Enriching Travel Guidebooks with Travel Blog Entries and Archives of Answered Questions

Kazuki Fujii[a],
Hidetsugu Nanba[a],
Toshiyuki Takezawa[a], and
Aya Ishino[b],

[a] Graduate School of Information Sciences, Hiroshima City University
[b] Department of Information Systems in Business, Hiroshima University of Economics
a) {fujii, nanba, takezawa}@hiroshima-cu.ac.jp, b) ay-ishino@hue.ac.jp
ay-ishino@hue.ac.jp

Abstract

Travellers planning to visit particular tourist spots need information about their destination and they often use travel guidebooks to collect this information. However, guidebooks lack specific information, such as first-hand accounts by users who have visited the specific destination. To compensate for the lack of such information, we focused on travel blog entries and archives of answered questions. In this paper, we propose a method for enriching guidebooks by matching and aligning the information with blog entries and questions answered (QA) archives. This is a three-step method. In Step 1, we classify pages of guidebooks, blog entries, and QA archives into five types of content, such as "watch," "dine," etc. In Step 2, we align each blog entry and QA archive with guidebooks by taking these content types into account. In Step 3, we match each blog entry and QA archive with individual pages in guidebooks. We conducted some experiments, and confirmed the effectiveness of our method.

Keywords: travel guidebook; blog; QA archive.

1 Introduction

Travellers planning to visit particular tourist spots need information about these tourist destinations, and they often use travel guidebooks to collect this information. Guidebooks give basic information about famous tourist spots, souvenir shops, and restaurants. As another method to collect the information, there are portal sites that are operated by travel companies and local governments for the benefit of the tourist. However, as there are also sites that are not updated frequently and there are large differences in the amount of information on each tourist destination, many travellers use guidebooks to get basic information.

Travellers who read guidebooks only get basic information from them and cannot determine how to travel between the destinations listed in the guidebooks or at which hotel to stay with their family. To help travellers read guidebooks, we focused on travel blog entries and archives of questions answered (QA), both of which contain valuable information, such as first-hand accounts by users who have visited the particular destination. In addition, blog entries and QA archives have more descriptive content than other SNSs, such as twitter or various review sites. Therefore, blog entries and QA archives are considered useful information sources for obtaining travel information.

In this paper, we propose a method for enriching guidebooks by matching them against blog entries and QA archives. In addition, we constructed a prototype system that enables guidebook enrichment. By using our system, users can obtain basic information from guidebooks and valuable information based on the personal experiences of travellers from blog entries and QA archives.

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

2 System Behaviour

In this section, we describe our prototype system, which can provide enriched guidebooks. For scanned and OCR-processed guidebooks, our system matches blog entries and QA archives to each guidebook page automatically.

Fig.1. An example of a page of an enriched guidebook. This page gives details about *Kakeromajima* Island in Japan.

Figure 1 shows an example of a page of an enriched guidebook; it gives details about tourist spots and accommodation facilities on *Kakeromajima* Island. In this figure, when a user clicks the "Blog" button (1), a list of blog entries related to the guidebook page is shown. In the same way, when a user clicks the "QA" button (2), all QA archives related to the guidebook page are shown. Figure 2 is an example of a QA archive that was aligned with the guidebook page shown in Figure 1. In this archive, the questioner asked recommendation of guesthouses in *Kakeromajima* Island. For this question, the answerer recommended guesthouses in *Ukejima* Island near to *Kakeromajima*. Users of our system can know basic information from guidebooks,

while additional valuable information from blog entries and QA archives, which were automatically aligned by our system.

There are several travel portals, such as "Rakuten Travel (http://travel.rakuten.co.jp/)" and restaurant review sites, such as "Tabelog (http://tabelog.com/)." Although these sites provide customer reviews, we focus on blog entries and QA archives, because they contain a lot of valuable information such as follows.

- Local information provided by local people, most of which are not written in guidebooks.
- Tourist spots or events influenced by transportation. (e.g. "I don't recommend to elderly people, because we need to go up and down a hill a couple of times.")
- Unrecommended information, such as shown in Figure 2.

