Regional Information Video Searches Using Word Searches Generated by Twitter Posts

Masahiro TAKEDA*, Nobuyuki KOBAYASHI†, Fumio KITAGAWA‡, and Hiromitsu SHIINA‡
*Graduate School of Informatics, Okayama University of Science, Okayama, Japan, Email: i15im02tm@ous.jp
†Faculty of Human Sciences, Sanyo Gakuen University, Okayama, Japan, Email: koba_nob@sguc.ac.jp
‡Faculty of Informatics, Okayama University of Science, Okayama, Email: {kitagawa,shiina}@mis.ous.ac.jp

***Abstract*—There are currently many services available on the Internet including map search sites such as Yahoo!Map, and video search sites such as YouTube . However, there is a limit to the information that can be obtained from each of these services. For example, on map search sites you can find the location of an establishment on a map; however, you cannot obtain information about the establishment itself or view videos of events being held there. On the other hand, although you can find videos on video sites, it is then difficult to find out the locations where the videos were taken in the first place. In addition, tourist information sites are able to provide information such as images and locations, but they offer few videos of events or detailed information about the establishment, limiting the overall amount of information that you are able to obtain at any one time. Our system is, therefore, based on the idea of combining multiple services to obtain many different types of useful information. However, completeness of the search results is low based on a search only using the facility name. By extracting words related to facilities from Twitter posts related to the searched facility and also using that word, many variations to the search results were given. In order to extract words related to facilities from Twitter posts, the results of a machine based learning classification were used in advance to determine whether or not the Tweet relates to a facility.**

***Keywords*—Mashup technology, Video search, Twitter, TF-IDF, Support vector**

I. Introduction

With the development of the environment surrounding the internet, a large volume of information is provided. For example, on YouTube[1], a video sharing service, individuals can watch and upload videos for free and can easily enjoy videos taken at tourist sites and events. Also, community sites like Twitter[2], which provide information services allowing for short text called Tweets to be posted, are used as a place to share information between users and for many companies to place advertisements.

In addition, development has been made to offer tourism information through computer systems. The tourist information system is one example, and specifically through a Japanese voice response, a system such as the tourism concierge that supports Kyoto tourism is being developed[3][4].

Further, by automatically collecting videos and blogs about the event, and mapping that on a map, an event participation support system where users can view event information has been developed [5][6][7]. However, there are problems, such as difficulties with accurately measuring position information solely based on blog of information as well as time consuming procedures associated with collection and analysis of blogs.

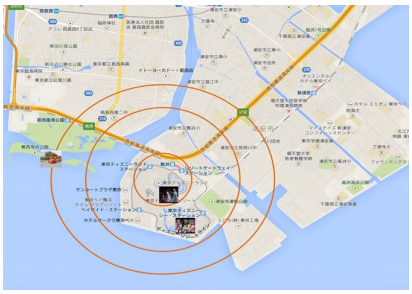

Fig. 1. Image of region search on the video search system

On the other hand, although several internet services such as the map search site Yahoo! Map[8] and the video search site YouTube[9] have been developed, information that can be obtained from each service is limited. For example, although the location on the map can be identified from the map search site, information on the facilities and videos of the events performed there cannot be obtained. Conversely, although videos can be obtained from the video site, it is difficult to discern the location of where the video was shot. In addition, on the tourism site, although it is possible to obtain information on the position and the image, there is little facility information and event videos, and the information that can be obtain at once is limited. So then, by combining multiple internet services and by obtaining numerous information that are useful, it is considered possible to offer tourism video information that is closer to the facility being searched.

In this system, multiple map and video internet services used frequently on a daily basis have been integrated, and a video search system for tourism information was developed. This systems assumes that travel and tourism information about events and facilities of areas surrounding the destination site will be offered via video to those who are planning on tourism. With regards to the increase in specialization and variations, other than the location and facility name for the desired search, it was realized by also using words related to tourism for the search. However, completeness of the search results is low based on a search only using the facility name. By extracting words related to facilities from Twitter posts related to the searched facility and also using that word, many variations to the search results were given. In order to extract words related to facilities from Twitter posts, the results of a machine based learning classification were used in advance to determine whether or not the Tweet relates to a facility.

