Advice Extraction from Web for Providing Prior Information Concerning Outdoor Activities

Shunsuke Kozawa, Masayuki Okamoto, Shinichi Nagano, Kenta Cho and Shigeki Matsubara

Abstract Conventional context-aware recommendation systems do not provide information before user action, although they provide information considering user's ongoing activity. However, users want to know prior information such as how to go to their destination or get necessary items when they plan to do outdoor activities such as climbing and sightseeing. It takes time to collect the prior information since it is not so easy to appropriately find them. This paper proposes a method for extracting prior advices from the web. The method first identifies whether a given sentence is an advice or not. Then the method identifies whether the sentence is a prior advice or not if the sentence is identified as advice. In this paper, we will show availability of the proposed method through our experimentation. We also developed a system for providing prior information using the proposed method.

Shunsuke Kozawa
Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, 464-8601, Japan, e-mail: kozawa@el.itc.nagoya-u.ac.jp

Masayuki Okamoto
Corporate R&D Center, Toshiba Corporation, 1 Komukai Toshiba-cho, Saiwai-ku, Kawasaki-shi, 212-8582, Japan, e-mail: masayuki4.okamoto@tohsiba.co.jp

Shinichi Nagano
Corporate R&D Center, Toshiba Corporation, 1 Komukai Toshiba-cho, Saiwai-ku, Kawasaki-shi, 212-8582, Japan, e-mail: shinichi3.nagano@toshiba.co.jp

Kenta Cho
Corporate R&D Center, Toshiba Corporation, 1 Komukai Toshiba-cho, Saiwai-ku, Kawasaki-shi, 212-8582, Japan, e-mail: kenta.cho@toshiba.co.jp

Shigeki Matsubara
Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, 464-8601, Japan, e-mail: matubara@nagoya-u.jp

1 Introduction

With the development of mobile devices such as smart phones and tablet PCs, we often access the Internet and get knowledge and information in outside. Therefore, context-aware systems have been researched in recent years [1]. In particular, information recommendation systems considering users' contexts have been developed [2, 11, 8, 13]. However, they did not fully consider contents of information provided to users since their main research interests were to recognize user's contexts and they provided users with prepared information or information based only on user's position.

Users want to know prior information such as how to go to their destination or get necessary items, when they plan to do outdoor activities such as climbing and sightseeing. For instance, when they plan to climb Mt. Fuji, they want to know the means of transportation to Mt. Fuji and the needed climbing gear before leaving. In outdoor activities, user's contexts are divided into two main categories: before or during activities. The existing information recommendation systems considering user's contexts do not provide information before user action, although they provide information considering user's ongoing activity.

Prior information on outdoor activities are written in blogs by many people and stored on the web. We could provide users with valuable information if we use the web as a collective information. However, it takes time to collect the prior information since it is not so easy to appropriately find them using existing web search engines and most web pages containing the prior information include information useful during actual activities. In conventional researches for text mining, sentences containing user's experiences and troubles were extracted[4, 7, 9, 10]. However, the extracted information could not be identified as prior information since they did not consider situations for using the extracted information.

In this paper, we propose a method for extracting prior advices from the web to provide prior information before user action. We first identify whether a given sentence is an advice or not. Then, we identify whether the sentence is a prior advice or not if the sentence is identified as an advice. We also developed a system for providing prior information using the proposed method.

2 Characteristics analysis of advices

2.1 The definition of advices

Advices are sentences containing information which is worthy to provide users before or during their activities. In the following sections, we assume that the target language is Japanese. Table 1 shows advices about climbing Mt. Fuji. The third column in Table 1 represents whether the sentence is an advice or not. If the sentence is an advice, the value in the column is 1, otherwise, the value is 0. There exist ap-

Table 1 Examples of advices.

