Providing Ad Links to Travel Blog Entries Based on Link Types

Aya ISHINO
Graduate School of Information Sciences, Hiroshima City University, Hiroshima, Japan
ishi-no@ls.info.hiroshima-cu.ac.jp

Hidetsugu NANBA
Graduate School of Information Sciences, Hiroshima City University, Hiroshima, Japan
nanba@hiroshima-cu.ac.jp

Toshiyuki TAKEZAWA
Graduate School of Information Sciences, Hiroshima City University, Hiroshima, Japan
takezawa@hiroshima-cu.ac.jp

Abstract

Content-targeted advertising systems are becoming an increasingly important part of the funding for free web services. These programs automatically find relevant keywords on a web page, and then display ads based on those keywords. We propose a method for providing links to ads for travel products (which we call ad links) automatically. We extract keywords from citing areas of travel information links, and provide appropriate ad links. To investigate the effectiveness of our method, we conducted experiments. We obtained a high precision for the extraction of keywords and provision of ad links.

1 Introduction

Online advertising is a form of promotion that uses the World Wide Web for the expressed purpose of delivering marketing messages to attract customers. Examples of such online ads are contextual advertising and listing advertising. Content-targeted advertising systems, such as Google's Ad Sense program and Yahoo's Contextual Match product, are becoming an increasingly important part of the funding for free web services. These programs automatically find relevant keywords on a web page, and then display ads based on those keywords. We propose a method for providing ad links for travel products to travel blog entries, which are travel journals written by bloggers in diary form. By specifying the travel domain, we aim to provide more appropriate ad links than existing content-targeted advertising systems.

To provide these ad links, we take account of the types of hyperlinks in each travel blog entry. Ishino *et al.* (2011) devised a method for constructing a collection of web links for travel information automatically. They extracted the hyperlinks by which bloggers describe useful web sites for a tourist spot from travel blog entries, and classified types of travel information link into the following four categories.

- S (Spot): The information is about tourist spots.
- H (Hotel): The information is about accommodation.
- R (Restaurant): The information is about restaurants.
- O (Other): Other than types S, H, and R.

By switching strategies for providing ad links according to these types, we attempt to provide more appropriate ad links for travel.

The remainder of this paper is organized as follows. Section 2 shows the system behavior in terms of snapshots. Section 3 discusses related work. Section 4 describes our methods. To investigate the effectiveness of our methods, we conducted some experiments, and Section 5 reports on these and the results. We present some conclusions in Section 6.

2 System Behavior

In this section, we describe our prototype system, which provides information about (1) useful web sites for tourist spots and (2) ads for travel products. The two steps in the search procedure are:

(Step 1) Input a keyword, such as "Okonomiyaki" (Japanese-style pancake), in the search form (shown as ① in Figure 1).

(Step 2) Click the "link" button (shown as ②) to generate a list of URLs for web sites related to the keyword together with automatically identified link types using Ishino's method (Ishino et al. 2011), the context of citations ("citing areas"), by which the authors of the travel blog entries describe the sites, and ads for travel products related to the citing areas. We classified link types into the four categories described above.

(Step 3) Click the "link"(shown as ③) to display detailed information of ads for travel products (shown in Figure 2).

We propose a method for providing ad links to travel products corresponding to the link type.