978-1-4799-9958-3/15 $31.00 © 2015 IEEE
DOI 10.1109/IIAI-AAI.2015.256

Fig. 2. Search screen of regional video search system (if the facility searched was Tokyo Disneyland)

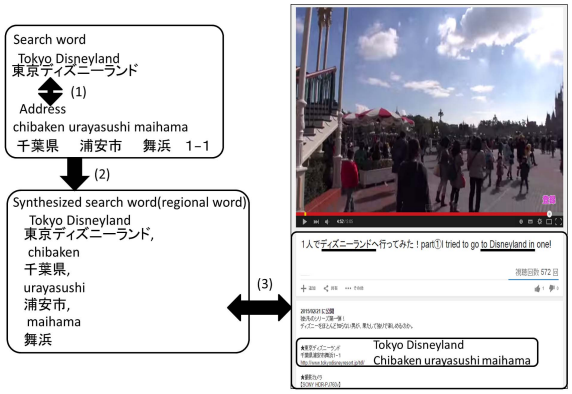

Fig. 3. Extraction of regional word (if the facility searched was Tokyo disneyland)

II. Video Search System for Tourism Information

The purpose of this system is to provide travel and tourism videos of the surrounding areas of the destination site (Figure1). For example, when the facility search name was noted as “Tokyo Disneyland”, facilities focusing on “Tokyo Disneyland” were identified, and videos of these main facilities and those in the surrounding areas such as the “Tokyo DisneySea” and “Kasairinkai Park(Tokyo Sea Life Park)” were searched.

Indicate operating screen for the video search system in Figure 2. The top left pane of the figure is the designated the search conditions, the middle left pane is the map of the facilities searched, and the right side pane is the search results. For the search results, the thumbnails of the videos and the comments in relation to the facilities and the surrounding facilities according to the search condition are displayed. When the thumbnail is clicked, the YouTube video starts.

A. Input Items for the Search

The input form for the search criteria consists of six items: (1) address, (2) facility name, (3) categories of surrounding facilities of search (4) maximum number of surrounding facilities in the search (5) search range and (6) maximum search video number.

B. Procedures for Video Search of Tourism Information

The following is the video search procedures for tourism information. Figure 3 illustrated extraction of regional words from search word, address and video comments.
(1) Based on the search items and conditions entered, obtain the facility name using Yahoo!OpenLocal Platform(YOLP)[8], address and longitude/latitude location information using Google Geocoding API[10].
(2) Based on the address and facility, extract from the nouns, the word determined to be location. Note that this word is referred to as the regional word using morphological analysis[11][12].
(3) Based on the facility name, obtain the YouTube video, video title, and the video details using Youtube API([9]).
(4) Extract the regional word from the video comments, and compare to the regional word obtained in (2), and present the video.
(5) Based on the facility name and the location information being searched, search the surrounding facilities using Yahoo!localSearchAPI[8], and take the facility name acquired as the facility name and repeat the video search for the maximum video search number as specified in (1).

III. Search Word Extraction of Regional Information Using Twitter

A. Process for Search Word Extraction of Regional Information

The method noted in section II, extracts regional words related to facility names focusing on the facilities subject to the search. For videos obtained from the search words, there are many videos that are useful for a particular region, but are unrelated to tourism or events. Therefore, by increasing the variations of the search results, volume and accuracy of the tourism information is improved. As a method to increase the variations of the video, an attempt was made to use a method that creates a new search word by adding tourism information such as event and trend information to the facility name of the search word. As a method to obtain tourism information, posts to Twitter, where many opinions are posted in real-time, were utilized. In addition, as a means for removing unnecessary posts, the support vector machine(SVM) machine[13][14][15] learning method was utilized.

(1) Obtain Post from Twitter Based on the facility name and the location information that a search will performed on, obtain multiple collective Twitter posts from the surrounding facilities.

(2) Calculating the characteristic index of a word and morphological analysis of Twitter post. For each word in the Twitter post, obtain the characteristic index of words arranged by TF-IDF, and using that as an element, create a feature vector(Figure 4).

