SUPPORTING INBOUND TOURISM IN HOKKAIDO: KEYWORD EXTRACTION AND FOCUS POINT ANALYSIS FROM SPOT REVIEWS

Zhenzhen LIU[1], Fumito MASUI[2] and Michal PTASZYNSKI[3]

[1]Doctoral Course in Manufacturing Engineering, Kitami Inst. of Tech.
E-mail: ryushinshintust@gmail.com
[2]Professor, Dept. of Computer Science, Kitami Inst. of Tech.
E-mail: f-masui@mail.kitami-it.ac.jp
[3]Associate Professor, Dept. of Computer Science, Kitami Inst. of Tech.
E-mail: michal@mail.kitami-it.ac.jp

(Koencho 165, Kitami, Hokkaido 090-8507, Japan)

Before the coronavirus pandemic, the number of inbound tourists in Japan had been on the rise during the whole previous decade. However, due to the influence of the coronavirus, the number of inbound tourists has now rapidly decreased. In order to support the post-pandemic tourism industry in Hokkaido, which is one of the popular travel destinations in Japan, it is important to clarify the needs of inbound tourists through the Internet and deliver valuable tourism information to support the flow of the inbound tourists.

This research concentrates on the information provided by the Chinese tourists, which are the largest group by the number of inbound tourists in Japan. We collected reviews of popular tourist spots in Hokkaido from Chinese travel industry website and extracted keywords from the reviews in order to find out the potential points of interests of Chinese tourists. The keyword extraction methods we used are TF-IDF(term frequency-inverse document frequency) and TextRank, and TF-IDF showed better results in this study. From the extracted keywords by TF-IDF we can clearly see what kind of elements Chinese tourists pay the most attention to when it comes to tourism spots. And this will be used for supporting the Japanese tourism recovery after the coronavirus pandemic.

Key Words : *Hokkaido, Tourism, Keyword Extraction*

1. INTRODUCTION

According to the Japan National Tourism Organization, the growth of inbound tourism to Japan has gradually increased during the whole previous decade. However, the coronavirus pandemic completely changed this situation in a few months. Since April 2020, inbound tourism has plummeted to -100% growth rate comparing to 2019[1)]. The coronavirus pandemic affected the traveling sector even more deeply than the 2011 earthquake and tsunami disaster in northern Japan.[2)]

In order to support the post-pandemic Japanese tourism industry, it is important to explore and further clarify the needs of inbound tourists and provide effective information for the tourism industry in Hokkaido, which is one of the popular travel destinations in Japan and famous for its beautiful nature and local culture. We want to extract the focus points of inbound tourists from online reviews by using keyword extraction methods and analyze the emotion of focus points to figure out what attracts inbound tourists the most and what the tourists are not interested in. Then use the obtained information to aid the Japanese tourism industry's recovery. Our goal is to build a system to automatically extract the focus points of inbound tourists from online reviews, analyze the potential interests of inbound tourists and then provide useful information for supporting the Japanese tourism industry.

This research mainly focuses on the online reviews provided by Chinese tourists, which are the largest group by the number of inbound tourists in Japan. For the first step, We collected reviews of popular tourist spots in Hokkaido from Chinese travel industry website, and extracted key-

words from the reviews. The keyword extraction methods we used in this study are TF-IDF and TextRank. TF-IDF is a numerical weighing factor based on word counts while TextRank is the application of PageRank algorithm to the field of natural language processing. From the extracted keywords we can find out the potential interests of Chinese tourists. These keywords will be used to determine specific focus points for attracting more inbound tourists in order to support the Japanese tourism industry.

The overall structure of this paper is as follows. Section 1 introduces the main idea of this study. Section 2 describes related works. In section 3, we describe the details of the experiments and analysis. In Section 4, we evaluate the keywords manually by checking how well they describe the features of each spot. We also compared the effectiveness of the different keyword extraction methods. Finally, we draw conclusions about this study in Section 5.

2. RELATED WORK

There are many studies on supporting Japanese tourism. For example, Shibata et al.[3] used the deep learning method LSTM to analyse the information of the users' visited countries, collected from social media. They attempted to predict more accurately which country the users will most likely visit in the future. Also, there are studies about the analysis of online reviews. Takamatsu et al.[4] extracted keywords and patterns from online hotel reviews and calculated the emotion polarity scores to recommend the most suitable hotels for the users.