id	sentences	advice or not	prior or not
1	上着は防寒用にウインドブレーカがお 薦め。 (I recommend a windbreaker for protection against the cold.)	1	1
2	登山にあたっては、落石の恐れがありますので、十分ご注意 ください。 (Please note that there is fear of falling rocks when climbing.)	1	0
3	ほぼ毎年富士山を訪れているが、2009 年も 8 月 1 日に富士に登った。 (I visit Mt. Fuji nearly every year and climbed Fuji on August 1 in 2009.)	0	-
4	ゆっくり歩いて、長時間休まないのが、楽 に登るコツですね。 (The knacks of easily climbing Mt. Fuji are to walk slowly and not to take a rest for a long time.)	1	0
5	私がお 薦め するルートがあります。 (There is routes I recommend.)	0	-
6	ただし、お盆休み近辺は マイカー 規制で入れません。 (However, you can't go to Mt. Fuji by driving your own car due to the regulation on private cars during the Bon Festival.)	1	1
7	万一、天候の急変により危険を感じたら最寄の山小屋に避難しましょう。 (If by any chance you feel a sense of danger by the sudden change in the weather, you should take shelter to near mountain lodge.)	1	0
8	体調に余裕を持って高度に慣れながら 登って いきましょう。 (Let's climb Mt. Fuji while keeping on the safe side and becoming accustomed to altitude.)	1	0

propriate situations for using the advices. However, the situations are various and depend on activities. In this paper, we assume that the situations of the advices can be categorized into two types; before and during activities. This is because we consider situations which are common to variety of activities. We discriminate between them by judging whether a given advice is needed before taking action. The fourth column in Table 1 represents whether the advice is needed before activities or not. If the advice is needed preliminarily, the value in the column is 1. Otherwise, the value is 0. For instance, the first example is a prior advice and the second one is an action advice.

2.2 Construction of development data

We constructed development data to analyze characteristics of advices. Firstly, we searched the web by using "富士山 (Mt. Fuji) & 登山 (climbing)" and "穂高岳 (Mt. Hotaka) & 登山 (climbing)" as queries and collected top 50 web pages from search results for each query. We used Yahoo! Web API [1] to search the web. Next, we extracted texts from the web pages and split them into sentences. Finally, we manually judged whether each sentence is an advice or not and then manually judged whether the sentence is a prior advice or not if the sentence is an advice. Note that we judged them by referring to the previous and next sentences of the sentence. The size of the development data is shown in Table 2.

[1] http://developer.yahoo.co.jp/

Table 2 Size of the development data.

location	action	# of sentences	# of advices	# of prior advices
Mt. Fuji	climbing	2,581	899	638
Mt. Hotaka	climbing	4,144	360	132
total		6,725	1,259	770

Table 3 Clue expressions which represent characteristics of advices.

class	clue expression
recommend	薦め，方が良い (recommend, have better, rather ... than, etc.)
warning	気を付ける，注意 (care, note, attention, etc.)
prepare	用意，準備，持参 (prepare, ready, equip, etc.)
inhibition	禁止，禁物，厳禁，避ける (inhibit, prohibit, forbid, avoid, etc.)
necessary	必要，必須，必需 (necessary, essential, require, etc.)
schedule	計画，工程 (schedule, route, plan, etc.)
crowd	渋滞，混雑，満員 (traffic jam, crowded, full, etc.)
business	営業，開店，閉店，閉鎖 (open, closed, in business, etc)
occasion	場合，際，とき，状況 (occasion, when, in case, scene, etc.)
emergency	万が一，いざというとき (if by any chance, when the chips are down, etc.)

2.3 Characteristics of advices

We manually investigated the development data to capture characteristics of advices. We carefully analyzed the data to capture general characteristics which do not represent a particular domain although the development data represents climbing. Consequently, we found the following five characteristics:

A. function word at sentence end
We found that advices are often described in a polite way. While most of sentences described in past tense and interrogative sentences are not advices. In Japanese, tense, aspect and modality of a sentence are often represented by function words at the end of the sentence. Therefore, we capture these writing styles by focusing on function words at the end of a sentence as shown in the underlined portions in the examples 2 and 3 in Table 1.

B. evaluation expressions
We found that advices often contain expressions having a semantic polarity. The example 4 is an advice containing positive or negative evaluation expressions and the underlined portions represent evaluation expressions.

C. clue expressions
We manually generated 470 expressions which were often contained in advices and classified them into 35 classes based on their meaning. Some of the expressions are shown in Table 3. The example 1 is an advice containing the clue expressions in the underlined portions.

