating a smart video in the e-learning domain. This smart video contains hyperlinks to other resources such as books, dictionaries, location information, people, and other videos, and these hyperlinks will be effective in deepening the interest and knowledge that the learner has gained in the museum's on-site education. Therefore, our proposed method extracts interest from learner's interaction with multimedia devices in the museum.

5. Conclusion

In chapter 2, we proposed a query generation method to extract location names from caption data of travel programs based on appearance frequency, appearance interval, and Wikipedia category structure, and to search related and detailed information based on spatio-temporal information. In addition, we evaluated related information and detailed information and evaluated the effectiveness of the query generation method. Besides, it was confirmed that pages with a large amount of information about the area in the scene, such as official websites and commentary websites, were effective in raising interest and understanding the contents.

In chapter 3, we defined seven editing features in a movie trailer based on definitions in related work and editing features of a movie. We then proposed a method that can extract seven editing features: scene reordering, scene length, background music, emphasis of characters, number of lines, number of topic words, and change in sentiment polarity. Then we analyzed the bias of editing features in movie trailers based on the results of a preliminary experiment. As a result, it became clear that length of scenes appears most strongly as bias among the seven editing features. Also, if other editing features could be used as parameters to adjust the strength of the editing features catering to the user's preferences, generation and recommendation of personalized trailers will be realized.

In chapter 4, we proposed a method to extract and visualize learners' interests from operation logs of multimedia devices in museums. As future work, we plan to propose a video content extraction method for post-learning support by analyzing the learner's learning styles based on classifications of important topics in textbooks and the tendency of the learner's operations. Bakkay et al. [2] developed a support protocol for the creation of educational video content for teachers without technical background. The goal of our research is to propose a dynamic scene extraction method using hyperlinks between video scenes based on an interest extraction method to recommend video content for post-education. If this goal is achieved, it is expected that teaching using educational video content will be easier for teachers and learners will be able to learn more scalable.

**References ***Cited references in this paper. Other references are listed in the master thesis.

[1] Seyoung Huh, Yoomi Park, onghyun Jang and Wan Choi. 2015. Making a video smart for smart e-learning. In Proceedings of the 2015 International Conference on Information and Communication Technology Convergence (ICTC'15). Jeju, Korea. pp. 858-863. DOI:https://doi.org/10.1109/ICTC.2015.735468

[2] Mohamed Chafik Bakkay, Matthieu Pizenberg, Axel Carlier, E. Balavoine, Graldine Morin, and Vincent Charvillat. 2019. Protocols and Software for Simplified Educational Video Capture and Editing. Journal of Computers in Education 6 (2). pp 25776. DOI:https://doi.org/10.1007/s40692-019-00136-6