Extracting Foreigner Interests for Japanese Culture from Interactive Digital Contents

Thi Ngoc Le and Hiromitsu Shimakawa

***Abstract*—In this paper, we propose a method to extract foreigner interests for Japanese culture from interactive digital contents, making use of quick exchange trait of instant messages (IM) on Social Network Service (SNS). It is difficult for foreigners who living in Japan to know their interest about Japanese culture because of language barrier and cultural differences. The method enables them to search how to enjoy Japanese culture based on their interests. From the experiment, we found that people from the same country tend to be more interested in the same topics. The result indicates we can provide excitements for foreigners from a specific country, preparing appropriate topics for individual countries in advance. It implies we can develop an automatic consultation system to introduce traditional Japanese culture to foreigners.**

***Index Terms*—Data mining, extract topic, machine learning, Naïve Bayes classifiers.**

I. Introduction

Japan has numbers of tangible and intangible cultural properties which attract attention from many foreigners. This country is holding various cultural activities and cultural experience programs [1]. By boosting awareness and interests of Japanese culture among foreigners living in Japan, both of the foreigners and Japanese would be benefit. First, foreigners' life will be enhanced from knowing there are various cultural properties around the place they are living in. Second, Japan's cultural maintenance, inheritance and development will be improved from insights of foreigners for Japanese cultural properties. However, language barrier and cultural differences have prevented foreigners from knowing their interests in Japanese culture [2]. It is difficult to suggest potentially interesting topics to a foreigner.

Foreigners usually have no knowledge on individual cultural properties, but they have unclear categories of cultural properties they want to experience. Their expectation for those categories is formed through reputations for the Japanese culture in their countries. Suppose we have topics on Japanese cultural properties to be presented to foreigners who have just come to Japan. We can assume foreigners of specific background (nationality, gender, etc.) will likely be interested in particular topics his peers from the same country are interested in. If we know topics popular to foreigners of a specific background, we can make the topics burst in conversation with them. It enables the Japanese culture to be spread and received more efficiently. To provide appropriate topics for individual foreigners, we need a proper method to find suitable topics as clues [3] for each foreigner to be inclined to experience Japanese cultural properties. With the method, we will take a huge leap in helping them understand their interests for Japanese cultural properties.

In this paper, we propose a method to extract foreigner interests of the Japanese culture from interactive digital contents. The method makes use of quick exchange trait of instant messages on SNS. It extracts interested topics in the Japanese culture. The method enables foreigners to search how to enjoy the Japanese culture based on their interests. From the experiment, we found that people from the same country tend to be more interested in specific topics. It implies we can introduce traditional Japanese culture which excites foreigners from every country, preparing topics suitable for each country.

Manuscript received November 3, 2016; revised May 6, 2017.

T. N. Le is with Ritsumeikan University, Shiga, 5258577 Japan (e-mail: le@de.is.ritsumei.ac.jp).

H. Shimakawa is with the Data Engineering Lab, Ritsumeikan University, Shiga, 5258577 Japan (e-mail: simakawa@cs.ritsumei.ac.jp).

II. Extracting User Interest Using IM

A. Familiarity for IM

The instant message (IM) on social network service (SNS) sites is being widely used as a primary means of contact. It is free and supports group chat. Besides, it allows users to exchange of digital contents (text messages, pictures, videos, icons, and so on) quickly and conveniently. Nowadays, many people are so familiar to use IM on SNS. They can communicate their opinion easily, even among far places, if the time is convenient for both participants.

B. Naïve Bayes Classifiers

Naive Bayes classifiers belong to a family of simple probabilistic classifiers based on the Bayes' theorem with strong independence assumptions between the features [4] [5]. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class. Using Naive Bayes' Theorem, $P\left(doc\middle|cat\right)$ the probability that the user has interests in class *cat* when we get document doc, is computed as:

$$P(cat|doc)=\frac{P(doc|cat)P(cat)}{P(doc)}$$

where $P\left(doc\middle|cat\right)$ is the conditional probability of the document occurring when we know the user is interested in class cat. Since each document is composed of multiple words, $P\left(doc\middle|cat\right)$ can be decompose as:

doi: 10.18178/ijiee.2017.7.3.671

$$P(doc \mid cat) = P(word_1 \wedge word_2 \wedge ... \wedge work_k \mid cat)$$
$$= \prod_i P(word_i \mid cat)$$

where $word_1, word_2 ..., word_k$ are words appearing in document doc . $P(word_i \mid cat)$ is he conditional probability of term $word_i$ occurring in document doc , when the user interested in cat writes the document.

