Event Extraction and Visualization for Obtaining Personal Experiences from Blogs

Yoko Nishihara[1], Keita Sato[2], and Wataru Sunayama[2]

[1] The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo
nishihara@sys.t.u-tokyo.ac.jp
[2] Hiroshima City University, 3-4-1 Ozuka-Higashi, Asa-Minami-ku, Hiroshima
{keita,sunayama}@sys.im.hiroshima-cu.ac.jp

Abstract. Internet users write blogs related to their personal experience, daily news, and so on. Though we can obtain blogs about personal experience using search engines on the Web, the search engines also output blogs about other topics unrelated to personal experiences. Therefore, we need to take too much time to read all blogs for obtaining those about personal experiences. This paper proposes a support system for obtaining blogs about personal experience efficiently. The system extracts three keywords that denote *place, object*, and *action* from a blog. The three keywords describe an event that leads a person to write a blog about personal experience. The system expresses the event with three pictures related to the extracted keywords. The pictures help users to judge whether personal experiences are written in the blog or not. We experimented with the system, and verified that it supports users to obtain personal experiences efficiently.

Keywords: personal experience, pictures expressing an event, place keyword, object keyword, action keyword.

1 Introduction

Since blogs have been written by many people, we can obtain personal experiences and/or reviews of commercial items from blogs. For example, if a man/woman need information about local area in Hokkaido, he/she can obtain such information by entering queries *Hokkaido AND sight seeing* into a search engine and reading Web pages of search results. The search results also contain other information, plans for sight seeing in Hokkaido, sight seeing in some areas near Hokkaido and so on, except those about personal experience in Hokkaido. Therefore, he/she has to spend much time to read all of the blogs. It causes the low efficiency of information acquisition.

This paper proposes a support system for obtaining information about personal experience from blogs. We consider that people obtain personal experiences through some events that they have experienced. Therefore, the proposed system uses the events for extracting personal experiences. The proposed system extracts some keywords denoting an event, and visualizes the event using pictures expressing the extracted keywords. Users can judge whether personal experiences are written in blogs or not by watching the pictures.

M.J. Smith and G. Salvendy (Eds.): Human Interface, Part II, HCII 2009, LNCS 5618, pp. 315–324, 2009.

Visualizing events using pictures makes users not to read all of blogs, and the amount of time to obtain personal experiences is reduced [6, 10]. The possibility that events and personal experiences are written together in a blog is usually high (In the experiment for the proposed system, the possibility was 93%). Therefore, we can support users obtaining personal experiences by extracting events.

We define an event as three keywords, *an action keyword, an object keyword,* and *a place keyword*. Other keywords, such as time keywords, subject keywords, and reason keywords, are generally used to visualize events. In case of time keywords, though some methods have been proposed [9], we do not use time keywords because we need a large amount of training data for machine learning to output the time keywords. In case of subject keywords, the subject of personal experience in a blog is always the writer of the blog. Therefore, we do not extract subject keywords from blogs. And we do not extract reason keywords because of the difficulty of extraction.

2 Related Work

2.1 Information Extraction from Web

There are many studies about information extraction from the Web [2]. In case of extracting information that is interesting to the public, there has been a method to extract human names, keywords, sentences attracted by the public from blogs [5]. There have been some methods for extracting noticed persons and noticed events [12, 13, 16]. Another method extracts keywords that will be attracted in the future using time series analysis of keyword frequencies [4, 8]. Though the proposed system also extracts information from the Web, the proposed system extracts information about personal experience instead of information attracted by the public.

2.2 Review Extraction

It is considered that personal experience is one of reviews. Some methods have been proposed to extract reviews of commercial items and movies from the Web. Some methods learn features of the reviews by machine learning [3, 11]. The methods proposed in [14, 15, 17] extract and show the reviews of commercial items using learned features. The proposed system extracts events leading persons to write blogs about personal experience, not extracts personal experiences themselves.

Events are different from commercial items and movies. Since each people feel the same event in each way, it is considered that the number of keywords for writing personal experiences is higher than the number of keywords for writing events [18]. Events and personal experiences appear together with high frequency in blogs. Therefore, the proposed system extracts events and supports users for obtaining personal experiences.

2.3 Support for Obtaining Personal Experience

The method proposed in [18] extracts personal experiences from blogs semi-automatically, and shows thumbnails of the blogs to users. The proposed system also shows pictures corresponding to events. Therefore, the proposed system is different from the method of [18] in showing pictures.