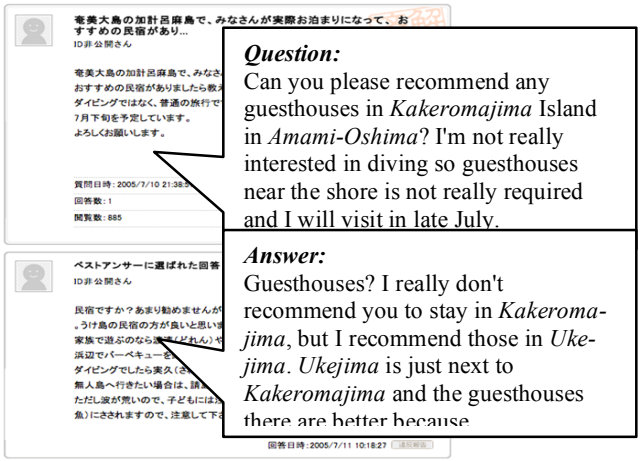

Fig. 2. An example of an automatically aligned QA archive for the guidebook page shown in Figure 1.

3 Related Work

In this section, we introduce some studies on enriching texts. Textbooks used in developing regions are largely text-oriented and lack good visual materials. Therefore, Rakesh *et al.* (2011) proposed a method that enhances the quality of textbooks by assigning images in Wikipedia to each section in textbooks. Nie *et al.* (2013) proposed a method that is able to enrich textual answers in QA archives. For some questions, such as "*What are the steps to make a weather vane,*" textual answers may not provide sufficient natural and easy-to-grasp information. Therefore, they automatically provide accompanying videos and images that visually demonstrate the process of the objects. Lu *et al.* (2009) proposed a method to visualize textual travel blog with images. Bressan *et al.* (2008) proposed a travel blog assistant system that facilitated the travel blog writing by selecting for each blog paragraph the most relevant images from an image set.

All these works focused on enriching texts by multimedia data for supporting to understand the texts, while we focus on enriching multimedia data (guidebooks

including both texts and images) with other textual data, such as QA archives and blogs. Especially significant characteristic of our work is that our target data are printed guidebooks, because our method is easily applicable to any printed materials, and has a potentiality to enrich them.

4 Enriching Guidebooks with Blog Entries and QA Archives

We propose a method for enriching guidebooks by aligning the content with blog entries and QA archives. In this section, we describe our method for implementing this alignment. The procedure of our method consists of the following three steps.

Step 1: Content-type (see Table 1) classification of guidebook pages, blog entries, and QA archives. One guidebook page is aligned with blog entries and QA archives, which have the same content type as the guidebook page. The results of the content-type classification are used in Step 3.

Step 2: Alignment of blog entries and QA archives with a guidebook.[1]

Step 3: Alignment of blog entries and QA archives with a guidebook page, having the same content-type, which was determined in Step 1. The results of Step 3 are the final output for enriched guidebooks.

These steps are described in detail in Sections 4.1, 4.2, and 4.3, respectively.

4.1 Content-Type-Classification of Guidebook Pages, Blog Entries, and QA Archives

4.1.1 Definition of Content Types

Generally, each guidebook page contains travel information in several content types, such as sightseeing spots, souvenir shops, or accommodation. Each guidebook page can be classified into one of the six content types that are typical for tourism (Table 1).

Table 1. Content types and their descriptions

Content Type	Criterion
Watch	Page about sightseeing for watching enjoyment
Experience	Page about experience (scuba diving, dance)
Buy	Page about shopping or souvenir stores
Dine	Page about drinking and dining
Stay	Page about accommodation
Other	None of the above applies to this page

It is possible to supplement a guidebook page with more relevant information by aligning blog entries and QA archives that have the same content type as the guidebook page.

The information published in guidebook pages was mainly content of type “watch,” “experience,” “buy,” “dine,” and “stay.” We defined guidebook pages that did not belong to these content types as content type “other.” Guidebook pages that were judged to belong to “other” contain a lot of information, not all of which is directly

[1] We employed a coarse-to-fine alignment, that is to align in a coarse (guidebook) level (Step 2), then in a fine (page) level (Step 3).

related to tourist destinations, for example, advertisement pages, transportation information for arriving at the tourist destination, and also how to get a passport. Therefore, we align blog entries and QA archives that have the same content type as the guidebook pages with the guidebook pages that were classified into "watch," "experience," "buy," "dine," or "stay." We classify guidebook pages, blog entries, and QA archives into each content type listed in Table 1, other than content type "other," to take advantage of alignment. Guidebook pages of each content type have already been published; however, it is not realistic to classify them manually because guidebooks are updated on a regular basis. Hence, guidebook pages were also subjected to automatic content-type classification. A guidebook page was classified into "watch" and "buy" if it contained information about "watch" and "buy." Blog entries and QA archives were also treated similarly. It is considered that by preforming such a content-type classification, we can align blog entries and QA archives of appropriate content type even if the guidebook page was classified into several types.