Currently, owing to the influence of COVID-19, tourism activities are decreasing rapidly. Considering that the potential foreign visitors to Japan are feeling stressed due to staying home for such a long time, it is expected that the number of overseas travelers will dramatically increase once the pandemic is over [5]. Therefore, the analysis of the habits of inbound tourists is considered to be very important. For example, Ohkubo et al.[6] investigated the images of foreign tourists visiting Japan by examining both English travel guidebook and travel site. They also discussed the differences in focus points of different nations of tourists. Claire et al.[7] examined the correlation between rating score of review and emotion analysis of the review content using Spearman and Kendall Correlation coefficients and Maximal Information Coefficient (MIC). Sugiyama et al.[8] using NLPIR analysed the review data in various languages of foreign visitors in the city of Hamamatsu, Japan. They clarified the characteristics and differences of foreign travelers visiting local cities by nationality. Suzuki et al.[9] made a questionnaire to Taiwanese tourists about Japan's travel spots. They discussed the relations between push factor, pull factor and satisfaction of travel.

3. METHODOLOGY

The main process proposed in this paper includes the following steps, as shown in Figure1:

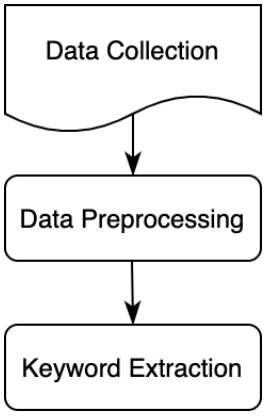

Fig.1 The Main Procedure of Analysis

Data Collection and Preprocessing: section 3(1) introduces the details of how we collected the reviews. We also describe the preprocessing steps necessary to clean the data.

Keyword Extraction: section 3(2) introduces how we extracted the keywords from reviews using TF-IDF and TextRank methods.

(1) Data Collection and Preprocessing

We selected spots that are popular and had a high number of reviews in Hokkaido from Ctrip (https://www.ctrip.com/), which is a well-known Chinese travel industry website. We scraped the website and collected a total of 10157 reviews. For each review, we extracted the useful information like content, score and date. Afterwards, we scanned the reviews for duplicates and removed them, this resulted in 10077 reviews remaining for our analysis.

In order to segment the words, we used a popular Chinese segmentation tool called Jieba[10]. As Jieba's dictionary did not include the words designed for Hokkaido spots, we updated the list manually to better suit our needs. The added words included '洞爺湖 (Lake Toya)', '五稜郭 (Goryokaku)', '富良野 (Furano)', etc. We also removed stop words from the segmented texts to reduce the amount of redundant words.

(2) Keyword Extraction and Analysis

In order to find the focus points of Chinese tourists, we extracted keywords from the collected reviews using TF-IDF and TextRank.

a) TF-IDF

TF-IDF or term frequency with inverse document frequency $tf * idf$ is a numerical weighing factor that can be used to extract keywords from texts. In TF-IDF, term fre-

quency $tf(t,d)$ refers to the number of times a term t (word, token) appears in a document d, while inverse document frequency $idf(t,D)$ is the logarithm of the total number of documents $|D|$ in the corpus divided by the number of documents containing the term n_t. Lastly, $tf * idf$ refers to the multiplication of these two as shown in equation (1).

$$idf(t,D) = log(\frac{|D|}{n_t}) \quad (1)$$

Compared to raw frequencies, TF-IDF helps adjust for the fact that some words appear more often in general as it is offset by the number of documents in the corpus that contain the word, while still increasing proportionally to the number of times a word appears in the document.

b) TextRank

TextRank[11] algorithm is derived from the classical PageRank algorithm[12]. PageRank is a famous algorithm by Google, which used to measure the importance value of particular web pages. The algorithm works by checking if there is a large number of web pages linking to a certain site, or if some important pages themselves link to the certain sites. These pages will result in a high value. In other words, the score of a page comes from the importance scores of all the pages that are linked to it through iteration calculation.

TextRank is the application of PageRank to the field of natural language processing and is widely used in keywords extraction. It is a graph structured in a similar way to PageRank. In the graph every sentence is a node that is linked to other nodes weighted by a similarity score.

Keywords can be extracted using TextRank in the following way. First split the document into component units like words or phrases and embed them into vector space. Then compute similarity scores between all pairs of nodes. Then run the PageRank algorithm to build the graph. Finally we can get top-n results from the generated graph[13]. Compared with TF-IDF which only considers word frequency itself, TextRank also considers the semantic relations between words in the document.

c) Experiments and Analysis

We used TF-IDF and Texrank methods to extract keywords from reviews. And for TF-IDF method, we used Jieba and Scikit-learn[14].