3 Proposed System

The system takes queries as input. The queries represent a theme that blogs are written based on. For example, if he/she need to obtain personal experiences about sight seeing in Hawaii, he/she should input *Hawaii AND sight seeing* into a search engine. The system downloads blogs including the queries and narrows the downloaded blogs to blogs including sentences about events. Since sentences are usually written as past tense to represent events, the system extracts blogs including sentences of past tense for narrowing the blogs. Next, the system separates blog texts every sentences including *place keywords*. The system extracts keywords (*object keywords* and *action keywords*) representing events from separated the texts. After the system sets out pictures representing the extracted keywords, the system outputs the set of the pictures.

3.1 Blog Text Separation

We explain how to separate blog texts every sentences including place keywords. It is considered that a blog has some descriptions about different places. Therefore, the system separates blog texts using sentences including place keywords, and looks on the separated text as a *block*.

For extracting place keywords, the system firstly extracts noun keywords following prepositions such as *at, on, in* representing places. The system extracts such nouns as candidates of place keywords, and decides one noun keyword as a place keyword.

3.2 Extraction of Keywords Corresponding to Event

The system extracts three keywords from a block. The keywords are a place keyword, an object keyword, and an action keyword. In the following section, we explain how to extract object keywords and action keywords.

3.2.1 Extraction of Object Keyword

We explain how to extract object keywords. The object keywords are noun keywords. If some noun keywords are in a block, the degree of relationship between each noun keyword and the extracted place keyword is evaluated the following equation,

$$relation(p,o) = \frac{hit(p \wedge o)}{hit(p)} \frac{hit(p \wedge o)}{hit(o)} \tag{1}$$

where p denotes a place keyword, and o denotes an object keyword. Eq.(1) calculates a rate of the number of Web pages including the place keyword to the number of Web pages including the object keyword. If the value of Eq.(1) is high, it is considered that two keywords are related to each other. The system decides a noun keyword with the highest value of Eq.(1) as an object keyword.

3.2.2 Extraction of Action Keyword

We explain how to extract action keywords. Action keywords are verb keywords that appear in the end of sentences including the extracted object keywords. This is because an action keyword and an object keyword appear in the same sentence.

3.3 Setting Out of Pictures for Visualizing Event

The system sets out of pictures representing the extracted keywords. The pictures are aligned crossly, a place picture, an object picture, and an action picture from the left to the right. If a blog has some blocks, pictures of the first block is visualized only.

We prepared a database of pictures. The database has been created using pictures of an image search engine [7]. We input keywords extracted blogs as queries into the search engine, and chose one pictures from top of 20 pictures of the search results for each keyword. We spent from two seconds to 30 seconds for choosing one picture. Now, the database has about 1,000 pictures for place keywords, about 700 pictures for object keywords, and about 200 pictures for action keywords. If the database does not have pictures corresponding to the extracted keywords, the system visualizes blanks instead of the pictures.

3.4 Output: Three Pictures Visualizing Event

The system outputs a set of pictures visualizing events (shown in Fig. 1). If titles and summaries of blogs have been obtained in downloading blogs, the system also outputs those with the set of pictures.

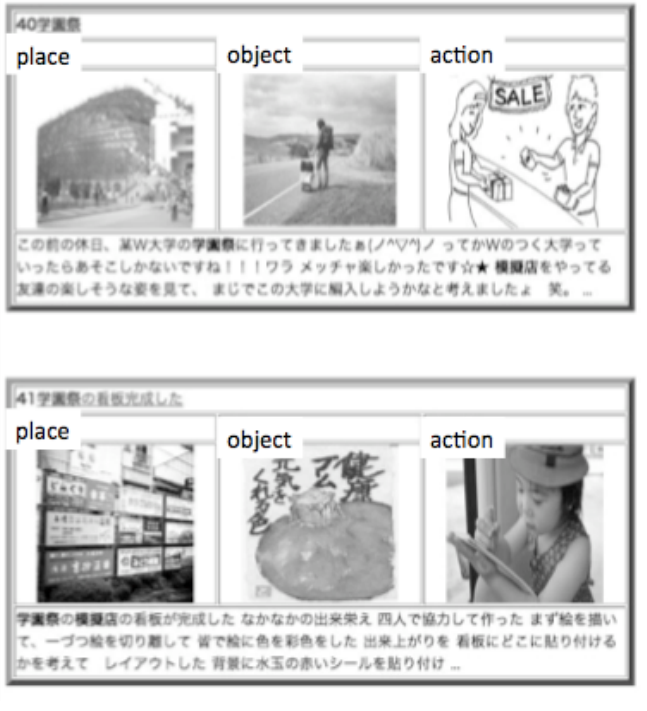

Fig. 1. Output of proposed system. A set of blogs with pictures.

