Query Keyword Extraction from Video Caption Data based on Spatio-Temporal Features

Honoka Kakimoto, Toshinori Hayashi, Yuanyuan Wang, Yukiko Kawai, and Kazutoshi Sumiya

***Abstract*—Recently, people have begun searching for relevant information of each scene of TV program videos with other devices, such as smartphones and tablets. While viewing TV programs, user's interests change with each scene of the video. When they try to get information related to the content of the scenes, users have to input appropriate query keywords for a Web search. However, it takes users time and effort to find their desired information. Although some data casting services suggest related information to TV programs, the related information does not synchronize enough with each scene of the videos. To solve this problem, our system proposes a novel query keyword extraction method for Web searches, based on spatio-temporal features of videos, using location names in the video caption data. We first extract the location names and classify them into two factors: geographical distance between locations and semantic distance between location names. The system recommends related information by setting the maximum range of each area. Subsequently, we determine the main topics and sub topics for generating web search queries based on the length of time. Therefore, through our system, suitable web pages for each scene can be found based on the generated query keywords.**

***Index Terms*—Closed caption, geographical metadata, geographical relationships, recommender system, topic extraction**

I. Introduction

Recently, people have begun searching for relevant information to each scene of TV programs with other devices, such as smartphones and tablets. While viewing TV programs, user's interests change with each scene of the video. When users want to get information related to the contents of the scenes, they have to input appropriate query keywords for a Web search. However, it takes users time and effort to find their desired web pages until they get the relevant information. In addition, there are certain users, including children and elderly people, who are not able to input appropriate query keywords. Further, it is difficult to search various types of information through the Web at once.

Some data casting services such as NHK Hybridcast [1] and other viewer participating program services recommend related information to TV programs on the interface. However, the recommended information does not synchronize with each scene of the videos. Therefore, it is necessary to recommend information related to each scene and user's concerns.

TV programs are often associated with closed captions, and many researchers proposed systems utilizing topics in closed caption data of videos. In this work, we develop an automatic location-based recommendation system using the concept of the automatic location-based image viewing system synchronized with video clips [2] by adding a real geographical distance and length of video times to generate web search queries.

II. System overview and Related work

A. System Overview

The system flow of our proposed method for generating a web search query based on location names in the closed caption is described as follows.

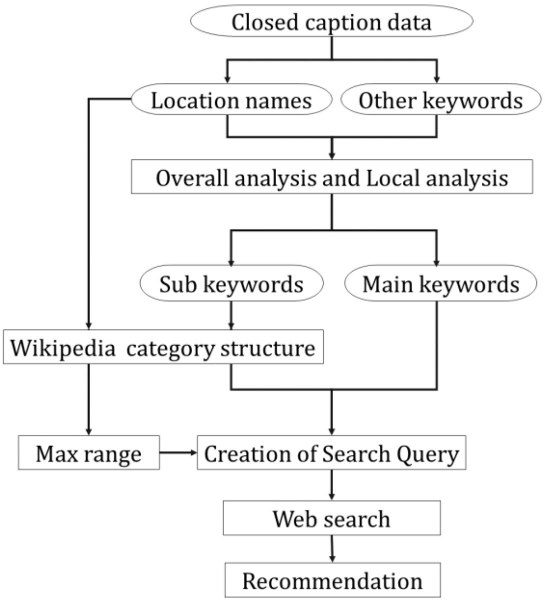


Fig. 1. System flow

First, the system extracts location names in closed captions from an MPEG file of a TV program based on the method of related work. It classifies location names and deletes unnecessary words as an outlier, and sets the maximum range of area that the system recommends based on the location names in closed caption data. Second, the system selects the main keywords and sub keywords of the TV program in two

Manuscript received December 28, 2017. revised February 11, 2018. This work was partially supported by JSPS KAKENHI Grant Numbers 16H01722.