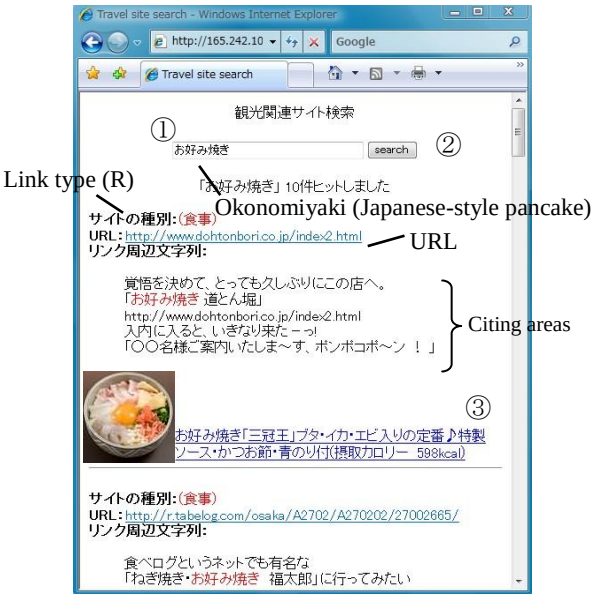

Figure 1. A list of travel information links together with automatically inserted ad links

Figure 2. An example of a web page for a travel product

3 Related Work

In this section, we describe some related studies. Recommendation systems provide a promising approach to ranking commercial products or various documents according to a user's interests. These systems can be classified into two categories by their underlying method of recommendation: (1) collaborative filtering (Goldberg *et al.* 1992) and (2) content-based filtering (Sarwar *et al.* 2000). We focus on the online advertising using the content-based filtering techniques.

Examples of online advertising are listing advertising and contextual advertising. Listing advertising is a method of placing online ads on web pages that show results from search engine queries. Search advertisements are targeted to match the entered keywords. Fujita *et al.* (2010) showed a system that automatically generates shop-specific listing ads by reusing textual data promoting each shop. They used a restaurant portal site as textual data. We use citing areas of travel information links, and provide ad links to related travel products.

In this paper, we focus on contextual advertising, which is based on keywords automatically extracted from the text of the web page. Keyword extraction is the core task of the contextual advertising system. There are a number of classical approaches to extracting keywords. TF*IDF uses a frequency criterion to select keywords. Yih *et al.* (2006) showed a learning-based technique using TF*IDF for contextual advertising. Recently, new methods based on the Wikipedia corpus have been proposed. The Wikify! system (Mihalcea and Csmomai, 2007) identifies keywords in a text, and then links these keywords to the corresponding Wikipedia pages. They identified keywords using a criterion called keyphraseness. They applied their method to English texts. We extract keywords from Japanese citing areas. Therefore, we cannot apply their method for Japanese citing areas.

4 Automatic Organization of Travel Information through Blogs

The task of organizing travel through blogs is divided into two steps: (1) classification of links in travel blog entries, and (2) providing ad links for travel information links. These steps are explained in Sections 4.1 and 4.2.

4.1 Link Classification

Ishino *et al.* (2011) devised a method for classifying links in travel blog entries into the four categories described in Section 1. The procedure is as follows.

1. Input a travel blog entry.
2. Extract a hyperlink and any surrounding sentences that mention the link (a citing area).
3. Classify the link by taking account of the information in the citing area.

In the following, we will explain Steps 2 and 3.

Extraction of citing areas

Ishino *et al.* manually created rules for the automatic extraction of citing areas. These rules use cue phrases.

Method of link type classification

They classified hyperlinks automatically. They employed a machine-learning technique using the following features. Here, a sequence of nouns (a noun phrase) was treated as a noun.

- A word.
- Whether the word is a cue phrase, detailed as follows, where the numbers in brackets shown for each feature represent the number of cues (shown in Table 1, Table 2, and Table 3).

Cue phrase	Number of cues
A list of tourist spots, collected from Wikipedia	17,371
Words frequently used in the names of tourist spots, such as “動物園” (zoo) or “博物館” (museum)	138
Words related to sightseeing, such as “見学” (sightseeing) or “散策” (stroll)	172
Other words	131

Table 1. Cues for type S

Cue phrase	Number of cues
Words frequently used in the name of hotels, such as “ホテル” (hotel) or “旅館” (Japanese inn)	9
Component words for accommodation, such as “フロント” (front desk) or “客室” (guest room)	29
Words frequently used when tourists stay in accommodation, such as “泊る” (stay) or “チェックイン” (check in)	14
Other words.	21