4 Experiment for Proposed System

In the experiment, we asked participants to extract texts written about personal experience from blogs using the proposed system. We collected blogs from a blog site, Yahoo! Blog [1]. We used top 100 blogs of the search results by entering queries for this experiment. The queries for blog collection are shown in Table 1. Five of the queries are about sight seeing, and one of the queries is about school festival. We chose those queries because it is considered that, in obtaining personal experiences, most of people search something about events that they will also experience in the near future.

Table 1. Queries for experiment of proposed system

Okinawa, Tokyo, Hiroshima, Nigata, Hokkaido	AND sight seeing
School festival	AND shop for eating

We prepared the other system (*baseline system*) that shows titles and summaries of blogs as shown in the proposed system. The same texts of blogs as shown in the proposed system can be read in the baseline system. The 100 blogs were separated into four subsets including 25 blogs and shown using a Web browser. The size of a window of a Web browser was 1,200 x 1,920 pixels.

We instructed the participants as follows:

1. (Proposed) Watch the set of pictures for each blog. / (Baseline) Read the summaries for each blog.
2. Judge whether events related the queries are written or not.
3. If you find blogs that events are written in, copy the blog texts about personal experience and paste them in a text editor.

The number of participants was 36 undergraduate/graduate students majoring information science. We assigned 18 participants to one query and one system. The time of one session was five minutes. This is because it spends about five minutes to search something. We compared the averages of personal experiences extracted using the prepared systems.

4.1 Experimental Results

Table 2 shows averages of blogs read by the participants. For all of the queries, the averages for the proposed system were higher than those for the baseline system ($P<.05$). This is because the time to understand the contents by watching pictures is shorter than the time to understand the contents by reading summaries. The result indicates that the proposed system supports users for reading more blog texts.

Table 3 shows averages of blogs from which personal experiences were extracted. The extracted texts were proper to personal experiences. The averages of blogs for the proposed system were higher than those for the baseline system ($P<.05$). This is because events were visualized by the pictures output by the proposed system, and

Table 2. Averages of blogs read by participants

	Okinawa	Tokyo	Hiroshima	Nigata	Hokkaido	School festival
Proposed	5.8	4.9	5.1	4.7	5.9	5.4
Baseline	4.9	4.7	4.3	4.6	4.7	4.7

Table 3. Averages of blogs from which personal experiences are extracted by participants

	Okinawa	Tokyo	Hiroshima	Nigata	Hokkaido	School festival
Proposed	2.2	3.1	2.9	1.8	3.2	3.4
Baseline	1.4	2.6	2.6	1.3	2.5	2.1

Table 4. Rate of blogs read by participants to blogs from which personal experiences were extracted by participants

	Okinawa	Tokyo	Hiroshima	Nigata	Hokkaido	School festival
Proposed	0.38	0.63	0.57	0.38	0.54	0.63
Baseline	0.29	0.55	0.60	0.28	0.53	0.45

most of the blogs chosen by the participants had texts about personal experience. This is also appeared in Table 4. In Table 4, except *Hiroshima*, the rates of read blogs to extracted blogs for the proposed system were higher than those for the baseline system ($P<.05$). In case of *Hiroshima*, since many pictures corresponding to place keywords were not visualized in the proposed system, it was difficult to judge whether events were written in each blog or not. However, in case of the other queries, the rates for the proposed system were higher than those for the baseline system. Therefore, we confirmed that the proposed system is more efficiently to obtain personal experiences.

4.2 Efficiency of Pictures for Choosing Blogs

In summaries of the used blogs for the experiment, some of them have descriptions about events (for example, *I went to a park*) and others do not have such descriptions. On the other hand, in texts of the used blogs for the experiment, some of them have descriptions about personal experience and others do not have such descriptions. Therefore, we divided the used blogs into four patterns by using the above two features. The results are shown in Table 5. Though 61% blogs for all of them (sum of pattern (1) and pattern (2)) can be judged using the baseline system, 39% blogs for all of them (sum of pattern (3) and pattern (4)) can not be judged using the baseline system. In case of pattern (3), users of the baseline system do not read blogs including personal experiences. In case of pattern (4), users of the baseline system read the blogs not including personal experiences. However, in case of pattern (3), the proposed system outputs some pictures like Fig. 2, therefore, the users can judge that personal experiences are written in the blog though it is not written in the summary. In case of pattern (4), the proposed system outputs some pictures like Fig. 3, therefore, the users can judge that personal experiences are not written in the blog.

Table 5. Number of blogs divided by two features. One feature is whether events are written in summaries or not. Another feature is whether personal experiences are written in blogs or not.