H. Kakimoto is with Kwansei Gakuin University, Sanda, Japan (e-mail: des44337@ kwansei.ac.jp).

T. Hayashi is with Kwansei Gakuin University, Sanda, Japan (e-mail: den82687@ kwansei.ac.jp).

Y. Wang is with Yamaguchi University, Ube, Japan (e-mail: y.wang@ yamaguchi-u.ac.jp).

Y. Kawai is with Kyoto Sangyo University, Kyoto, Japan (e-mail: kawai@ cc.kyoto-su.ac.jp).

K. Sumiya is with Kwansei Gakuin University, Sanda, Japan (e-mail: sumiya@kwansei.ac.jp).

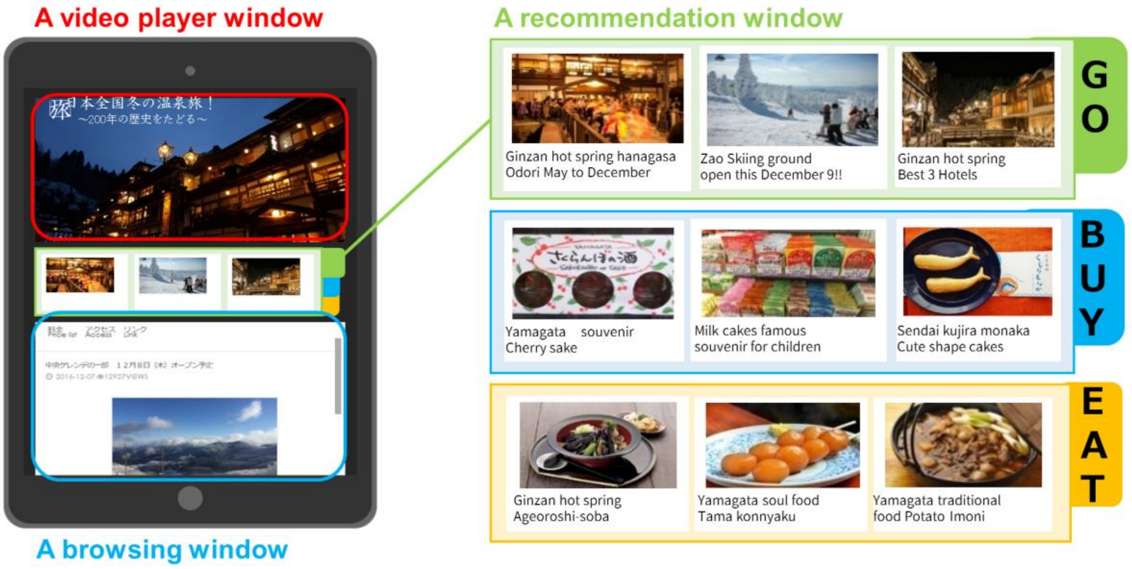

A video player window
A recommendation window
A browsing window
旅 日本全国冬の温泉旅！
~200年の歴史をたどる~
GO
Ginzan hot spring hanagasa Odori May to December
Zao Skiing ground open this December 9!!
Ginzan hot spring Best 3 Hotels
BUY
Yamagata souvenir Cherry sake
Milk cakes famous souvenir for children
Sendai kujira monaka Cute shape cakes
EAT
Ginzan hot spring Ageoroshi-soba
Yamagata soul food Tama konnyaku
Yamagata traditional food Potato Imoni

Fig. 2. User interface

analyses, based on the length of time. Then the system creates a web search query by combining the main keywords and sub keywords. Third, the system searches suitable web pages for the scenes with the search query. Then it recommends several web pages on each tab of the user interface as detailed information and related information to the scenes. Web pages are recommended on three tabs of user interface: “go”, “eat,” and “buy.” An example of system flow is described as follows: A TV program related to the tourist spot along Hokuriku Shinkansen Line has the following flow:

(1) Extract location name: “Niigata,” “Kanazawa,” etc.
(2) Max range: Hokuriku area
(3) Select “Hokuriku Shinkansen Line” as the main keyword.
(4) Select “Hot spring,” “Skiing,” etc. as sub keywords.
(5) Create search query: (Skiing or Hiking) and Hokuriku Shinkansen Line.
(6) Search web pages with the queries.
(7) Recommend web pages about skiing ground or hot spring along Hokuriku Shinkansen Line on the tab “go” of the user interface.

B. *Related Work*

Nishizawa et al [2] extracted a semantic structure of location names in closed caption data by utilizing Wikipedia categories, and detected relevant topics of location names in the semantic structure. In our work, we extract a semantic structure of location names in closed caption data based on their method. In addition, Nishizawa et al [2] proposed location-based image viewing system synchronized with video clips. In their work, the system recommends the images and maps information related to the scenes based on location names in closed caption data of travel video clip. To recommend more suitable information for user’s interests in each scene, we recommend contents from the web pages of e-commerce, travel, and restaurant search sites. Wang et al [3] proposed an automatic video reinforcing system based on popularity rating of scenes and level of detail controlling scenes based on closed caption data. They proposed a novel automatic video reinforcing system with a media synchronization mechanism and a video reconstruction mechanism based on closed caption data. Their work recommended some web content such as YouTube video clips and images related to the scene of a video clip. To recommend suitable information for a travel TV program, this work recommends web pages related to tourism and local specialties.