For content-type classification, we employed a machine-learning technique using text information and image information; these are explained in Sections 4.1.2 and 4.1.3, respectively.

4.1.2 Content-Type-Classification using Text Information

In a guidebook page belonging to "watch," words such as "display" and "museum" appeared frequently. A page judged as belonging to "experience" contained words such as "instructor" and "beginner." Namely, the guidebook page, which was classified into each content type, tended to contain words specific to the content type. Moreover, words in guidebooks that were not judged as "watch," "experience," "buy," "dine," or "stay," were used as the words specific to the content type "other." Therefore, we collected words that were peculiar to each content type for machine learning, and employed information gain (IG)[2] as the feature selection method.

Our method computed the value of the words in each content type by the IG. The words used were nouns, verbs, and adjectives. In addition, out of these words, we removed stop words longer than 15 letters, shorter than 2 letters, alphanumeric characters, and words with appearance frequency of less than two. We used MeCab (http://mecab.sourceforge.net/) as a Japanese, morphological analysis tool to identify the part of speech. We collected cue words that have a value over the threshold of the IG. The threshold value was determined by the preliminary experiment. In the same way, we collected cue words from blog entries and QA archives.

4.1.3 Content-Type-Classification Using Image Information

Guidebooks contain many images. For example, pages judged to belong to content-type "watch" may include many images of mountains or seashores. Pages judged as belonging to content-type "dine" may include many images of various foods. For this reason, we assumed that images were good clues for classifying the content type by analysing what the images contain. Therefore, for the content-type classification of pages of guidebooks, we also employed image information as features for machine

[2] For example, if we collect words that are peculiar to "watch" type, we compare words in type "watch" blog entries with other ones.

learning. We used one guidebook page as one image. We adopted the Bag of Visual Words (BoVW) (Csurka *et al.*, 2004) for image information. BoVW represents images by vectors of the frequency of appearance of local features; it is also a popular method for the task of object recognition. BoVW was applied Bag of Words, which represents documents by vectors of appearance frequency of words in natural language processing to images.

First, we extracted local features from the training image data by dense sampling. Second, we extracted local features and performed clustering to create a representative vector (visual word). We opted for k-means as the method for clustering feature vectors of images. The many feature vectors were clustered into 1,000 visual words by k-means (1,000 visual words were created in a preliminary experiment). Finally, we generated BoVW by histogram by counting the number of occurrences of visual words as an approximation of guidebook pages. BoVW features were generated for each page. The generated features were used for machine learning.

Blog entries also include many images, which are important clues. However, we could not collect images in blog entries in Yahoo! Blog (http://blogs.yahoo.co.jp/), because images are embedded by JavaScript, and we were not allowed to crawl from the blog server. For this reason, the type classification in blog entries uses only text information.

4.2 Alignment of Blog Entries and QA Archives with Guidebooks

In this section, we first explain why we employed a coarse-to-fine alignment. Let us consider aligning blog entries and QA archives about *Miyajima* Island with the corresponding page in a guidebook of Hiroshima. The island is located in Hiroshima, but the page does not necessarily include the word "Hiroshima" despite the fact that each page contains information about Hiroshima. In this case, it is difficult to align blog entries and QA archives that are relevant to information for the guidebook's page contents and for *Miyajima*. Therefore, it is necessary to take account of both global and local contexts of a guidebook for more accurate alignment of a guidebook page with blog entries and QA archives. There is a related work focusing on global and local contexts of a document. He *et al.* (2010) removed cited documents from an article, and estimated cited documents from text surrounding the citation or the placeholder. They considered that some parts of the document such as the title and abstract have a global context and allow words related to the article to be extracted. The text surrounding a citation or placeholder comprises local context and allows characteristic words related to the citation to be extracted. They enhanced the precision of estimation of the bibliography by using extracted words. We can consider that our framework is similar to their framework by regarding the local context as one page of the guidebooks. In particular, we consider global context as the words related to the guidebook such as "Hiroshima," and local context as the words that appear on the guidebook page. Therefore, we first align blog entries and QA archives with the guidebook to improve the precision of alignment, and then align with the guidebook page.