Table 2. Cues for type H

Cue phrase	Number of cues
Dish names such as “omelet”, collected from Wikipedia	2,779
Cooking styles such as “Italian cuisine”, collected from Wikipedia	114
Words frequently used in the names of restaurants, such as “レストラン” (restaurant) or “食堂” (dining room)	21
Words used when taking meals, such as “食べる” (eat) or “おいしい” (delicious).	52
General words that indicate food, such as “ご飯” (rice) or “料理” (cooking)	31
Other words	31

Table 3. Cues for type R

To investigate the effectiveness of their method, Ishino *et al.* conducted an experiment. The evaluation results are shown in Table 4. We applied this model to 17,266 travel blog entries, and classified 4,155 links. The numbers of automatically classified links are shown in Table 5. We call these links travel information links. We describe the method for providing ad links to travel products in the following section.

Link types	Recall (%)	Precision (%)
S	62.5	72.7
H	64.9	81.3
R	71.9	76.7
O	71.6	48.6

Table 4. Evaluation results for link classification (Ishino *et al.* 2011)

Link types	S	H	R	O
Number of links	1,174	123	921	2,061

Table 5. The number of links of each type

4.2 Providing Ad Links for Travel Information Links

The procedure for providing ad links for travel information links is as follows.

1. Input a link type and the citing areas of a travel information link.
2. Extract keywords from the citing areas.
3. Extract product data containing all keywords, and calculate the similarity be-

tween the citing areas of a travel information link and the product data.

4. Provide an ad link to the product data having the highest similarity for the travel information link.

In the following, we will explain Steps 2 and 3.

Keyword extraction based on link types

We extract keywords for travel products corresponding to the link type. We use the cues used by Ishino's method for classifying travel information links, and extract keywords from citing areas of link types S and R.

First, we describe the method for extracting keywords from citing areas of link type S. The cues for type S shown in Table 1, such as tourist spots collected from Wikipedia and words frequently used in the names of tourist spots, tend to become keywords. Therefore, we register these cues as candidate keywords for link type S. If citing areas of link type S contain candidate keywords, we extract them as keywords. In addition, if citing areas contain names of places, we extract them as keywords. Candidate keywords for link type S are shown in Table 6.

Next, we describe a method for extracting keywords from citing areas of link type R. The cue for type R shown in Table 3, such as dish names and cooking styles, tend to become keywords. Therefore, we register these cues as candidate keywords for link type R. If citing areas of link type R contain candidate keywords, we extract them as keywords. Candidate keywords for link type R are shown in Table 7. We used CaboCha software (http://chasen.org/~taku/software/cabocha/) to identify location names.

Candidate keywords	Number of candidate keywords
A list of tourist spots, collected from Wikipedia	17,812
Words frequently used in the names of tourist spots, such as "動物園" (zoo) or "博物館" (museum)	138
Names of places	

Table 6. Candidate keywords for link type S

Candidate keywords	Number of candidate keywords
Dish names such as "omelet", collected from Wikipedia	2,779
Cooking styles such as "Italian cuisine", collected from Wikipedia	114

Table 7. Candidate keywords for link type R

Product data extraction based on the link types

We extract product data, which contain all keywords, and calculate the similarity between citing areas of a travel information link and the product data.

Product data

We provide ad links to travel products of Rakuten Shopping Mall (Rakuten Ichiba) or facility data of Rakuten Travel released through the Rakuten Institute of Technology for travel information links. An example of product data of Rakuten Ichiba is shown in Figure 3. The product data contain 50 million items. An item has a name, a code, a price, descriptive texts, URL, picture, shop code, category ID and registration date. The facility data contain 11,468 facilities. A facility data entry has a name, ID number, and user review.

