Construction of a System for Providing Travel Information along Hiroden Streetcar Lines

Aya Ishino Hidetsugu Nanba Toshiyuki Takezawa
Graduate School of Information Sciences, Hiroshima City University
Hiroshima, Japan
{ishino, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp

***Abstract*— We propose a method for collecting Hiroshima Electric Railway (Hiroden) blog entries in a blog database. Hiroden blog entries are defined as travel journals that provide regional information at Hiroden streetcar stations. The task of collecting Hiroden blog entries is divided into two steps: (1) collection of blog entries, and (2) identification of Hiroden blog entries. To investigate the effectiveness of our method, we conducted some experiments. From the experimental results, we obtained precision of 82.4%, recall of 64.5%, and F-measure of 72.4% in automatic identification of Hiroden blog entries. Finally, we have constructed a system for providing travel information along Hiroden streetcar lines.**

Keywords-component; blog; information extraction; travel information

I. INTRODUCTION

In Japan, the Basic Act for Promoting a Tourism-oriented Country became effective in January 2007. The Japan Tourism Agency was launched in October 2008, and aims to promote domestic and overseas tourism. At the moment, various events relating to NHK's period drama "Taira no Kiyomori" are being held in Hiroshima, Japan. From April 2013, the biggest confectionery exposition in Japan is scheduled to be held in Hiroshima (Hiroshima Kashihaku 2013). These are examples of the numerous efforts to increase the number of travelers who visit Hiroshima.

Our focus is on the Hiroshima Electric Railway, known as "Hiroden" for short. Its routes in Hiroshima are tourist resources. In this paper, we propose a method for collecting Hiroden blog entries in a blog database. Hiroden blog entries are defined as travel journals that provide information along Hiroden streetcar stations. In addition, we construct a system for providing Hiroden blog entries as travel information at Hiroden streetcar lines. We consider that the system can promote tourism using Hiroden and provide regional information not found in guidebooks.

Recently, several methods to identify bloggers' attributes such as residential area [1], gender, and age [2, 3], have been proposed. By combining this research with Hiroden blog entries, it is possible to recommend a local product that is popular among females, for example, or a travel spot, that young people often visit.

The remainder of this paper is organized as follows. Section II shows the system behavior in terms of snapshots. Section III describes related work. Section IV explains our methods. To investigate the effectiveness of our methods, we conducted some experiments, and Section V reports on these and the results. We present some conclusions in Section VI.

II. SYSTEM BEHAVIOR

In this section, we describe our prototype system. Fig. 1 is a snapshot of its behavior on the iPad. The system operates with a browser.

Figure 1. A snapshot from our system on the iPad.

Fig. 2 shows a route map of the system for providing travel information along Hiroden streetcar lines. The route map contains Hiroden streetcar stations and major tourist spots. These are the steps in the search procedure.

(Step 1) Click a part of the route map, such as "Kamiya-cho area" (shows as ① in Fig. 2), to show an enlarged route map (Fig. 3).
(Step 2) Click the Hiroden streetcar station, such as "Atomic Bomb Dome" (shown as ② in Fig. 3), to generate a list of links to Hiroden blog entries (Fig. 4).
(Step 3) Click the link to a Hiroden blog entry to display it.

In this paper, we propose a method for collecting the Hiroden blog entries that the system provides.

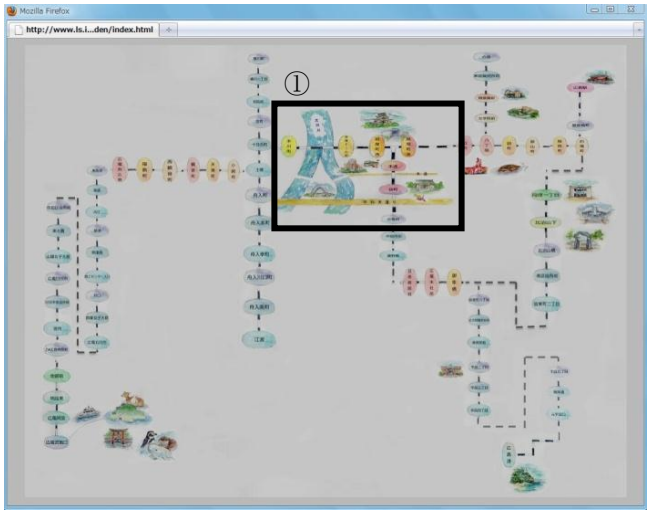

Figure 2. The opening page of our travel information system: a route map of the Hiroden system.