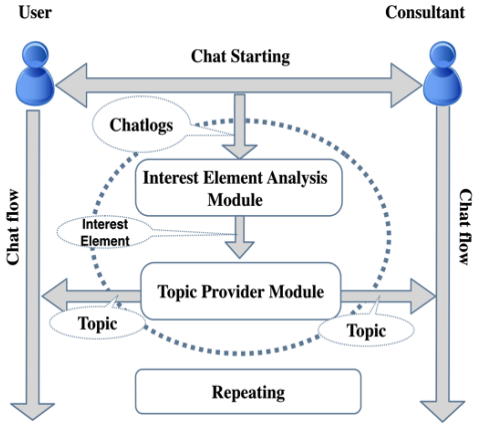

Fig. 1. User interest extracting.

III. Interest Element Analysis

A. Method Overview

We propose a method to analyze chat logs to extract user's interests and suggest a topic related to the extracted interests. Our target user is a foreigner who wants to know the Japanese traditional culture. He/she chats with a consultant using IM functions of SNS. The consultant introduces the variety of Japanese traditional culture to the user. We regard a chat log as a document representing the interest of the user. We define a chat segment is a four-round-trip of chat between the user and the consultant. To extract user's interests, we define an interest field and an interest element. An interest field is a genre in Japanese culture, for instance, nature and history of Japan, or the Japanese traditional paintings and pottery. It is associated with related keywords. Each interest field is assumed to be mutually exclusive. Hence, we regard that an interest field has a set of related keywords, as well as all of the sets are mutually exclusive. An interest element is the interested field with the highest possibility that the user is most interested in. The interest element is considered to appear in the most recent chat segment.

The method analyzes chronologically user's interest through the chat conversation with the Naive Bayes classifier**.** The consultant introduces topics expected to be interesting to the user. The topics are selected among ones representing the interested elements. Since our method estimates the interested element using a Naive Bayes classifier, the user is notified of a topic most likely to represent his/her interest.

Figure 1 shows the chat flow in our method. The Interest Element Analysis Module analyzes each four-round-trip (one chat segment), to derive the interest element. After that, the Topic Provider Module suggests a topic related to the interest element. Here, the user is a foreigner who is unclear with his/her interest for Japan culture. The consultant is a person who assists the user to discover his/her interest. The user chats with the consultant via IM on SNS.

At first, the consultant uses his/her own knowledge to provide topics to the user. For each chat segment, the method stores related keywords from each interest field with their appearance frequency. Any chat segment identified with the followings procedure will be stored in the burst database.

1) The consultant proposes an initial topic.
2) The user discusses with the consultant on the initial topic, therefore chat segment *S* is produced.
3) The method marks the chat segment, *S*, as an exciting chat segment if *S* makes user interested. Chat segment *S* is broken down into keywords *K*
4) The method derives a new interest element, *I*.
5) The method orders keywords *K* {k_i}, while k_i is the *i*-th keyword of interest element *I*. It stores these keywords into **burst database.**

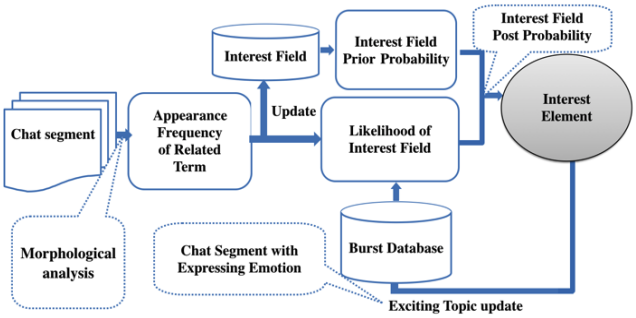

Fig. 2. Interest element analysis module overview.