	Product data
Name	[marine diving] Earthly paradise OKINAWA 2009
Code	seasir-umi:10001011
Price	980
Descriptive text	*The guidebook contains extended information about diving in Okinawa! *It introduces 140 dive spots!
URL	http://item.rakuten.co.jp/seasir-umi/r023/
Picture	@0_mall/seasir-umi/cabinet/09shohin/rakuenokinawa_2009.jpg
Category ID	101922
Registration date	2010/03/24 15:08:58

Figure 3. Product data

We provide ad links to product data having associations with travel information links by using the link types of the travel information links. We take account of the characteristics of link types, and assign each category to link types. Categories of product data are shown in Table 8.

Facility data do not contain descriptive texts about the facilities. Therefore, we proposed a method for providing ad links for travel information links of link types S and R.

Link types	Category	Number of product data items
S (Spot)	Product data related to travel.	51,516
H (Hotel)	Facility data.	11,468
R (Restaurant)	Product data related to food.	830,807

Table 8. Category of product data assigned to each link type.

We describe a method for collecting product data related to travel. Product data has a category ID. A category master, shown in Table 9, was released through the Rakuten Institute of Technology. The category master has a hierarchic structure. First, we collect subcategories of the category "Travel, Study abroad, Outdoor amusement". In this way, we collect category IDs related to travel, and show the category IDs in Table 10. Next, we collect product data with category ID related to travel as product data related to travel. In the same way, we collect product data related to food.

Category ID	Category name	Subcategory ID
101242	Travel, Study abroad, Outdoor amusement	200162
209835	SANYO	209830
503221	Panasonic	503218

Table 9. The category master

Category ID	Category name
208922	Hot spring
208924	Theme park
208925	Guidebook
208930	Travel book, Travel essay
411387	Mountain climbing, Outdoor amusement, Camp

Table 10. Category IDs related to travel

Similarity calculation

We describe the method for calculating the similarity between citing areas and product data. We extract product data that contain all keywords *K*, and calculate the similarity *Score* between citing areas of a travel information link and the product data. *Score* is calculated as follows:

$$Score = \sum_{k_i \subseteq K} Link\_Score(k_i) * Advertising\_Score(k_i) \quad (1)$$

Link_Score(k_i) is the number of times the given keywords k_i appears in that citing area. *Advertising_Score(k_i)* is the number of times the given keyword k_i appears in the descriptive text of the product data. We provide an ad link to the product having the highest score for the travel information link.

5 Experiments

To investigate the effectiveness of our methods, we conducted several experiments.

5.1 Experimental Method

To generate the test data for providing ad links, we randomly selected 50 travel information links of type S and 50 of type R and manually classified them. To investigate the effectiveness of our method, we extracted keywords using the following two methods for evaluation of keywords in Section 5.2 and ad links in Section 5.3.

- Our method: Extract keywords by our method, as described in Section 4.2.
- TF*IDF: TF*IDF is a conventional baseline used in the scientific literature for comparison of keywords extraction algorithms. IDF was calculated using an open API (http://developer.yahoo.co.jp/webapi/search/websearch/v2/websearch.html). We extract the top-X words as keywords.

In addition, we evaluated our method for ad links for the limited categories described in Section 4.2 and all categories of product data.

5.2 Evaluation of Keywords

We evaluated the method for keywords extraction. We used precision as the evaluation measure, calculated as follows:

$$\text{Precision} = \frac{\text{the number of appropriative keywords}}{\text{the number of extravted keywords}} \quad (2)$$

The evaluation results are shown in Table 11. Our method attained higher precision than baseline methods.

	Link Type S (%)	Link Type R (%)
Our method	**81.8 (72/88)**	**98.0 (49/50)**
TF*IDF(X=1)	46.0 (23/50)	48.0 (24/50)
TF*IDF(X=2)	38.0 (38/100)	37.4 (37/99)
TF*IDF(X=3)	36.0 (54/150)	36.5 (54/148)
TF*IDF(X=4)	34.0 (68/200)	31.5 (62/197)
TF*IDF(X=5)	34.0 (85/250)	32.5 (80/246)

Table 11.Precision of extracted keywords

There were two typical errors in keyword extraction: (1) the lack of cues and (2) a problem with the method of keywords extraction. We describe these errors as follows.