We first used Jieba's TF-IDF implementation with its own dictionary to get top 50 keywords from all the collected Hokkaido reviews, which are from 18 most popular spots. The results are sorted by the TF-IDF weight and shown in Table 1. In order to delete the keywords that are not specific to a certain spot, we then extract top 50 keywords from reviews of each spot and count how many times the keywords of Hokkaido (Table 1) appeared in all of the spots. This is also shown in Table 1, sorted by the frequency from least to most. For example, the frequency of the keyword '地獄谷 (Hell Valley)' is 1, which means this keyword only appeared in one spot. On the other hand, the frequency of the keyword '日本 (Japan)' is 18, which means it appeared in all of the spots.

Table 1: Top 50 keywords sorted by TF-IDF (left) and Bottom 50 keywords sorted by frequency (right)

Rank	Keyword		Freq	TF-IDF	Keyword		Freq
1	北海道	Hokkaido	18	0.1444	地獄谷	Hell Valley	1
2	札幌	Sapporo	16	0.1083	巧克力	chocolate	1
3	公園	park	8	0.0693	洞爺湖	Lake Toya	1
4	音楽盒	music box	3	0.0682	企鵝	penguin	1
5	函館	Hakodate	3	0.0672	神宮	shrine	1
6	夜景	night view	6	0.0576	大学	university	1
7	白色恋人	Shiroi Koibito	2	0.0552	工厂	factory	1
8	運河	canal	3	0.0547	餅乾	biscuits	1
9	地獄谷	Hell Valley	1	0.0406	白色恋人	Shiroi Koibito	2
10	日本	Japan	18	0.0405	動物園	zoo	2
11	温泉	spa	4	0.0398	函館山	Mt. Hakodate	2
12	動物園	zoo	2	0.0381	電視塔	TV Tower	2
13	景点	attractions	17	0.0370	美瑛	Biei	2
14	巧克力	chocolate	1	0.0334	富良野	Furano	2
15	洞爺湖	Lake Toya	1	0.0324	纜車	cable car	2
16	登別	Noboribetsu	3	0.0300	朝市	morning market	2
17	企鵝	penguin	1	0.0292	音楽盒	music box	3
18	地方	local	17	0.0290	函館	Hakodate	3
19	大通	Odori	3	0.0282	運河	canal	3
20	JR	JR	12	0.0279	登別	Noboribetsu	3
21	函館山	Mt. Hakodate	2	0.0276	大通	Odori	3
22	狸小路	Tanukikoji	3	0.0270	狸小路	Tanukikoji	3
23	電視塔	TV Tower	2	0.0267	時間	time	3
24	不錯	not bad	17	0.0261	温泉	spa	4
25	遊客	tourist	14	0.0260	海鮮	seafood	4
26	海鮮	seafood	4	0.0252	旭川	Asahikawa	4
27	景色	view	13	0.0249	浪漫	romantic	5
28	旭川	Asahikawa	4	0.0243	好吃	delicious	5
29	冬天	winter	13	0.0239	夜景	night view	6
30	美瑛	Biei	2	0.0239	参観	visit	6
31	富良野	Furano	2	0.0234	札幌市	Sapporo City	7
32	浪漫	romantic	5	0.0215	酒店	hotel	7
33	札幌市	Sapporo City	7	0.0214	公園	park	8
34	酒店	hotel	7	0.0212	建築	building	8
35	値得	worth	12	0.0209	拍照	take pictures	8
36	神宮	shrine	1	0.0199	晚上	night	8
37	喜歓	like	9	0.0197	喜歓	like	9
38	建築	building	8	0.0197	**美麗**	**beautiful**	**10**
39	特別	especially	13	0.0195	**JR**	**JR**	**12**
40	参観	visit	6	0.0189	**値得**	**worth**	**12**
41	大学	university	1	0.0180	**景色**	**view**	**13**
42	工厂	factory	1	0.0171	**冬天**	**winter**	**13**
43	拍照	take pictures	8	0.0170	**特別**	**especially**	**13**
44	餅乾	biscuits	1	0.0166	**遊客**	**tourist**	**14**
45	美麗	beautiful	10	0.0166	**札幌**	**Sapporo**	**16**
46	好吃	delicious	5	0.0164	**景点**	**Attractions**	**17**
47	纜車	cable car	2	0.0162	**地方**	**local**	**17**
48	朝市	morning market	2	0.0162	**不錯**	**not bad**	**17**
49	時間	time	3	0.0161	**北海道**	**Hokkaido**	**18**
50	晚上	night	8	0.0159	**日本**	**Japan**	**18**