The remainder of this section explains how to align blog entries and QA archives with a guidebook. As the location names of tourist spots frequently appear in blog entries, QA archives, and guidebooks, the appearance of the location name is important for

alignment. We align blog entries and QA archives with guidebooks by using the appearance frequency of the location name. To extract location names from guidebooks, blog entries, and QA archives, we used CaboCha (http://code.google.com/cabocha/) as the Japanese syntactic parser.

We used the results of content-type classification in Step 1 to align blog entries and QA archives with the guidebook page. Here, we explain why the results of content-type classification are necessary. Generally, it is known that context has typical content-type constituents. For example, academic articles have typical constituents, namely "background," "purpose," "method," "conclusion," or "consideration." Kando (1997) reported on a method that creates an index by using only sentences that have a particular type and found that it is more accurate than the method that uses academic article detail in the task of academic articles retrieval. The constituents of guidebooks are the content types shown in Table 1. We attempted to achieve a more accurate alignment by using the results of content-type classification.

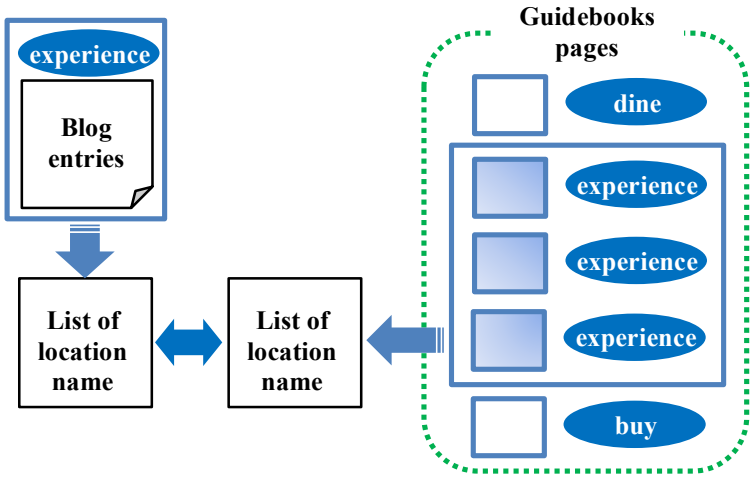

Fig. 3. Alignment of blog entries with guidebook pages using content-type classification

Figure 3 shows how to use the results of content-type classification for aligning blog entries and QA archives with guidebooks. When we selected the guidebooks that were aligned with blog entries classified into "experience," we used location names that were extracted from the blog entries and the guidebook page having the same content-type, "experience." A similar flow of operations was performed in the case of other content types: "buy," "watch," "dine," and "stay." Using extracted location names for aligning blog entries and QA archives with the guidebook page, we proposed two methods, namely the k-Nearest Neighbour (KNN) method and the Support Vector Machines (SVM) method. In the KNN method, we used the SMART measure to calculate the similarity between guidebooks and blog entries or QA archives. Blog entries and QA archives can be aligned with the guidebook page when the similarity is higher than the threshold. The threshold was determined to be the value when the recall was above 15% and the precision was the highest in training data by the

twofold cross-validation test. In the SVM method, we constructed models for each guidebook. Models classify whether intended blog entries and QA archives can be aligned with guidebooks.

4.3 Alignment of Blog Entries and QA Archives with a Guidebook Page

We extracted location names from guidebooks, blog entries, and QA archives, and then calculated the similarity between them for the same content type and aligned with the guidebook page having the highest similarity. The similarity calculation uses the cosine degree. We used the tf*idf for term weighting. The idf scores were calculated by using the number of hits in the web search engine.