(1) The lack of cues:
For keywords extraction, we used manually selected cues, as described in Section 4.2. To improve the coverage of cues, a statistical approach, such as applying n-gram statistics to a larger blog corpus, will be required.

(2) The problem with the method of keywords extraction:

In the following example, our method mistakenly extracted "天然温泉" (a natural hot spring) as a keyword, because the candidate keywords for link type S "天然温泉" (a natural hot spring) appears in the citing areas.

[original]
お湯のヌルヌル感と温まり具合は最高です。
高浜の「湯っぷる」
詳しいことは→
http://www.seaside-takahama.com/
ここは天然温泉じゃないんですが
[translation]
The quality and the temperature of spring water are great.
"YUPPLE" in Takahama.
For details, access the following web page:
http://www.seaside-takahama.com/
YUPPLE is not a natural hot spring.

5.3 Evaluation of Ad Links

We evaluated the method for providing ad links for travel information links. We used precision and coverage as evaluation measures, which were calculated as follows:

$$\text{Precision} = \frac{\text{the number of appropriative travel information links provided ad links}}{\text{the number of travel information links provided ad links}} \quad (3)$$

$$\text{Coverage} = \frac{\text{the number of travel information links provided ad links}}{\text{the number of travel information links}} \quad (4)$$

The evaluation results are shown in Table 12 and Table 13. We obtained a higher precision than the baseline methods.

	Precision (%)	Coverage (%)
Our method	**79.3 (23/29)**	**58.0 (29/50)**
TF*IDF(X=1)	39.1 (18/46)	92.0 (46/50)
TF*IDF(X=2)	37.8 (14/37)	74.0 (37/50)
TF*IDF(X=3)	35.9 (7/20)	40.0 (20/50)
TF*IDF(X=4)	36.4 (4/11)	22.0 (11/50)
TF*IDF(X=5)	0.0 (0/2)	4.0 (2/50)

Table 12. Precision and coverage of providing ad links for travel information links of type S

	Precision (%)	Coverage (%)
Our method	**91.3 (21/23)**	**46.0 (23/50)**
TF*IDF(X=1)	22.0 (11/50)	100.0 (50/50)
TF*IDF(X=2)	29.5 (13/44)	88.0 (44/50)
TF*IDF(X=3)	25.0 (7/28)	56.0 (28/50)
TF*IDF(X=4)	6.3 (1/16)	32.0 (16/50)
TF*IDF(X=5)	25.0 (2/8)	16.5 (8/50)

Table 13. Precision and coverage of providing ad links for travel information links of type R

First, we discuss the results for link type S. As shown in Table 12, our method attained higher precision than the baseline methods. For coverage, TF*IDF (X=1) and TF*IDF (X=2) were better than our method, while we obtained more appropriate travel information links provided as ad links than the baseline method.

Next, we discuss the results for link type R. As shown in Table 13, our method attained higher precision than the baseline methods. For coverage, TF*IDF (X=1), TF*IDF (X=2) and TF*IDF (X=3) were better than our method, while we obtained larger numbers of appropriate travel information links provided as ad links than the baseline methods. Therefore, our experimental results confirm the effectiveness of our methods.

We discuss a typical error in providing links. In the following example, we could not provide ad links. Our method extracted an appropriative keyword "Nihon-heso-kōen" (The Navel Park,

NISHIWAKI, JAPAN). Nihon-heso-kōen is a park in Hyogo prefecture, Japan. However, there are no product data containing this word, and we could not provide any ad links for the travel information link. To solve this problem, we took the address of Nihon-heso-kōen from web pages, and provided an ad link to a guidebook that describes tourist spots near Nihon-heso-kōen for the travel information link.