(3) Creating a search word Based on the SVM that had been previously learned, classify the feature vectors, and extract the Twitter post classified as facility information. From the posts obtained, extract the feature words by using the Yahoo! key phrase extraction. Then, perform a AND search for the

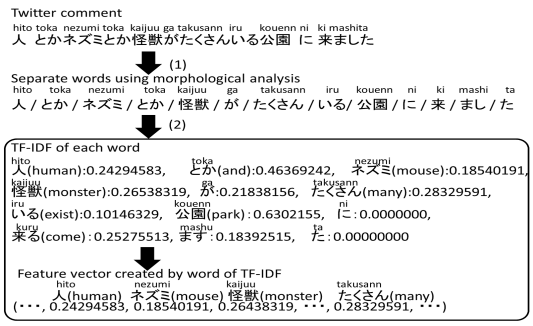

Fig. 4. Process of TF-IDF vector (if the facility searched was Tokyo Disneyland)

TABLE I. SCORE OF YAHOO! KEY PRASE EXTRACTION

word	score
ピグレット (Piglet)	41
スプラッシュマウンテン (Splash Mountain)	30
フローズンファンタジー (Frozen Fantasy)	39
ティーパーティ(Tea Patty)	26
ネズミ (Rat)	25
アリス (Alice)	22
怪獣 (Monster)	20

extracted feature words with the facility names. For example, the Tweet that can be obtained with “Tokyo Disneyland”, the feature word in noted in Table I.

Figure 5 shows created new search word and Figure 6 shows video search results.

B. Calculating the Characteristic Index of Words

In calculating the characteristic index of words, use those arranged by TF-IDF. In the present system, retrieve from Wikipedia[17] the data description section, and set the number of Wikipedia sentences as N, the number of words in the entire Wikipedia sentences as S, the number of occurrences of the word t of in the entire Wikipedia sentences as $n(t)$, and the number of sentences the word t appears as $df(t)$ of the Wikipedia sentences word t appears, and define as follows.

$$\text{TF-IDF(t)} = \frac{n(t)}{S} \cdot (log\frac{N}{df(t)} + 1)$$

Example of score of TF-IDF shows in Table II.

C. Learning the Classification of Twitter Posts

In section III.*A* (2), the Twitter posts are classified by SVM, and the SVM learning will be described. In this system, for the words that appear in every sentence of Wikipedia and Aozora Bunko[18], the TF-IDF(t) representing the characteristic index of words was determined, and the feature vector was created. This is the same method used to create the feature vector in section III.*A* (2). There process of feature vecitor for SVM learning shows in Figure 7. Then the feature vectors created from Wikipedia sentences were identified as positive cases, and feature vectors created from Aozora Bunko were identified as negative cases, and learning was carried out based on SVM teacher data. This was because Wikipedia text data included

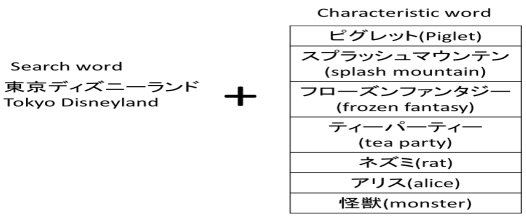

Fig. 5. Created new search word (if the facility searched was Tokyo Disneyland)

Fig. 6. Vido search using twitter posts(if the facility searched was Tokyo Disneyland)

TABLE II. EXAMPLE OF TF-IDF

東京 (tokyo):0.15329, ディズニーランド (disneyland) : 0.50876,
千葉 (chiba): 0.27813, 県 (prefecture):0.25516, パーク (park):0.15009,
シー (sea): 0.23617, など (nado) : 0.26544, 共 (tomo) : 0.32301,
リゾート (resort) : 0.15009, を (wo) : 0.11005, 形成 (create) : 0.17191,
する (suru) : 0.0050, 主役 (main charactor) : 0.35309,
ネズミ (mouse) : 0.28794, モチーフ : 0.45065,
ミッキーマウス (mickey mouse) : 0.48290, 岡山 (okayama): 0.17647,
城 (castle): 0.15307, 日本 (Japan) : 0.07578,
国 (nation) : 0.24210599, 指定 (assign) : 0.2493475,
史跡 (historic sites) : 0.10008282

many characteristic words related to facility information, while on the other hand Aozora Bunko text data included many slang and colloquial expressions irrelevant to facility information.

The classification accuracy of Wikipedia and Aozora Bunko text was 97.8%. An example of classified Twitter post about the Tokyo Disneyland in SVM is shown in Table III. In addition, for the evaluation of the classification accuracy, the official Tweets of the tourism association were considered as positives cases and Tweets from celebrities as negative cases, and 1,500 cases respectively were classified as positive SVM. Accuracy in this case was 60.0%.

IV. EVALUATION OF SEARCH RESULTS

A. Evaluation of Video Search

Using regional words in section II facilitated the determination of the region from the title and the details of the videos. In comparison to the conventional video search for YouTube, it is considered possible to perform a search for videos suitable for or close to the region that corresponds to the facility name being searched.