5 Experiments

We conducted three experiments: (1) content-type classification of pages of guidebooks, blog entries and QA archives, (2) alignment of blog entries, and QA archives with guidebooks, and (3) alignment of blog entries and QA archives with pages of guidebooks. We report on these in Sections 5.1, 5.2, and 5.3, respectively.

5.1 Content-Type-Classification of Guidebook Pages, Blog Entries, and QA Archives

Datasets and Experimental Setting

We used 2,897 pages of 20 guidebooks, 1,000 blog entries that were collected by the method of Nanba *et al.* (2009), and 9,388 QA archives registered in the "region / travel" category of Yahoo! Answers. Then we classified them manually, and used them for our examination. It is possible to classify a guidebook page, blog entry, and QA archive into more than one content type. Table 2 shows the number of each content type. We performed a twofold cross-validation test. We used TinySVM (http://chasen.org/~taku/software/TinySVM/) software as the machine-learning package, and precision and recall as evaluation measures. Here, precision is the fraction of classified instances that are relevant, and precision is the fraction of relevant instances that are classified. We considered that precision is more important than recall because there are a large number of blog entries and QA archives on the web, and low recall would not become a serious matter in our case. In addition, misclassification of content-type causes wrong alignment of blog entries and QA archives.

Table 2. The number of pages of guidebooks, blog entries, and QA archives for each content-type

Content Type	Guidebook	Blog	QA Archive
Watch	102	395	620
Experience	78	241	412
Buy	418	163	191
Dine	741	382	502
Stay	278	134	257
Other	975	193	7888

Experimental Methods

To investigate the effectiveness of our method, we classified content types using the following four methods.

- The IG method: Used cue words collected by IG as features for machine learning.
- The IG+BoVW method: Used cue words collected by IG and BoVW as features for machine learning for guidebooks.
- BoVW method: Used BoVW as features for machine learning.
- Baseline method: Used all words that appeared in datasets.

Table 3. Evaluation results for content-type classification of pages of guidebooks

Feature	Measure	Watch	Experience	Buy	Dine	Stay	Average
Word (Baseline)	Precision	46.0	16.9	25.1	41.7	39.1	46.9
	Recall	53.6	15.4	23.3	49.2	17.0	30.9
IG	Precision	73.3	91.7	81.5	80.5	74.0	75.6
	Recall	32.1	17.0	20.0	32.4	32.8	27.9
IG+BoVW	Precision	74.1	91.7	76.2	77.6	75.4	**75.8**
	Recall	37.3	17.0	28.7	35.3	36.8	**33.7**
BoVW	Precision	61.9	—	—	72.0	—	—
	Recall	14.7	—	—	5.6	—	—

Table 4. Evaluation results for content-type classification of blog entries

Feature	Measure	Watch	Experience	Buy	Dine	Stay	Average
Word (Baseline)	Precision	68.4	55.3	50.7	75.1	49.8	**66.4**
	Recall	60.6	37.0	20.8	67.4	19.2	48.7
IG	Precision	66.7	60.2	54.9	77.2	58.9	65.9
	Recall	64.0	33.7	31.8	69.9	34.3	**51.0**

Table 5. Evaluation results for content-type classification of QA archives

Feature	Measure	Watch	Experience	Buy	Dine	Stay	Average
Word (Baseline)	Precision	72.7	56.6	77.5	72.2	82.0	70.9
	Recall	70.6	43.3	41.7	69.6	65.4	**61.1**
IG	Precision	71.1	81.4	90.1	71.8	90.8	**80.7**
	Recall	44.4	11.9	33.3	47.5	20.4	32.4

Results and Discussion

The evaluation results for the content-type classification of pages of guidebooks, blog entries, and QA archives are shown in Tables 3, 4, and 5. In Table 3, "-" indicates that we could not conduct machine learning because of the lack of training data. As shown in Tables 3, 4, and 5, the IG method obtained high precision in each dataset. However, the average precision score of blog entries was smaller than that of guidebooks and QA archives. Blog entries use informal expressions more frequently than guidebooks and QA archives. Therefore, we considered that the method based on cue words could not obtain high precision. To investigate the validity of this guess, we calculated the unique word ratio in guidebooks, blog entries, and QA archives, using the following.

$$\text{Unique word Ratio} = \frac{\text{the number of unique words in the content}}{\text{the total number of words in the content}}$$

The unique word ratio of guidebooks, blog entries, and QA archives was 0.082, 0.112, and 0.078, respectively. The unique word ratio of blog entries was higher than that unique word ratio of guidebooks or QA archives. It is considered that words that seldom appear in training data appear frequently in test data, if the unique word ratio is high. In addition, the number of blog entries that were classified manually into "buy" and "stay" is too small to train. That is, the IG method cannot cyclopedically collect cue words, and improve the precision of content-type classification in blog entries.

