Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 35 (2014) 894 – 901

Procedia Computer Science

18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014

Evaluation of communication and travel behavior extraction with latent topics

Nobuo Suzuki[a*], Kazuhiko Tsuda[b]

[a]KDDI Corporation, Iidabashi 3-10-10, Chiyoda, Tokyo 102-8460, Japan
[b]University of Tsukuba, Otsuka 3-29-1, Bunkyo, Tokyo 112-0012, Japan

Abstract

This study proposed the habitual behavior information extraction method from the data on Internet to build effective behavioral change support system so far. It is well known that habitual behavior improvement is important to avoid risk behaviors for a safety driving and a health improvement. It used Latent Dirichlet Allocation approach and evaluated by using telecommunication behaviors in Question and answering Web sites. This paper describes another evaluation by using travel behavior information. On the other hand, the dependency relation is often used to extract valuable information from text data. It also shows the comparative evaluation between our proposed method and the dependency relation method. It is realized the proposed method is more accurate than the dependency relation method according to the result of the evaluation.

Peer-review under responsibility of KES International.

Keywords: Habitual behavior; Behavioral modification; LDA; PMI

1. Introduction

Large text data are generated daily at SNS and Question and answering sites on the Internet [14]. The method for extraction habitual behavior information from such data has been proposed to build a behavioral modification supporting system affected to safety driving and health improvement [9]. It is well known that the improvement of habitual behavior is significantly important for a habitual behavior subject to improve health by prohibiting smoking and avoiding a dangerous behavior such as using a mobile phone while driving. Therefore, this study has tried to extract the information focused on habitual behaviors.

* Corresponding author
E-mail address: suzu3nobu@gmail.com

Peer-review under responsibility of KES International.
doi:10.1016/j.procs.2014.08.163

Actual extraction method of habitual behaviors presumed latent topics included in text by Latent Dirichlet Allocation, LDA, which was one of topic models and decided the words suited for habitual behaviors by Point-wise Mutual Information, PMI, from candidate words included the topics. This method is called the proposed method in this paper. This paper reports the result of the evaluation experiment at travel behaviors by using transport facilities on Question and answering sites, in addition to the evaluation experiment at telecommunication behaviors so far. The dependency relation is frequently used as a method to extract the valuable information from text data. Therefore, our proposed method has been evaluated by comparing with the dependency relation method with data same as the evaluation experiment of the proposed method before. It was realized that the proposed method obtained higher accuracy rate than the dependency relation method by the result of the comparative evaluation.

2. Methods of extracting behavior information with latent topics and PMI

2.1. How the extraction method works

The proposed method extracts the habitual behaviors information with LDA and PMI as below. First, the habitual behavior was defined as frequently appearing human's common behavior that was not only physiological habitats such as tooth brushing and sleeping. The habitual behavior, therefore, includes three elements those are an action, an object and periodic frequency information. The habitual behavior HB is defined as combination of the formula (1).

$$HB = \{Frequency, Action, Object\} \tag{1}$$

The habitual behaviors were extracted by using LDA that was one of the topic model technologies [3]. The feature of the topic model is to express one document as a mixture of one or more topics. Canini et. al showed it could model documents with high accuracy. Our method prepared frequently keywords used as periodical expressions such as "Yoku", "Mai" and "Itsumo". A morpheme analysis is carried out by using the extracted sentences. It selects bag-of-words with adjectives, verbs, nouns and adverbs that are easily used as objects of the frequency, the action and the object. Then, LDA processing is executed. As a result, the topics constructed with one or more words are extracted and then the topics that have periodic expressions in those topics are also extracted. Some words are included at the time. They become the candidates of actions and objects expressed the habitual behaviors other than the periodic expressions in extracted each topics. The habitual behaviors, however, cannot be extracted more accurately because extracted topics include some words other than habitual behaviors in extracted topics even as they are. Therefore, the method assumes the words related to habitual behaviors with PMI as an index from the words in the topics. Then, the keywords are applied to "Frequency" of habitual behaviors. PMI between the words for "Action" and frequency keyword is calculated. They include Verb-independent, Noun-Sa-changing and Noun-adverb-available. Then two largest words are extracted as Action. It calculates PMI between the nouns and the periodical keywords, and then selects the best three words. The nouns includes Noun-Sa-changing that wasn't selected at Action and excludes Noun-independent. PMI is calculated with formula (2) and (3). It is an index that can express the strength among words and shows strength of relation between Action and Object words with the periodical words.

$$PMI(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{2}$$