[original]
今回初めてお会いした、Kamon さん奥さんは、モデルさんのようなとても綺麗な方でした！
このメンバーで、第一目的地”日本へそ公園”へ
ここでの目的は、やっぱり”ぶーにゃん”に会う事・・・
が、今回も会えませんでしたつ
Д｀)・゜・。・゜゜・*:.。
[translation]
I met with Mrs. Kamon for the first time. She was very beautiful like a model!
We came to visit “Nihon-heso-kōen” (The Navel Park, NISHIWAKI, JAPAN).
We expected to see “Bu-nyan”, but we missed again :-(

5.4 Evaluation of Limited Category

We evaluate our method for ad links for the limited category shown in Section 4.2 and for all categories of product data. We used precision and coverage as evaluation measures, as described in Section 4.3. The evaluation results are shown in Table 14 and Table 15.

	Precision (%)	Coverage (%)
Our method (particular category)	79.3 (23/29)	58.0 (29/50)
Baseline (all categories)	38.2 (13/34)	68.0 (34/50)

Table 14. Precision and coverage of ad links for travel information links of type S (evaluation particular category)

	Precision (%)	Coverage (%)
Our method (particular category)	91.3 (21/23)	46.0 (23/50)
Baseline (all categories)	70.8 (17/24)	48.0 (24/50)

Table 15. Precision and coverage of ad links for travel information links of type R (evaluation of particular category)

As shown in the tables, our method attained higher precision than the baseline method. For coverage, baselines of types S and R were better than our method, while we obtained a larger number of appropriate travel information links provided as ad links than the baseline method. Therefore, our experimental results confirm the effectiveness of our methods.

As examples of more appropriate ad links than our method could provide, the baseline method could provide ad links to product data not directly related to travel, such as books, CDs and DVDs that describe tourist spots. These are not categorized as product data related to travel in our system. In our future work, we plan to classify such product data as related to travel automatically so that we can provide more appropriate ad links.

6 Conclusion

We have proposed a method for providing ad links for travel information links automatically. For the extraction of keywords, we obtained 81.8% precision for link type S and 98.0% precision for link type R. For providing ad links, we obtained 79.3% precision for link type S and 78.3% precision for link type R. Our method also provided larger numbers of appropriate travel information links as ad links than the baseline method. Our experimental results have confirmed the effectiveness of our methods.

7 Future Work

For our method, we used manually selected cues. To increase the number of cues, a statistical approach is required.

In this paper, we have focused on travel information links written in Japanese. In our future work, we will translate cue phrases from Japanese into other languages, and apply our method to travel information links in various languages.

Reference

Aya Ishino, Hidetsugu Nanba, and Toshiyuki Takezawa. 2011. Automatic Compilation of an Online Travel Portal from Automatically Extracted Travel Blog Entries. *Proceedings of ENTER 2011*, 113-124.

Hidetsugu Nanba, Haruka Taguma, Takahiro Ozaki, Daisuke Kobayashi, Aya Ishino, and Toshiyuki Takezawa. 2009. Automatic Compilation of Travel Information from Automatically Identified Travel Blogs. *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing, Short Paper*, 205-208.

Atsushi Fujita, Katsuhiro Ikushima, Satoshi Sato, Ryo Kamite, Ko Ishiyama, and Osamu Tamachi. 2010. Automatic Generation of Listing Ads by Reusing Promotional Texts. *Proceedings of the 12th International Conference on Electronic Commerce (ICEC)*, 191-200.

David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. 1992. Using Collaborative Filtering to Weave an Information Tapestry. *Communications of the ACM*, 35(12), 61-70.

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2000. Analysis of Recommendation Algorithms for E-commerce. *Proceedings of the 2nd ACM Conference on Electronic Commerce*, 158-167.

Wen-tau Yih, Joshua Goodman, and Vitor R. Carvalho. 2006. Finding Advertising Keywords on Web Pages. *Proceedings of the 15th International Conference on World Wide Web*.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: Linking Documents to Encyclopedic Knowledge. *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, 233-242.