B. Interest Element Analysis

When a chat segment is provided, the interest element analysis module analyzes interest fields to presume the interest element of the provided chat segment. According to its result, the topic provider module provides topics, which relate to presumed interest element, to both parties (the user and the consultant).

Fig. 2 shows the overview of the working flow in the interest element analysis module. In the first step, the module performs the morphological analysis on related keywords appearing in the chat segment. The module picks up only nouns. In the last step, interest element is determined through calculation of the posterior probability based on Bayes' theorem. We can calculate each related keyword frequency in each interest field over all the related keywords in the whole chat segment history. The problem is how we can identify the interest field with the highest probability of user being interested in it. In another word, we are calculating the posterior probability of a particular interest element.

The method proposes an initial topic every chat segment. The consultant starts interactions from the initial topic with the user. In the interactions, chat segment S is produced. If chat segment S makes the user interested, the method will mark S as an exciting chat segment. Chat segment S is broken down into keywords {K}. Keywords in {K} is stored into the burst database. Whenever the method derives a new interest element, the method calculates the probability of the interest

element which caused the user excited based on the historical data from the burst database. We refer to this probability as the likelihood. The value of the posterior probability is the product of the prior probability and the likelihood. The posterior probability indicates that the chat segment becomes interesting to the user because of a particular interest element. The interest element can be regarded as the cause of excitement.

TABLE I: INTEREST FIELD DATABASE STRUCTURE

Interest field H_i	Experience H_1	Japan castle H_2	Temple H_3
{keyword: frequency}	{karate-do: f_{11}, ikebana: f_{12}, …}	{Osaka castle: f_{21}, Hikone castle: f_{22}, …}	{Kiyomizudera: f_{31}, Kinkakuji: f_{32}, …}

C. Interest Field and Exciting Topics

Table I shows an example of the interest field database structure. The interest filed database stores featured keywords with frequency of the interest field.

Besides, the method also stores emotional words appearing in the entire chat segment history. Emotional words could be "wow", "definitely", "uhm", etc. We presume in the past others users have discovered their interests, using the method. If a chat segment contains any of these emotional words, the chat segment will be stored in the burst database.

Application of Bayesian Modelling

This section explains how the Bayesian modelling is applied in the method to derive interest elements or to calculate the posterior probability. Bayesian formula is computed as

$$P(cat \mid doc) = \frac{P(doc \mid cat)P(cat)}{P(doc)} \quad (1)$$

where doc is a chat segment, cat is an interest field. When the method obtains doc , the posterior probability of occurring cat , which is $P\left(word_i \middle| cat\right)$, will be calculated by formula (1). Since each interest element is determined, comparing the posterior probabilities of interest fields,

$$P(cat|doc) \propto P(doc|cat)P(cat) \quad (2)$$

where cat indicates the portion of each interest field in the entire chat segment history. In this study, every chat segment history is treated as training data. We define cat_i is the *i*-th interest field in the interest field database and f_i is the frequency of related keywords in cat_i . In our research, we applied the bag-of-words model. Under the circumstances, those keywords are treated independently. If $\{word_1, word_2, .., word_k\}$ is a set of keywords we collected after the morphological analysis on a chat segment, the order of these keywords will be disregarded. $P\left(doc \middle| cat\right)$ will be calculated by the following formula

$$\begin{aligned} P(doc \mid cat) &= P(word_1 \wedge word_2 \wedge ... \wedge work_k \mid cat) \\ &= \prod_i P(word_i \mid cat) \end{aligned} \quad (3)$$

here, $P\left(word_i \middle| cat\right)$ is the ratio of the frequency of $word_i$ to the total frequency of all keywords (V) in all chat segments in all training data. If $T\left(\text{cat} \middle| word_i\right)$ is the frequency of $word_i$ in interest field cat, then $P\left(word_i \middle| cat\right)$ would be calculated by the following formula

$$P(word_i \mid cat) = \frac{T(cat, word_i)}{\sum_{word_j} T(cat, word_j)} \quad (4)$$