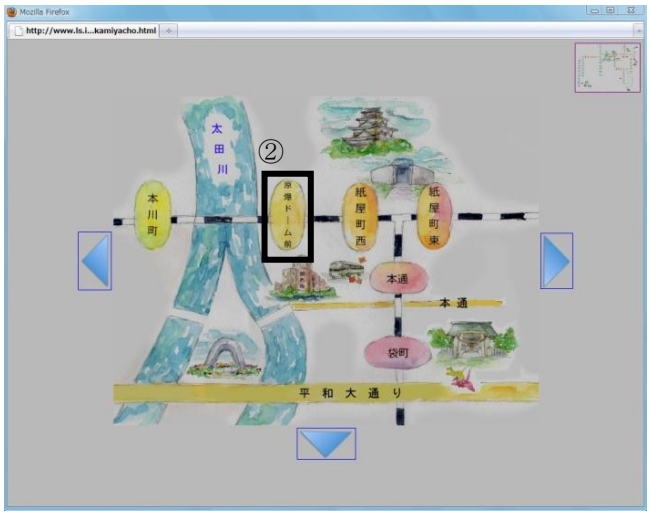

Figure 3. A portion of the route map, enlarged.

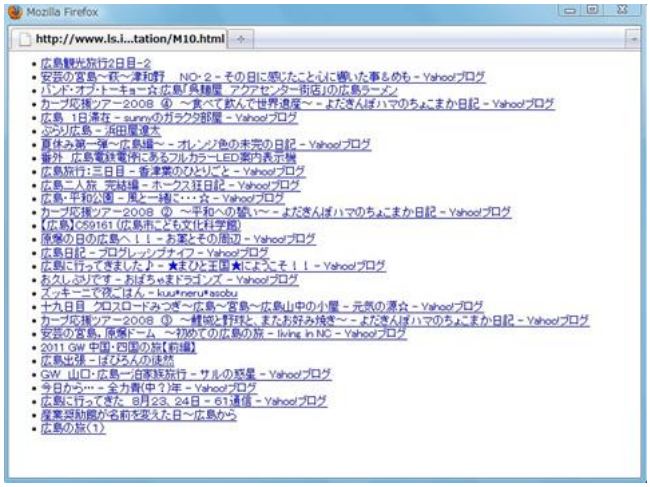

Figure 4. A list of links to Hiroden blog entries.

III. Related Work

In this section, we describe some related studies. GeoCLEF (http://ir.shef.ac.uk/geoclef/) is the cross-language geographic retrieval track run as part of the Cross Language Evaluation Forum (CLEF http://www.clef-initiative.eu/), and it has been operating since 2005 [4]. The goal of this task is to retrieve news articles relevant to particular aspects of geographic information, such as "wine regions around the rivers in Europe." In our work, we focus on Hiroden blog entries rather than news articles, because bloggers' travel experiences tend to be written as blog entries.

In this paper, we propose a method for collecting Hiroden blog entries as regional information. Some other studies have also proposed methods for collecting regional information from the Web. Otsuki and Sato [5] proposed a system that edits a Web directory of regional information automatically. They focused on regional Web sites provided by local governments and classified the pages in the regional web sites into eight categories. Okamoto *et al.* [6] proposed a method for extracting occasional or volatile local events from blogs using search queries. In this paper, we focus on blog entries.

A Hiroden blog entry is a travel blog entry. Both "www.travelblog.org" and "travel.blogmura.com" are portal sites for travel blogs. At these sites, travel blogs are manually registered by bloggers themselves, and the blog entries are classified by their destinations. However, there are many more Hiroden blog entries in the blogosphere. There are few travel blogs about a confined geographical area like Hiroden. Aiming to construct an exhaustive database of Hiroden blog entries, we have studied their automatic identification in the blogosphere. Nanba *et al.* [7] identified travel blog entries in a blog database. They formulated the identification of travel blog entries as a sequence-labeling problem, and solved it using machine learning. They extracted pairs comprising a location name and a local product from these entries and showed the effectiveness of travel blog entries as a source of travel information. In this paper, we propose a method for identifying Hiroden blog entries in a blog database.