D. sentence end information
We found that appearance position of evaluation and clue expressions in a sentence is important. For example, a clue expression “薦め (recommend)” is appeared at the end of the sentence in the example 1 which is an advice. While the

Table 4 clue expressions which represent situations of advices.

class	clue expression
traffic	車，電車，駅，国道 (car, train, station, national road, etc.)
prepare	用意，準備，持参 (prepare, ready, equip, etc.)
preliminary	事前，予め，前もって，未然に (preliminary, on ahead, before happens, etc.)
weather	天気，気温，雨，暑い (weather, rain, temperature, hot, etc.)
impossible	不可，できない (impossible, can not, unable)
careful*	慎重，きちんと，しっかり (carefully, neatly, accurately, etc.)
knack*	コツ，ポイント，仕方 (knack, point, know-how, how to, etc.)
emergency*	万が一，いざというとき (if by any chance, when the chips are down, etc.)
possible*	可能，できる (can, possible, able, etc.)

example 5 is not an advice even though it contains the same clue expression. As seen in these examples, the advices often contain evaluation and clue expressions at the end of a sentence since a content word at the end of a sentence often plays important role in representing the sentence meaning in Japanese.

E. context information
It is often the case that the previous and next sentences of an advice are also advices since advices are often collectively described in a web page. Thus, the evaluation and clue expressions frequently appear in the previous and next sentences of an advice.

2.4 Characteristics of advices suitable for situations

We manually investigated the development data to capture characteristics of prior and action advices, and found the following four characteristics:

A. clue expressions
We manually generated 457 expressions which often appeared in prior advices and classified them into 17 classes based on their meaning. We also manually generated 109 expressions which were often contained in action advices and classified them into nine classes based on their meaning. Some of the expressions are shown in Table 4. The class names attached with * in the first column in Table 4 represent clue expressions for action advices. The examples 6 and 7 are advices containing the clue expressions in the underlined portions.

B. action verbs
The action advices often contain action verbs which represent bodily movement. The example 8 contains an action verb “登る (climb)”.

C. sentence end information
We found that prior advices often contain clue expressions for prior advices at the end of sentence and action advices often contain clue expressions for action advices and action verbs at the end of sentence.

D. context information
The advices are frequently described collectively. Therefore, the clue expressions for prior advices are often contained in the previous and next sentences of a

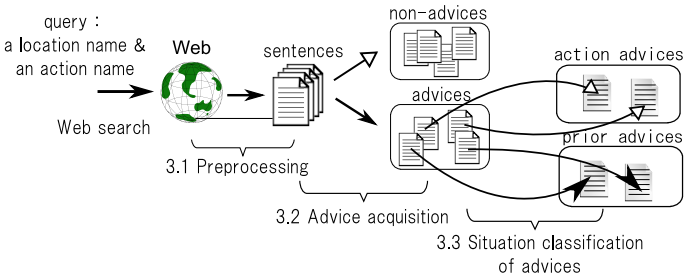

Fig. 1 Flow of prior advice acquisition

prior advice. The clue expressions for action advices and action verbs are often contained in the previous and next sentences of an action advice in the same way.

3 Prior advice acquisition

Figure 1 shows the flow for extracting prior advices. Firstly, we search the web by using a location name and an action name as a query and get HTML texts whose title includes the location name. Yahoo! Web API is used for searching the web. Secondly, we preprocess the HTML texts and extract sentences from them. Thirdly, each sentence is identified whether it is an advice or not by a classification learning. Finally, each extracted advice is identified whether it is a prior advice or not.

3.1 Preprocessing

Sentences are extracted from HTML texts by the following method. The texts enclosed by the HTML tags are extracted if “id” or “class” attributes in HTML tags contain any of the following strings.

content, entry, main

Otherwise, the texts enclosed by the body tags are extracted. Note that the texts enclosed by the HTML tags are eliminated from the targets if “id” or “class” attributes in HTML tags contain any of the following strings.

head, foot, menu, copy, list, comment

Then each sentence extracted from the texts is segmented into morphological elements using MeCab[6] and the sentences which meet the following requirements are extracted. The stopword list is used to eliminate advices for browsing web pages.

- The number of morphemes in the sentence is more than five.

- The sentence contains any of verbs, adverbs, adjectives or auxiliary verbs.
- The sentence does not contain any of the stopwords (ex. submit, account, browser, Adobe, JavaScript, spam, comment).