In fact, the frequency of keywords not appearing in any interest field is *zero*. To avoid the division with the zero frequency, we will only deal with keyword appearing in interest fields. We define cat_m is the interest element we need to derive from input chat segment. cat_m is calculated by the following formula (5):

$$\begin{aligned} \text{cat}_m &= agr\max P\left(cat \middle| doc\right) \\ &= \underset{cat}{agr\max} P\left(cat\right) \prod_i P\left(word_i \middle| cat\right) \end{aligned} \quad (5)$$

In formula (5) $\prod_i P\left(word_i \middle| cat\right)$ will cause an underflow if there are too many keywords in the chat segment. To deal with this problem, we will get the logarithm (5) and derive in formula (6):

$$\begin{aligned} cat_m &= agr\max \log P\left(cat \middle| doc\right) \\ &= \underset{cat}{agr\max} \left(\log P\left(cat\right) + \sum_i \log P\left(word_i \middle| cat\right)\right) \end{aligned} \quad (6)$$

Application of the logarithm on (5) does not affect achieving the interest element which maximizes the posterior probability. The result from (6) is the interest element that we expect.

IV. EXPERIMENT AND EVALUATION

A. Experimental Overview

In order to evaluate the validity of the proposed method, we investigate the following three purposes. First, we inspect features of interest element in practical chat segments between foreigners and consultants. Second, we verify the effectiveness of introducing other factors such as images and web links to the chat session. Finally, we want to exploit the features of exciting chat segments.

There are 2 participant groups of this experiment: examinee group and consultant group. In the examinee group, we recruited 7 foreigners (3 Vietnamese, 3 Indians and 1 Indonesian; 3 females, 4 males) from 23- 28-year-olds. All are students of Ritsumeikan University, living in Shiga, Japan. Their average time of stay in Japan is 5.6 months. In

the consultant group, we recruited 2 Japanese, who are tour guide in contact with foreigners before. They supported the examinees independently in this experiment.

In this experiment, we created a group chat on LINE—the number one in SNS market share in Japan. The participants are required to participated in the group on Line to have conversations. There was one chat session between 1 examinee and 1 consultant. All participants communicated in English. Duration of each chat session was 15 minutes. To initial topic to start a conversation, examinee brought up a topic related to their living environment. Examinee would say "I have just arrived in Shiga recently and I don't have much knowledge of this place. Please guide me to know more about the culture of this place".

B. Evaluation Result

To evaluate collected data from this experiment, we used qualitative evaluation method. We recruited 3 Japanese evaluators to analyze the data. They read through the chat segments collected from experiment. Then, they extracted topics (could be multiple topics per chat segment) that they think they could be interest elements.

From evaluation result of topics from 3 evaluators, we found that even though 2 different consultants carry out the chat sessions with examinees, the results show that there is the majority in interested topics of examinees. 5 out of 7 topics fall into the major category. It implies that a certain topic already exists in the mind of each examinee.

TABLE II: LIKELIHOOD VALUE FOR EACH COUNTRY

Likelihood	Experience	Traditional exhibition	Festival	Food	Scenery	Watching
Indian	**0.0037619**	0.0024629	0.0006209	**0.0049875**	0.0013957	0.0006209
Vietnamese	0.0022074	0.0046915	0.0011352	0.0015247	**0.0159867**	0.0023458
Indonesian	**0.3018868**	0.0566038	0.0188679	0.2452830	**0.3584906**	0.0188679

Furthermore, the results show that examinees are interested in different topics. Some are interested in scenery, while others are interested in Japanese foods as well as their home country foods. Another topic is to experience Japanese traditional activities. There are some examinees interested in agriculture. We found that most Indians are interested in foods and experiences, while most Vietnamese are interested in scenery. Only the Indonesian is interested in experience and scenery.