5.2 Alignment of Blog Entries and QA Archives with Guidebooks

Datasets and Experimental Setting

We used 90 guidebooks, 1,000 blog entries, and 1,998 QA archives. A guidebook was aligned manually with blog entries and QA archives that were thought to be related to the guidebook. We used precision and recall as evaluation measures.

Experimental Methods

To investigate the effectiveness of our methods, we conducted tests using the following. The KNN_TYPE method was only used for results of content-type classification. This experiment used results of the content type that were obtained by the IG+BoVW method for the guidebook page and the IG method for blog entries and QA archives. Except for the KNN_TYPE method, the methods did not use the results of content-type classification.

- KNN_TYPE: Calculated the similarity by KNN using the appearance frequency of location names. Location names were extracted from datasets of travel guidebooks, travel blog entries, and QA archives having the same content type.
- KNN_LOC: Calculated the similarity by KNN using the appearance frequency of location names. Location names were extracted from the datasets regardless of the content type.
- BASE_KNN: Calculated the similarity by KNN using the appearance frequency of nouns extracted from the datasets.
- BASE_SVM: Used SVM. Features were the appearance frequency of nouns extracted from the datasets.

Results and Discussion

The evaluation results are shown in Tables 7 and 8. As can be seen from these tables, the KNN methods were more useful than the SVM methods. The KNN_TYPE method obtained the best performance. It had high precision but low recall. However, the low recall was not a serious matter in this case, because the KNN_TYPE method could align 99 blog entries and 1,561 QA archives for a guidebook page, when applying our method to real datasets. We considered that precision was more important than recall, because enriching guidebooks having appropriate information was more beneficial than enriching guidebooks having incorrect information. Our methods allow the flexible alignment of appropriate information with guidebooks by using location names and results of content-type classification. Therefore, it is possible to align with blog entries and QA archives, which have various information if the recall will low.

The main cause of the failure was that location names, which were not the destination of the trip, were extracted from blog entries and QA archives. This was because of the use of CaboCha, which is the Japanese syntactic parser, for extracting location names from datasets. These web contents contain location names before moving, such as the traveller's home.

Table 7. Evaluation results for an automatically aligned blog entry for a guidebook

		Precision	**Recall**
Our methods	KNN_TYPE	**81.1**	**20.4**
	KNN_LOC	76.7	20.1
Baseline methods	BASE_KNN	27.3	15.5
	BASE_SVM	21.6	3.0

Table 8. Evaluation results for an automatically aligned QA archive for a guidebook

		Precision	Recall
Our methods	KNN_TYPE	**85.8**	**21.0**
	KNN_LOC	78.3	20.6
Baseline methods	BASE_KNN	48.7	16.6
	BASE_SVM	40.5	18.4

5.3 Alignment of Blog Entries and QA Archives with a Guidebook Page

Datasets and Experimental Setting

We performed experiments that aligned 100 blog entries and 100 QA archives with 90 guidebooks.

Experimental Methods

To investigate the effectiveness of our method, we conducted tests using the following methods. Our methods selected guidebooks that were aligned with blog entries and QA archives by Step 2. In this experiment, we used guidebooks that were aligned manually with blog entries and QA archives by the KNN_TYPE method of Section 5.2.

- Our method 1 (coarse-to-fine, without content type): We selected guidebooks that were aligned with blog entries and QA archives by the KNN_TYPE method described in Section 5.2. We calculated cosine similarity between selected guidebook pages and blog entries or QA archives, and aligned them with the guidebook page having the highest similarity.
- Our method 2 (coarse-to-fine, with content type): We selected guidebooks that were aligned with blog entries and QA archives by the KNN_TYPE method. We calculated cosine similarity between the selected guidebook pages and blog entries of the same content type as the guidebook pages or QA archives, and aligned them with the guidebook page having the highest similarity.
- Baseline method (not coarse-to-fine, without content type): Calculated the cosine similarity between the guidebook pages and blog entries or QA archives. Blog entries and QA archives were aligned with the guidebook page having the highest similarity.