3.2 Advice acquisition

We identify whether a given sentence is an advice or not by using a classification learning. The characteristics of advices described in Section 2.3 are used as features of the classification learning. The features for acquiring advices are shown as follows. The features through *a* to *e* correspond to the characteristics through *A* to *E* in Section 2.3, respectively.

a. An auxiliary verb at the end of the sentence is any of the following auxiliary verbs.

‘ぬ, ん, ない’, ‘べし’, ‘だ’, ‘たい’, ‘た’, ‘う’, ‘です’, ‘ござる’, ‘ます’, ‘らしい’, etc.

b. The frequencies of evaluative expressions (13,590 expressions for four classes) The evaluative expressions in the dictionaries [5, 3] are used.
c. The frequencies of clue expressions (470 expressions for 35 classes)
d. Whether the content word (noun, verb, adjective and adverb) at the end of the sentence is evaluative and clue expression or not.
Note that we checked its previous content word if the content word is one of the following words

‘下さる’, ‘思う’, ‘する’, ‘いる’, ‘くれる’, ‘できる’, ‘おく’, ‘れる’, ‘られる’, ‘の’, ‘こと’, ‘もの’

e. Through *b* to *d* for the previous and next two sentences of the target sentence
f. The morpheme length
g. The ratio of the number of morphemes to the morpheme length for each part-of-speech

3.3 Situation classification of advices

We identify whether a given advice is a prior advice or not by using a classification learning. The characteristics of advices described in Section 2.4 are used as features of the classification learning. The features for classifying advices are shown as follows. The features through *a* to *d* correspond to the characteristics through *A* to *D* in Section 2.4, respectively.

a. The frequencies of clue expressions for prior (457 expressions for 17 classes) and clue expressions for action (109 expressions for nine classes)

Table 5 Size of evaluate data.

# of sentneces	# of advices	# of prior advices
1,335	172	107

b. The frequencies of action verbs (447 verbs for two classes)
We used 76 verbs which represent movement and 371 verbs which represent body movement in a thesaurus of predicate-argument structure [12].
c. Whether the content word (noun, verb, adjective and adverb) at the end of the sentence is the clue expressions and the action verbs or not.
Note that we checked its previous content word for some of the content word in the same way as the feature *d* in Section 3.2.
d. Through *a* to *c* for the previous and next two sentences of the target sentence

4 Experiment

We carried out experiments to show the performance of the proposed method and to validate our observed features.

4.1 Evaluation data

We constructed evaluation data. Firstly, we searched the web by using "高尾山 (Mt. Takao) & 登山 (climbing)" as a query and extracted top 20 web pages whose title included the location name "高尾山 (Mt. Takao)". We used Yahoo! Web API to search the web. Secondly, sentences were extracted from the web pages by the preprocessing method described in Section 3.1. Finally, we manually judged whether each sentence was an advice or not and then judged whether the advice was a prior advice or not if the sentence was an advice. The size of the evaluation data is shown in Table 5.

4.2 Experiment for acquiring advices

We carried out experiments by training the development data as training data and testing the evaluation data as testing data. As features for a classification learning, the features described in Section 3.2 were used. We used Support Vector Machines (SVM) as a machine learning model and trained SVM with a linear kernel using LibSVM[2]. We used the precision (ratio of the number of successfully acquired advices

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Table 6 Experimental results for acquiring advices.

feature	precision	recall	f-measure
word uni-gram (baseline)	49.4%	26.2%	34.2
proposed method (features through *a* to *e*)	61.2%	35.7%	45.1
proposed method - function word at sentence end (feature *a*)	56.4%	31.5%	40.5
proposed method - evaluate expressions (feature *b*)	56.5%	36.3%	44.2
proposed method - clue expressions (feature *c*)	-	0	0
proposed method - sentence end information (feature *d*)	43.4%	25.6%	32.2
proposed method - context information (feature *e*)	53.7%	34.5%	42.0

Table 7 Experimental results for classifying situation of advices.

feature	precision	recall	f-measure
word uni-gram (baseline)	68.7%	76.7%	72.5
proposed method (feature through a to d)	75.0%	81.3%	78.0
proposed method - clue expressions (feature a)	62.0%	94.4%	74.8
proposed method - action verbs (feature b)	73.1%	73.8%	73.5
proposed method - sentence end information (feature c)	72.9%	80.4%	76.4
proposed method - context information (feature d)	66.9%	75.7%	71.1

to the number of automatically acquired advices), the recall (ratio of the number of successfully acquired advices to the number of advices in the evaluation data) and the f-measure as metrics. In the baseline method, we used word uni-gram as features.