There are some studies of extracting of transportation information. Davidov [8] presented an algorithm framework that enabled automated acquisition of map-link information from the Web, based on surface patterns such as "from X to Y." Given a set of locations as initial seeds, they retrieved an extended set of locations from the Web and produced a map-link network that connects these locations using transport-type edges. Ishino *et al.* [9] extracted transportation information automatically from travel blog entries via machine-learning techniques. By combining this research with our method, it is possible to recommend an optimum route using Hiroden.

IV. Collection Of Hiroden Blog Entries

In this section, we describe our method for collecting Hiroden blog entries. This task is divided into two steps: (1) collection of blog entries; and (2) identification of Hiroden blog entries. These steps are explained below.

A. Collection of Blog Entries

The blog entries to be collected should include information about the 78 Hiroden streetcar stations. We therefore searched blogs using Yahoo! search (blog) for each Hiroden streetcar station. As a result, we collected 1,748 blog entries.

B. Identification of Hiroden Blog Entries

In this paper, we define a Hiroden blog entry as one in which the author alighted at a Hiroden streetcar station and saw sights around the station. The blog entries collected in Section IV-A contain non-Hiroden blog entries, for example, blogs about other means of transportation or blogs whose authors did not see sights. Then, we propose a method for identifying Hiroden blog entries among the blog entries collected in Section IV-A.

Fig. 5 shows an example of a Hiroden blog entry. It contains the names of Hiroden streetcar stations, words related to Hiroden such as "広電" (Hiroden) or "市電" (streetcar in Hiroshima city), and words related to sightseeing such as "観光" (sightseeing) or "散策" (stroll). In addition, when blog authors see a sight, they tend to take pictures there and post them to their blog entries. Therefore, Hiroden blog entries contain pictures that authors took in seeing sights around a Hiroden streetcar station. We employed a machine-learning technique to identify Hiroden blog entries. We used the following features for machine learning. The feature words were selected manually.

- Whether the word is a name of a Hiroden streetcar station (78).
- Whether the word is related to Hiroden, such as "広電" (Hiroden) or "市電" (streetcar) (5).
- Whether the word related to other means of transportation, such as "新幹線" (Sinkansen bullet train) or "フェリー" (ferry) (6).
- Whether the word related to sightseeing, such as "観光" (sightseeing) or "散策" (stroll) (9).
- The number of pictures.

[original]
ココに車を停めて後の移動は市内電車**広電**
市電と宮島行の船が一日中乗り放題の一日乗車乗船券を買ってまずは**原爆ドーム前**へ
周辺のテキヤを満喫し**広電**で宮島口まで移動

[translation]
I parked here. After this, I used **Hiroden** for transportation.
I got a 1-Day Trip Card for enjoying **streetcar**s and a ferryboat to Miyajima and went to **Atomic Bomb Dome Station**.
I visited the adjacent stands.
Next, I went to Hiroden-miyajima-gichi Station by **Hiroden**.

Figure 5. An example of a Hiroden blog entry.

V. EXPERIMENTS

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

A. Data Sets and Experimental Setting

To generate the test data for identifying Hiroden blog entries, we used the 1,748 blog entries collected in Section IV-A. Then, we manually identified Hiroden blog entries among 1,748 blog entries. The number of manually identified Hiroden blog entries is shown in Table I.

TABLE I. NUMBER OF MANUALLY IDENTIFIED HIRODEN BLOG ENTRIES

Hiroden Blog Entries	Other	Total
568	1,180	1,748

B. Machine Learning and Evaluation

We performed a four-fold cross validation test. We used Tiny SVM (http://chasen.org/~taku/software/TinySVM/) software as the machine-learning package and used precision, recall and F-measure as evaluation measures, calculated as follows:

$$Precision = \frac{The\ number\ of\ Hiroden\ blog\ entries\ that\ the\ system\ could\ detect\ correctly}{The\ number\ of\ Hiroden\ blog\ entries\ that\ the\ system\ detected} \quad (1)$$

$$Recall = \frac{The\ number\ of\ Hiroden\ blog\ entries\ that\ the\ system\ could\ detect\ correctly}{The\ number\ of\ Hiroden\ blog\ entries\ that\ should\ be\ detected} \quad (2)$$

$$F\text{-}measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (3)$$

To confirm our method, we also examined a baseline method that identified all 1,748 blog entries as Hiroden blog entries.