In Table 1, The keywords sorted by frequency clearly show more specific features of each spot on the top, such as '巧克力 (chocolate)', '企鵝 (penguin)' and '白色恋人 (Shiroi Koibito)'. The more general words like '美麗 (beautiful)', '特別 (special)' and '日本 (Japan)', mainly go to the bottom. We highlight the words in the bottom, which appeared in over half of the spots, and removed them from the keyword list.

Some spots like '大通公園 (Odori Park)', '函館 (Hakodate)' and '美瑛 (Biei)' show lots of the distinct features. For example in Table 2, spot '大通公園' shows many keywords, such as '電視塔 (TV tower)', '冰雪節 (ice festival)', '噴泉 (fountain)', which described the features well. While spot '小樽音楽盒堂 (Otaru Music Box Museum)', '登別地獄谷 (Noboribetsu Hell Valley)' and '北海道旧道庁 (Former Hokkaido Govt. Office)' show more common words such as '喜歓 (like)', '味道 (smell)', '大楼 (buildings)', etc.

Similarly, we used Scikit-learn's TF-IDF function and Jieba's TextRank function to extract keywords from the collected reviews. In Scikit-learn's case however, we

Table 2: Examples of the top 10 keywords of some individual spots

Rank	大通公園	Odori Park	函館	Hakodate	美瑛	Biei
1	公園	park	函館	Hakodate	美瑛	Biei
2	大通	Odori	夜景	night view	富良野	Furano
3	電視塔	TV Tower	函館山	Mt. Hakodate	之樹	tree
4	札幌市	Sapporo City	海鮮	seafood	四季	four seasons
5	冰雪節	ice festival	朝市	morning market	自行車	bicycle
6	噴泉	fountain	三大	three biggest	接布	patchwork
7	街心公園	city center park	温泉	spa	花田	flower field
8	冰雕	ice sculpture	倉庫	warehouse	衣草	lavender
9	雪雕	snow sculpture	海胆	sea urchin	丘陵	hills
10	雪祭	snow festival	金森	Kanamori	騎行	cycling

Rank	小樽音楽盒堂 Otaru Music Box Museum		登別地獄谷 Noboribetsu Hell Valley		北海道旧道庁 Former Hokkaido Government Office	
1	音楽盒	music box	地獄谷	Hell Valley	紅磚	red brick
2	音楽	music	登別	Noboribetsu	旧道	old road
3	各式各樣	various	温泉	spa	建築	building
4	博物館	museum	硫黄	sulfur	巴洛克	Baroque
5	精緻	exquisite	硫黄味	sulfur smell	免費参観	free visit
6	喜歓	like	火山	volcano	歴史	history
7	蒸汽	steam	地獄	hell	札幌市	Sapporo City
8	琳瑯滿目	dazzling	味道	smell	大樓	building
9	建築	building	酒店	hotel	風格	style
10	童話	fairy tale	噴出	erupt	参観	visit

created the IDF dictionary directly from our review dataset.

4. KEYWORD EVALUATION AND COMPARISON

In order to compare those different methods of extracting keywords, we evaluate the top 10 keywords of each spot manually as shown in Table 3:

Table 3: Evaluation results of the top 10 keywords of the "旭山動物園 (Asahiyama Zoo)" spot

Rank	Keyword		Evaluation
1	動物園	zoo	good
2	企鵝	penguin	good
3	旭川	Asahikawa	good
4	動物	animal	acceptable
5	旭山	Asahiyama	good
6	北極熊	polar bear	good
7	散步	take a walk	good
8	可愛	lovely	good
9	海豹	seal	good
10	近距離	close range	acceptable
Evaluation Score			80%

The keywords were labeled by a Chinese native speaker using the following criteria.
(1)If the word shows the distinct features of the spot, or it can otherwise describe the spot very well, it is labeled as ‘good’.
(2)If the word does not show any features of the spot, or has no relation to the spot, it is labeled as ‘bad’.
(3)If there is a possibility that the word could be a feature of the spot, or the annotators are not sure, it is labeled as ‘acceptable’.