III. Query Keyword Extraction Based on Spatio-Temporal Features

A. *Extraction and Classification of Location Names*

We extract location names and other keywords from closed captions from an MPEG file of TV program based on the method of related work [2]. Then, we classify location names into two factors: geographical distance between locations and semantic distance between location names. First, we delete the location names that extremely deviate from a group of other location names in a closed caption. Subsequently, we set a maximum range of recommendation with the rest of the names. Second, to recommend relevant web pages effectively, we set a maximum range of the recommendation of web pages. We measure the semantic distances from location to location by using the semantic structure based on the Wikipedia category structure for the creation of a web search query, as described in detail in section C.

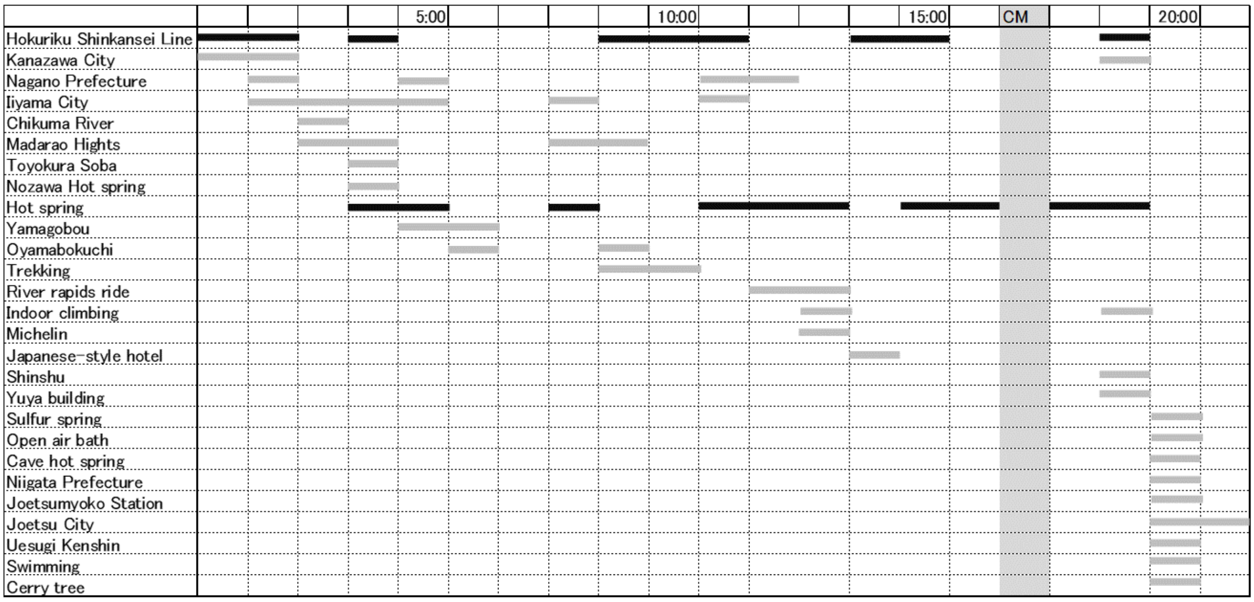
B. *Determination of Main Topics and Sub Topics*

With the keywords in closed captions, we select main topics and sub topics of the TV program to generate the web search query. Table 1 shows the appearance of location names and keywords related to tourism extracted from closed caption data of a 20-min TV program [4]. In this work, closed captions of TV commercials are not extracted. The first row of Table 1 shows the time sequence. The black lines show keywords that appear periodically in the TV program, and gray lines show other keywords. The keywords that appear periodically in closed captions are determined as main topics, and other keywords are determined as sub topics based on the length of time.

In this work, we consider the keywords that appear for a long time as the overall analysis. The main objective of the overall analysis is to define the keywords that appear periodically in the closed captions as the main topics: the words that present the main theme or the atmosphere of the

TABLE I

LOCATION NAMES AND KEYWORDS RELATED TO TOURISM IN CLOSED CAPTION DATA



TV program potentially. In this analysis, we select at least one location name for recommendation related to the location in the closed caption. Therefore, we increase the priority of main topics to enhance the accuracy of Web search. It means that semantic interpretation of the main topics changes based on each scene of the video. As a result, the system can generate various web search query keywords.