TABLE III. EXAMPLE OF FEATURE WORDS THAT CAN BE OBTAINED FROM THE TWITTER POSTS (THE CASE OF "TOKYO DISNEYLAND")

	Twitter post
necessary posts	ピグレットとイーヨー。@東京ディズニーランド (Piglet and Eeyore. @ Tokyo Disneyland) 人とかネズミとか怪獣がたくさんいる公園に来ました。 (I came to the park to have a lot of people and rats and monster) 動画を投稿しました@東京ディズニーランド【スプラッシュマウンテン】 (I posted the video) 高校生バミ in Disney たのしい @東京ディズニーランド【アリスのティーパーティー】 (High school students Bami in Disney Happy @ Tokyo Disneyland [Alice's Tea Party]) フローズンファンタジー (@東京ディズニーランド (TokyoDisneyland) (Frozen Fantasy (@ Tokyo Disneyland (TokyoDisneyland)) . . .
unnecessary posts	今日はこっち。(Today here) また来たよヽ(・∀・)ノ (I've come again.) いつも楽しそうなスティーブンさん @東京ディズニーランド (Always fun likely Stephen's @TokyoDisneyland) 楽しすぎた～#Disneyland#love@東京ディズニーランド (I'm having so much fun #Disneyland # love @ Tokyo Disneyland) いい天気激混み#TDL#祝日@東京ディズニーランド (TokyoDisneyland) (Nice weather, super crowded #TDL # holiday @ Tokyo Disneyland) . . .

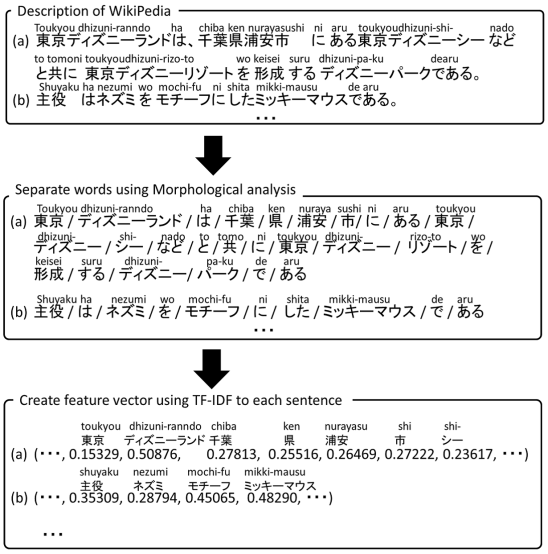

Fig. 7. Example of feature vector(if the facility searched was Tokyo Disneyland)

The problem is that if the morphological analysis of the facility name of in the process of creating the regional word has failed, a search is performed for videos that differs in address and region of the video. In order to improve search accuracy, not only is the analysis of video comments and the comparison of the word required, but also comparison of the region code is considered necessary. Also, there may be cases where inter-conversion of the address and facility name cannot be performed, and there were instances where similar videos could not be obtained. It is believed that an error in the location information held among the WebAPI was the cause.

With regards to the contents of the video, in case a search is performed for a tourist site, many landscapes are presented. In case buildings are searched, there are many videos of events, and videos principally on people rather than scenery and buildings were often displayed.

With regards to the increase of variations in the search results using the Twitter posts of section III, for example, in the case of searching for "Tokyo Disneyland", based on the SVM classification, words that include unique expressions such as "Splash Mountain" and "Frozen Fantasy" were classified as being useful Tweets. Particularly, "Frozen Fantasy" indicated an event that was held that from January 13, 2015 at "Tokyo Disneyland", "Anna and Elsa ' s Frozen Fantasy", and as of March 2015 although it is an event that started recently, Twitter posts that include trend information are useful in the creation of the search words. It is considered that many words related to Tokyo Disneyland, such as Tokyo Disneyland attractions of Splash Mountain and Frozen Fantasy, were extracted from the Wikipedia text. In fact, pages on Splash Mountain and Frozen Fantasy exist on Wikipedia, with detailed descriptions indicated for each.