From the evaluation of topics, we proceed to calculate the probability a user is interested in a particular interest field if we know where the user comes from. We could conclude 6 typical interest fields from the evaluation:

1) **Experience field** including activities which user could participate and experience. For instances: cycling, climbing, ropeways, and agriculture experiences.
2) **Traditional exhibition field** including traditional exhibitions of Japan such as temples, shrines, ninja villages, and movie villages.
3) **Festival field** indicating a day or time of religious or celebration, marked by feasting, ceremonies, or other observances, for example, the Gion-matsuri festival.
4) **Food field** including culinary cultures such as miso, ra-men noodles, and rice cakes.
5) **Scenery field** containing natural landscapes considered in terms of their appearance, especially picturesque places like the lake Biwa, fuji flower, old-towns.
6) **Watching field** showing events which users could watch and perceive, such as boat-races.

At first, we carry out morphological analysis, where we only analyze nouns. From the collected nouns, we take out insignificant words, such as "me", "this", "there". We put the collected nouns into 6 interest fields. In the end, we calculate the sum of the word frequency for each interest field. Table II summarizes the likelihood of 3 groups of examinees by country for 6 interest fields.

Table II has shown that the value of likelihood which varies for each country. Indians have the highest likelihood of experiences and foods. The participated Indians are interested in experience and food. Vietnamese have the highest likelihood of scenery. The Indonesian participant is interested in experiences and scenery.

From the above results, we could prepare specific topics for each group of people from the specific countries. In the past, we know that group of people from the same country tend to be interested in specific topics. There is a high possibility that people from the countries will be interested in the same topics. Because of that, chat sessions could be exciting from the beginning without proposing many undesired topics. Since there are differences among the groups of people, we apply the Bayesian model to provide topics for each chat session with high possibilities that the user finds the topics interesting. However, groups of Indians and Vietnamese are friends. There might be chances that they share the common view of Japanese cultures, which result in the common topics they are interested in.

V. CONCLUSION

In this paper, we have presented a method to extract foreigner interests for Japanese cultures from interactive digital contents using IM. An experiment has shown clearly that people from a specific country tend to be more interested in the same topics. The result has shown that it is possible to apply the Bayes model to derive foreigner's interests. The next steps are to complete the method and to collect more data, so that we could verify the method.

REFERENCES

[1] The Agency for Cultural Affairs Japan. (March 2009). Cultural communication to improve understanding and interest in Japanese culture. [Online]. Available: http://www.bunka.go.jp/seisaku/bunkashingikai/sokai/sokai_9/48/pdf/shiryo_10. pdf

[2] S. Ozdemir-Cagatay & A. Kullu-Sulu. (2013). An investigation of intercultural miscommunication experiences. *International Online Journal of Education and Teaching (IOJET).* [Online]. 1(1). pp. 39-52. Available: http://iojet.org/index.php/IOJET/article/view/43/63

[3] M. Zhang, R. K. E. Bellamy, and W. A. Kellogg, "Designing information for remediating cognitive biases in decision-making," *CHI 2015*, Crossings, Korea: Seoul, pp. 2211-2220.

[4] A. McCallum and K. Nigam, "A comparison of event models for Na ïve Bayes text classification," in *Learning for Text Categorization*, Paper from the AAAI Workshop, AAAI Press, 1998, pp. 41-48.

[5] Christopher D. Manning, "An Introduction to Information Retrieval, 1st ed.," Cambridge University Press, U. K.: Cambridge, 2009, ch. 13, pp. 234-265.

T. N. Le was born in Vinh Phuc, Vietnam, in 1989. She received B.E degree in information technology from Hanoi University of Science and Technology, Hanoi, Vietnam in 2012.

She advanced graduate School of Ritsumeikan University, Shiga, Japan.

Ms. Le engages in the research on data engineering.

H. Shimakawa was born in Osaka, Japan, in 1961. He graduated the department of information engineering, Kyoto University, Kyoto, Japan. He received Ph.D of engineering from Kyoto University in 1999.

He worked at Mitsubishi Electric Corporation, Japan. He has been a professor of Ritsumeikan University, Shiga, Japan since 2002. His main interests are data engineering, user interface, and educational engineering.

Prof. Shimakawa is a member of IEEE and ACM.