In addition, we deal with keywords that appear for short time as the local analysis. To generate search query keywords to recommend detail information, we combine the overlapping topics. Similar to overall analysis, main topics also take priority in the local analysis.

C. Generation of Web Search Query

To search web pages related to the scenes, we create a query in two ways based on the location names and other keywords that are classified: AND search and AND-OR search. First, we generate AND search query with the main location name and some sub topics based on local analysis to search and recommend detail information. Second, we generate AND-OR search query with the main location name and other location names and some sub topics based on the overall analysis. Other location names and sub topics are in parallel relationships based on Wikipedia category structure. Finally, the system searches web pages related to each scene of TV program from travel, e-commerce, and restaurant search sites based on the search queries. Then the results of the search are recommended.

D. Recommendation of Web Pages

In this work, web pages searched with a search query are recommended on three tabs of the user interface, as shown in Figure 2. At least three web pages are recommended on each tab. Tab “go” displays web pages related to tourism and events. Tab “buy” and “eat” display web pages that are searched from e-commerce and restaurant search sites. Users change the tabs based on their interests. To enhance operability, this work assumes the use of the system on a tablet. Users tap on a thumbnail image and browse the web pages on a browsing window while watching the TV program on a video player window.

IV. PRELIMINARY EXPERIMENT

We plan to conduct a questionnaire survey with several short videos of TV program related to tourism, and the group of the keywords in the closed caption data. The purpose of the survey is to analyze the trend of combinations of the keywords that viewers use while watching TV programs. Each video is divided based on the scene changes. To analyze the trend of search queries for relevant information to the scenes, viewers add at least one keyword to make a search query. We plan to provide the questionnaires to the students. After they watch several videos, they answer the following questions:

Q1: Please circle all keywords that you are interested in.

Q2: Please make web search keywords with the keywords that you circled in Q1, and at least one keyword that you want to add.

V. CONCLUSION

In this paper, we proposed query keyword extraction method for Web search based on location names in the closed caption data of videos. First, we extracted the location names and classified them into two factors: geographical distance between locations and semantic distance between location names. Then, we determined main topics and sub topics based on the length of time. Next, we generated a web search query by combining main topics and sub topics, and the suitable web pages for scenes were found based on the generated web search query. Then, we recommended detail information and related information through the user interface of our proposed system.

In the future, we plan to evaluate our system with a questionnaire survey. For this, we plan to make several demonstration videos with TV programs about travel and tourism. In addition, we plan to construct a dictionary to extract appropriate keywords to TV programs related to travel and tourism. Another future direction is to recommend information with various types of media such as comments on SNS, videos, reviews, etc. to expand the user’s interests.

REFERENCES

[1] NHK Hybridcast (2013) Accessed 7 September 2017. Retrieved from http://www.nhk.or.jp/hybridcast/online/

[2] Y. Wang, M. Nishizawa, Y. Kawai, K. Sumiya, "Location-based image viewing system synchronized with video clips," in *Proc. of the 13th International Conference on Location Based Services (LBS 2016)*, pp. 233–238, Vienna, Austria, November 2016.

[3] Y. Wang, Y. Kawai, K. Sumiya, Y. Ishikawa, "An automatic video reinforcing system based on popularity rating of scenes and level of detail controlling," in *Proc. of the 2015 IEEE International Symposium on Multimedia (ISM 2015)*, pp. 529–534, 2015.

[4] "Let's go somewhere. It is not only Kanazawa! The best Hokuriku Shinkansen Line stations you should get off," TV TOKYO, April. 2015.

[5] H. Fleites, H. Wang, S. Chen, "Enabling enriched TV shopping experience via computational and temporal aware view-centric multimedia abstraction," in *Proc. Of IEEE Transactions on Multimedia*, vol. 17, no. 7, pp. 1068–1080, 2015.

[6] K. Tanaka, K. Tajima and T. Sogo, "Algebraic retrieval of fragmentarily indexed video," *New Generation Computing*, vol. 18, no. 4, pp. 359–374, December 2000.

[7] S. Pradhad, "An Algebraic Query Model for Effective and Efficient Retrieval of XML fragments," in *Proc. of VLDB*, pp. 295–306, 2006.