Evaluation Methods

We used a questionnaire survey as the evaluation method. The questionnaire contents asked whether blog entries and QA archives that were aligned with the guidebook page were "Appropriate" or not for travellers collecting information. Each guidebook page was judged by 11 judges. The evaluation results for each guidebook page were decided by a majority vote.

Results and Discussion

The questionnaire results are shown in Table 9, and indicate that our method could appropriately align blog entries and QA archives with guidebook pages. For the results of blog entries, our methods 1 and 2 obtained better results than the baseline method. Therefore, our method that determines a guidebook-related blog entry first and then aligns the blog entry with the guidebook page could align more appropriately than the baseline method that aligns a blog entry with a guidebook page at one time. There was not much difference between our methods 1 and 2 because most blog

entries had been written about lots of things and had been classified into multiple content types.

In the results for QA archives, our method 2 gave the best performance. QA archives had posted questions focusing on specific content types such as "Could you recommend a restaurant in [location name]?" Therefore, the method that used the results of content-type classification was effective when QA archives were aligned with the guidebooks.

Table 9. Ration of enriched guidebook pages that were judged to be "Appropriate"

Method	Blog Entry (%)	QA Archive (%)
Baseline method	72	53
Our method 1	82	57
Our method 2	**82**	**77**

6 Conclusions

In this paper, we proposed a method for enriching guidebooks by aligning them with blog entries and QA archives. We conducted three experiments: (1) content-type classification of guidebook pages, blog entries, and QA archives, (2) alignment of blog entries and QA archives with guidebooks, and (3) alignment of blog entries and QA archives with a guidebook page. In the experiment (1), we found that our method using both textual and image information (IG+BoVW) obtained the best performance. In the experiment (2), our method KNN_TYPE obtained best precision scores in both aligning with blog entries and QA archives. In the experiment (3), judges considered 82.0% of blog entries and 77.0% of QA archives to be helpful in out method 2 (coarse-to-fine, with content type). With these steps, it was possible to align relevant information with pages of guidebooks and help travellers collect tourist information. We mentioned the method for enriching guidebooks by SNS, but we can easily apply this method to various other documents, such as travel leaflets or booklets.

References

Rakesh, A., Sreenivas, G., Anitha, K., and Kishnaram, K. 2011. Enriching Textbooks with Images. *Proceedings of the 20th ACM Conference on Information and Knowledge Management*. 1847-1856.

Nie, L., Wang, M., Gao, Y., Zha, Z.-J., and Chua, T.-S. 2013. Beyond Text QA: Multimedia Answer Generation by Harvesting Web Information. *IEEE Transactions on Multimedia*, 15, 2. 426-441.

Lu, X., Pang, Y., Hao, Q., and Zhang, L. 2009. Visualizing Textual Travelogue with Location-Relevant Images. *Proceedings of the 2009 International Workshop on Location Based Social Networks*. 65-68.

Bressan, M., Csurka, G., Hoppenot, Y., and Renders, J.M. 2008. Travel Blog Assistant System (TBAS) - An Example Scenario of How to Enrich Text with Images and Images with Text using Online Multimedia Repositories. *Proceedings of VISAPP Workshop on Metadata Mining for Image Understanding*.

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C. 2004. Visual Categorization with Bags of Keypoints. *Proceedings of ECCV International Workshop on Statistical Learning in Computer Vision*. 1-22.

He, Q., Pei, J., Kifer, D., Mitra, P., and Giles, C.L. 2010. Context-aware Citation Recommendation. *Proceedings of the 19th International WWW Conference*. 421-430.

Kando, N. 1997. Text-level Structure of Research Papers: Implications for Text-Based Information Processing Systems. *Proceedings of British Computer Society Annual Colloquium of Information Retrieval Research*. 68-81.

Nanba, H., Taguma, H., Ozaki, T., Kobayashi, D., Ishino, A., and Takezawa, T. 2009. Automatic Compilation of Travel Information from Automatically Identified Travel Blogs.

Proceedings of Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing. 205-208.