And we use the percentage of ‘good’ as the evaluation score, which represents how well they describe the features of each spot. The evaluation scores of TF-IDF (Jieba, Scikit-learn) and TextRank are compared in Table 4.

Table 4: Comparison of evaluation results

Spot name		TF-IDF (Jieba)	TF-IDF (sklearn)	TextRank
旭山動物園	Asahiyama Zoo	80%	80%	60%
札幌電視塔	Sapporo TV Tower	80%	60%	70%
小樽	Otaru	90%	80%	70%
小樽音楽盒堂	Otaru Music Box Museum	70%	70%	50%
小樽運河	Otaru Canal	90%	80%	70%
大通公園	Odori Park	100%	90%	70%
狸小路商店街	Tanukikoji Shopping Street	80%	80%	60%
登別地獄谷	Noboribetsu Hell Valley	70%	70%	70%
洞爺湖	Lake Toya	90%	80%	60%
白色恋人公園	Shiroi Koibito Park	90%	90%	80%
函館	Hakodate	100%	90%	80%
函館山	Mt. Hakodate	80%	80%	50%
函館朝市	Hakodate’s morning market	80%	70%	80%
美瑛	Biei	100%	100%	60%
富良野	Furano	90%	80%	60%
北海道旧道庁	Former Hokkaido Govt. Office	70%	70%	50%
北海道神宮	Hokkaido Shrine	90%	80%	70%
北海道大学	Hokkaido University	80%	80%	60%
Average		85%	79%	65%

From Table 4, we can find out that TF-IDF (Jieba) shows the best result with an average score of 85%, while TextRank shows the worst result with a score of 65%. The differences in the results of Jieba and Scikit are most likely due to the fact that Jieba uses a prebuilt IDF dictionary trained on a huge corpus, while Scikit’s dictionary was directly trained on the reviews themselves. With a larger review dataset, the evaluation ranking of the two TF-IDF methods could be the opposite. In the future we are planning to utilize a larger review dataset.

Table 5: Top 10 keywords of the "旭山動物園 (Asahiyama Zoo)" spot

Rank	TF-IDF (Jieba)		TF-IDF (sklearn)		TextRank	
1	動物園	zoo	動物園	zoo	動物園	zoo
2	企鵝	penguin	企鵝	penguin	企鵝	penguin
3	旭川	Asahikawa	動物	animal	動物	animal
4	動物	animal	旭川	Asahikawa	散步	take a walk
5	旭山	Asahiyama	旭山	Asahiyama	可愛	lovely
6	北極熊	polar bear	散步	take a walk	北極熊	polar bear
7	散步	take a walk	可愛	lovely	展示	display
8	可愛	lovely	北極熊	polar bear	設計	design
9	海豹	seal	海豹	seal	生活	life
10	近距離	close range	展示	display	日元	JPY

Table 5 shows the top 10 keywords of ‘旭山動物園 (Asahiyama Zoo)’ spot, which are extracted by TF-IDF(Jieba, Scikit-learn) and TextRank. We can see that TF-IDF(Jieba) and TF-IDF(Scikit-learn) show more words related to the features of the spot, like‘旭川 (Asahikawa)’, ‘海豹 (seal)’ and ‘近距離 (close range)’, which contains the information of location, popular animals and the design of zoo. While TextRank shows more general words, like ‘設計 (design)’, ‘生活 (life)’ and ‘日元 (JPY)’, which are not so strong features associated with the spot. Comparing to TextRank, TF-IDF performed better in our data set. It not only considered the frequency of words, but also the unique factor of the spot.

And from the extracted keywords, we can get an idea what Chinese tourists pay attention to. For example in Table 5, we can see that ‘企鵝 (penguin)’, ‘散步 (take

a walk)’, ‘動物 (animal)’, ‘可愛 (lovely)’, ‘北極熊 (polar bear)’ and ‘海豹 (seal)’ could be the things that attract Chinese tourists to Asahiyama zoo the most. In the next step, we will extract the sentences, which contain the keywords and use them to analyze the topics in order to extract the main focus points.

5. CONCLUSIONS

In this study, we illustrated the research plan that realizes a new method to extract focus points to attract inbound tourists. We want to build an automatic system to extract the focus points of inbound tourists from online reviews and analyze what attracts inbound tourists, then provide useful tourism information for supporting the post-pandemic tourism industry in Hokkaido.