Experimental results are shown in Table 6. In comparison with the baseline method, the proposed method increased in both the precision and the recall. The f-measure decreased when each feature was eliminated from the proposed method. This results show that our features obtained by the analysis described in Section 2.3 are valid.

4.3 Experiment for classifying situation of advices

We carried out experiments by training the development data as training data and testing the evaluation data as testing data. As features for a classification learning, the features described in Section 3.3 were used. We used SVM as a machine learning model and trained SVM with a linear kernel using LibSVM. We used the precision (ratio of the number of successfully acquired prior advices to the number of automatically acquired prior advices), the recall (ratio of the number of successfully acquired prior advices to the number of prior advices in the evaluate data) and the f-measure as metrics. In the baseline method, we used word uni-gram as features.

Table 7 shows the experimental results. Both the precision and recall obtained by using the proposed method showed more increase than the baseline method. The f-measure decreased when each feature was eliminated from the proposed method. This results show that our features obtained by the analysis described in Section 2.4 are avalible for classifying situations of adivces.

5 Conclusion

In this paper, we proposed a method for extracting prior advices from the web to provide prior information before user action. The experimental results show the availability of our method. We also developed the system for providing prior advices.

For future works, we will identify whether a given advice is related to the target location and the target action or not. In addition, we would like to consider more detailed situations.

References

1. Baldauf, M., Dustdar, S., Rosenberg, F.: A survey on context-aware systems. International Journal of Ad Hoc and Ubiquitous Computing **2**(4), 263–277 (2007)
2. Cheverst, K., Davies, N., Mitchell, K., Friday, A.: Experiences of developing and deploying a context-aware tourist guide: the GUIDE project. In: Proceedings of the 6th annual international conference on Mobile computing and networking, pp. 20–31. ACM (2000)
3. Higashiyama, M., Inui, K., Matsumoto, Y.: Acquiring noun polarity knowledge using selectional preferences. In: Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing, pp. 584–587 (2008)
4. Inui, K., Abe, S., Morita, H., Eguchi, M., Sumida, A., Sao, C., Hara, K., Murakami, K., Matsuyoshi, S.: Experience mining: Building a large-scale database of personal experiences and opinions from web documents. In: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 314–321 (2008)
5. Kobayashi, N., Inui, K., Matsumoto, Y., Tateishi, K., Fukushima, T.: Collecting evaluative expressions for opinion extraction. In: Proceedings of the 2nd International Joint Conference on Natural Language Processing, pp. 584–589 (2004)
6. Kudo, T.: Mecab: Yet another part-of-speech and morphological analyzer. URL http://mecab.sourceforge.jp/
7. Kurashima, T., Fujimura, K., Okuda, H.: Discovering association rules on experiences from large-scale blog entries. Advances in Information Retrieval **5478/2009**, 546–553 (2009)
8. Oku, K., Nakajima, S., Miyazaki, J., Uemura, S., Kato, H.: A recommendation method considering users' time series contexts. In: Proceedings of the 3rd International Conference on Ubiquitous Information Management and Communication, pp. 465–470. ACM (2009)
9. Park, K.C., Jeong, Y., Myaeng, S.H.: Detecting experiences from weblogs. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 1464–1472. Association for Computational Linguistics (2010)
10. Saeger, S.D., Torisawa, K., Kazama, J.: Looking for trouble. In: Proceedings of the 22nd International Conference on Computational Linguistics, pp. 185–192. Association for Computational Linguistics (2008)
11. van Setten, M., Pokraev, S., Koolwaaij, J.: Context-aware recommendations in the mobile tourist application COMPASS. In: International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, pp. 235–244 (2004)
12. Takeuchi, K., Inui, K., Takeuchi, N., Fujita, A.: A thesaurus of predicate-argument structure for japanese verbs to deal with granularity of verb meanings. In: Proceedings of the 8th Workshop on Asian Language Resources, pp. 1–8 (2010)
13. Zheng, V.W., Zheng, Y., Xie, X., Yang, Q.: Collaborative location and activity recommendations with gps history data. In: Proceedings of the 19th international conference on World wide web, pp. 1029–1038. ACM (2010)