C. Results and Discussion

Table II shows the experimental results. The recall score by our method was 0.355 lower than that of the baseline method. However, our method improved the precision and F-measure scores of the baseline method by 0.499 and 0.233, respectively. We have thus confirmed the effectiveness of our method for identifying Hiroden blog entries.

TABLE II. IDENTIFICATION OF HIRODEN BLOG ENTRIES

	Precision (%)	Recall (%)	F-measure (%)
Our method	82.4	64.5	72.4
Baseline method	32.5	100.0	49.1

We now discuss the low recall of our method. A typical reason for low recall is the lack of cues. For the machine learning, we used manually selected cues, as described in Section IV-B. Most cues are words related to Hiroden, such as "広電" (Hiroden) or "市電" (streetcar). However, various cues were needed for collecting more Hiroden blog entries, for example, the names of the restaurants, tourist spots, and souvenir shops that often appear in Hiroden blog entries. To improve the coverage of cues, a statistical approach, such as applying n-gram statistics to a larger blog corpus, will be required.

VI. CONCLUSION

In this paper, we propose a method for collecting Hiroden blog entries from a blog database. The task is divided into two steps: (1) collection of blog entries, and (2) identification of Hiroden blog entries. To investigate the effectiveness of our method, we conducted some experiments. From the experimental results, we obtained precision of 82.4%, recall of 64.5% and F-measure of 72.4% in automatic identification of Hiroden blog entries. Finally, we have constructed a system for providing travel information along Hiroden streetcar lines.

In our future work, we will classify Hiroden blog entries into some categories of travel, such as "sightseeing," "eating," or "staying." Users can then check Hiroden blog entries more efficiently.

At present, our system considers only Hiroden blog entries as travel information. In our future work, we will collect Yahoo! Answers (http://answers.yahoo.com/) and news Websites related to Hiroden. Yahoo! Answers (formerly known as Yahoo! Q&A) is a community-driven question-and-answer (Q&A) site. Such information is useful for travelers. Our system will provide rich information.

REFERENCES

[1] N. Yasuda, T. Hirao, J. Suzuki, and H. Isozaki, "Identifying Bloggers' Residential Areas," Proc. AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, pp.231-236, 2006.

[2] D. Ikeda, H. Takamura, and M. Okumura, "Semi-supervised Learning for Blog Classification," Proc. the 23rd AAAI Conference on Artificial Intelligence, pp.1156-1161, 2008.

[3] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker, "Effects of Age and Gender on Blogging," Proc. AAAI Symposium on Computational Approaches for Analyzing Weblogs, pp.199-205, 2006.

[4] F. Gey, F, R. Larson, M. Sanderson, H. Joho, P. Clough, and V. Petras, "GeoCLEF: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview," *Lecture Notes in Computer Science*, LNCS4022, pp.908-919, 2005.

[5] Y. Otsuki, and S. Sato, "Automated Editing of a Web Directory of Regional Information", J. Transactions of Information Processing Society of Japan, Vol. 42, pp.2310-2318, 2001.

[6] M. Okamoto, and K. Masaaki, "Discovering Volatile Events in Your Neighborhood: Local-Area Topic Extraction from Blog Entries," Proc. Asia Information Retrieval Symposium, Lecture Notes in Computer Science 5839, pp.181-192, 2009.

[7] H. Nanba, H. Taguma, T. Ozaki, D. Kobayashi, A. Ishino, and T. Takezawa, "Automatic Compilation of Travel Information from Automatically Identified Travel Blogs," Proc. 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, pp.205-208, 2009.

[8] D. Davidov, "Geo-mining: Discovery of Road and Transport Networks Using Directional Patterns," Proc. the 2009 Conference on Empirical Methods in Natural Language Processing: pp.267-275, 2009.

[9] A. Ishino, H. Nanba, and T. Takezawa, "Automatic Compilation of an Online Travel Portal from Automatically Extracted Travel Blog Entries," Proc. the 18th international Conference on Information Technology and Travel & Tourism, 2011.