For the first step, we collected Hokkaido’s tourism spot reviews and used TF-IDF and TextRank to extract keywords from the collected reviews. For TF-IDF, we used two different implementations, Jieba and Scikit-learn. To compare the extraction methods, we evaluated the top 10 keywords from each spot by checking how the keywords show the distinct features of each spot. The evaluation results indicate that TF-IDF(Jieba) shows the best result compared to the other two methods. Also, the keywords show the main topics Chinese tourists discuss most often in the reviews. We think these helps in finding the focus points of Chinese tourists.

In the future, we will increase the number of evaluators to get more objective evaluation results and evaluate the top 50 keywords from each spot. Then, we plan to extract n-gram patterns from the reviews which contains the keywords, aiming to clarify the focus points of Chinese tourists towards Hokkaido travel spots.

REFERENCES

1) Honichi gaikyaku su 2020-nen 12-gatsu suikei-chi oyobi nenkan suikei-chi [Number of foreign visitors to japan, estimated values for December 2020 and whole year 2020] (in Japanese). `https://www.jnto.go.jp/jpn/news/press_releases/pdf/210120_monthly.pdf`. Accessed: 2021-01-20.
2) The COVID-19 pandemic impact on Japanese inbound tourism. `https://www.wakayama-u.ac.jp/ctr/news/20201014000019/`. Accessed: 2021-03-28.
3) Shibata, Y., Ishino, A., Nanba, H., and Takezawa, T. Kanko no keitai o koryo shita shorai no homon-koku no yosoku [Predicting future visiting countries considering the form of tourism] (in Japanese). In *Society for Tourism Informatics 21st Research Presentation (2020)*, pages 37–40. Specified non-profit organization Society for Tourism Informatics, 2020.
4) Takamatsu, K. and Okuno, T. Kuchikomi kara chushutsu shita hyoka kanten no jushi-do ni motodzuku shukuhaku shisetsu suisen shuho no teian [Proposal of accommodation facility recommendation method based on importance of scoring category extracted from word of mouth] (in Japanese). In *Society for Tourism Informatics 21st Research Presentation (2020)*, pages 49–52. Specified non-profit organization Society for Tourism Informatics, 2020.
5) Sharma, G. D., Thomas, A., and Paul, J. Reviving tourism industry post-COVID-19: A resilience-based framework. *Tourism Management Perspectives*, 37:100786, 2021.
6) Okubo, T. and Muromachi, Y. Ryoko gaidobukku to kuchikomi no gengo kaiseki ni yoru honichi gaikoku hito no kanko-chi imeji ni kansuru kenkyu [Research on tourist destination images of foreigners visiting Japan by linguistic analysis of travel guidebooks and word-of-mouth] (in Japanese). *Journal of the City Planning Institute of Japan*, 49(3):573–578, 2014.
7) Claire, A. C. E., Nonaka, H., and Hiraoka, T. Honichi chugokujin kankokyaku no onrainhoterurebyu no kanjo bunseki to hyoka-ten no kankei-sei bunseki [Sentiment analysis of online hotel reviews of Chinese tourists visiting Japan and analysis of the relationship between score and sentiment] (in Japanese). *Journal of The Institute of Industrial Applications Engineers*, 6(2):95–99, 2018.
8) Sugiyama, Y., Zheng, J., Matsuo, T., Iwamoto, H., and Hochin, T. Internet review analysis of foreign visitors to regional cities in Japan. *Information Engineering Express*, 5(2):73–82, 2019.
9) Suzuki, T. and Zhou, W. Taiwanjin no Nihonryoko ni okeru ryoko doki to manzoku-do to no kankei ni tsuite [About the relationship between Taiwanese travel motivation and satisfaction in traveling to Japan] (in japanese). In *Society for Tourism Informatics 21st Research Presentation (2020)*, pages 33–36. Specified non-profit organization Society for Tourism Informatics, 2020.
10) "Jieba" Chinese text segmentation. `https://github.com/fxsjy/jieba`.
11) Mihalcea, R. and Tarau, P. TextRank: Bringing order into texts. In *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*, July 2004.
12) Langville, A. N. and Meyer, C. D. *Google’s PageRank and beyond: the science of search engine rankings*. Princeton: Princeton University Press, 2006.
13) Yujun Wen, Hui Yuan, and Pengzhou Zhang. Research on keyword extraction based on Word2Vec weighted TextRank. In *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, pages 2109–2113, 2016.
14) Pedregosa, F., Varoquaux, G., Gramfort, A., Michel,

V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.