B. Evaluation of Questionnaire

A questionnaire was carried out in regards to the video search results based on using regional words and Twitter posts. The questionnaire was conducted for five items,
Q1:Able to search for intended video,
Q2:Is there a wide variety of videos (completeness),
Q3:Were the key words for the search accurate? (accuracy),
Q4:Were you interested in the search results? (interest),
Q5:Did you find it to be enjoyable? (enjoyableness),
using a with a five evaluation scales of 1. Very bad. 2. Bad, 3. Normal, 4. Good. 5. Very Good.

The evaluation of system by questionnaire is shown in Table IV. With regards to completeness, using Twitter posts is better, and with regards to precision, using regional words leads to better results. In addition, as for interest and enjoyableness, using Twitter posts lead to better evaluations. It is considered that achieved the objective for increased variation in the results has been achieved.

V. CONCLUSION AND FUTURE WORKS

Our regional information video search system was able to obtain similar videos related to tourism information by entering either an address or an establishment name. Based on this, it would be possible for us to develop a system that takes either an address or establishment name to search for

TABLE IV. EVALUATION OF REGIONAL VIDEO SEARCH SYSTEM

	System using regional words (Section II)			System using Twitter posts (Section III)		
	Minimum	Maximum	Average	minimum	Maximum	Average
Q1	3.0	5.0	3.9	2.0	5.0	3.6
Q2	2.0	3.0	2.4	3.0	5.0	4.0
Q3	2.0	5.0	3.6	2.0	4.0	2.9
Q4	3.0	4.0	3.4	3.0	5.0	4.0
Q5	2.0	4.0	3.1	4.0	5.0	4.1

videos on relevant establishments and nearby establishments to provide tourist information in the form of videos. However, the classification accuracy of Twitter posts based on SVM learned from Wikipedia and Aozora Bunko is approximately 60 percent, which is not necessarily high. This may be due to the fact that heavy weighting was placed on words included in a Tweet appearing in the incorrect classification. As there are several effects on the search accuracy and variation, building a method to improve accuracy is a challenge.

REFERENCES

[1] Youtube, https://www.youtube.com getting_started/

[2] Twitter, http://https://twitter.com/

[3] T. Misu, E. Mizukami, C. Hori, H. Kashioka, ''Construction and Language Portability of Sightseeing Guidance System with Spoken Dialog Interfaces : Knowledge and Problems Obtained through the Development and Operation of Spoken Dialog System AssisTra", Japanese Society for Artifical Intelligence, Vol. 28 No.1, pp68-74, 2013.

[4] K. Ohtake, T. Misu, C. Hori, H. Kashioka, S.Nakamura, "Dialogue acts annotation for NICT Kyoto tour dialogue corpus to construct statistical dialogue systems", Proc. of the seventh international conference on Language Resources and Evaluation (LREC), 2010

[5] H. Nanba, H. Taguma, T. Ozaki, D. Kobayashi, A. Ishino, T. Takezawa, "Automatic Compilation of Travel Information from Automatically Identified Travel Blogs", Proc. Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing, pp. 205-208, 2009.

[6] A.Ishino, H. Nanba, T. Takezawa "Automatic Compilation of Travel Information from Automatically Identified Travel Blog Entries", journal of Japan Society for Fuzzy Theory and Intelligent Informatics, Vol.22, No. 6 pp.667-679, 2010.

[7] M. Okamoto, M. Kikuchi, "Local-area Event Extraction from Blog Entries", IPSJ, 51(1), pp14-17, 2010.

[8] Yahoo! Open Local Platform(YOLP), http://developer.yahoo.co.jp/webapi/map/

[9] Google Developers Youtube Data API, https://developers.google.com/youtube/ getting_started/

[10] Google Developers Google Geocoding API, https://developers.google.com/maps/docume ntation/geocoding/

[11] T.Kudo,K.Yamamoto,Y.Matsumoto, "Applying Conditional Random Fields to Japanese Morphological Analysis," Proeeding of the 2004 Conference on Empirical Methods in Natural Language Processing(EMNLP-2004), pp.230-237, 2004.

[12] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.googlecode.com/svn/trunk/me cab/doc/index.html

[13] V.N.Vapnik, V.N., Statistical Learning Theory, Wiley: New York, 1998.

[14] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, NewYork, 2006.

[15] UCI Machine Learning Repository, archive.ics.uci.edu/ml/

[16] K. Spärck, Jones, "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". Journal of Documentation 28, pp.11—21, 1972.

[17] Wikipedia, http://en.wikipedia.org/wiki/Wiki pedia/

[18] Aozora Bunko, http://www.aozora.gr.jp/