Pattern Event : Personal Experience	(1) O : O	(2) X : X	(3) X : O	(4) O : X
Okinawa	31	30	20	19
Tokyo	15	60	20	5
Hiroshima	29	28	21	22
Nigata	11	61	8	20
Hokkaido	20	26	28	26
School festival	35	20	33	12
Average	23.5	37.5	21.6	17.3
Rate to all of the blogs	61%		39%	

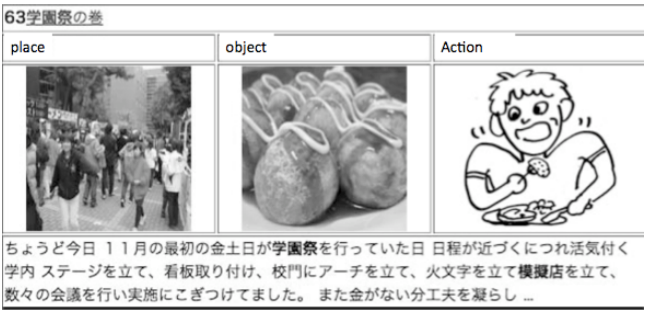

Fig. 2. Example of a blog. Events were not written in the summary, but personal experiences were written in the blog. The queries were *school festival AND shop for eating*.

Fig. 3. Example of blog downloaded using queries *Hokkaido AND sight seeing*. An event is written in the summary, but personal experiences are not written in the blog.

Table 6 shows the number of blogs divided two features: whether pictures visualize events or not, and whether personal experiences are included in blogs or not. The blogs divided into pattern (3) and pattern (4) in Table. 5 are corresponded to pattern (c) and pattern (d) in Table 6. The sum of pattern (c) and pattern (d) was 22.2%. The sum was lower than the sum of pattern (3) and pattern (4) (39.0%). It is considered that the participants can obtain more texts about personal experience by watching the pictures.

Table 6. Number of blogs divided with two features. The first feature is whether the pictures visualize events or not. The second feature is whether personal experiences are written in blogs or not.

Pattern	(a)	(b)	(c)	(d)
Event pictures: Personal experience	O : O	X : X	X : O	O : X
Okinawa	40	42	5	13
Tokyo	28	52	7	13
Hiroshima	52	35	7	16
Nigata	9	70	7	14
Hokkaido	39	41	5	15
School festival	41	28	4	27
Average	34.8	43.0	5.8	16.3
Rate to all of blogs	77.8%		22.2%	

Table 7. Number of blogs including personal experiences

School festival	Okinawa	Hiroshima	Hokkaido	Tokyo	Nigata	Average
68	51	50	48	35	19	45.1

Even when the number of the blogs including personal experiences was low, the participants for the proposed system extracted texts about personal experience efficiently. Table 7 shows the number of blogs including texts about personal experience. In case of *Nigata*, the number of blogs including personal experiences was 19. Therefore, the rate of extracted blogs to read blogs was low using the baseline system (the rate was 0.28 shown in Table 4). However, using the proposed system, the rate was 0.38, and bigger than that of the baseline system ($P<.05$). This is because the pictures of the proposed system support users for judging whether events are written in blogs or not. The results indicate that the proposed system can extract texts about personal experience even if the number of personal experiences is low.

4.3 Overview Efficient by Event Visualizing Pictures

Users of the proposed system watched the outputs in different way from users of the baseline system. In questionnaire to 36 participants of the proposed system, 22 participants answered they watched the whole of outputs, four participants answered they watched the outputs sequentially, and 10 participants answered they watched both of them, the whole of outputs and watch the outputs sequentially. This is because users of the proposed system can understand the contents of blogs at a glance. On the other hand, all of the users of the baseline system answered they watched the outputs sequentially. This is because users of the baseline system can not understand the contents of blogs at a glance. The result indicates that users of the proposed system tend to watch the whole of outputs.

The proposed system did not output all of pictures corresponding to extracted keywords. 17 pictures for place keywords, 12 pictures for object keywords, and 9 pictures for action keywords were not output by the proposed system. The number of blogs without pictures was 81. If there are blanks in the output, the system should

Table 8. Number of participants using combination of pictures

Place	Object	Action	Number of participants
O	X	X	19
O	X	O	6
O	O	X	4
O	O	O	4
X	X	O	2
X	O	O	1
X	O	X	0

output the extracted keywords only. However, it will cause the low efficiency of obtaining personal experiences with two reasons. The first reason is that users can not watch the whole of outputs and can not quickly judge whether texts about personal experiences are written or not if the extracted keywords and pictures are shown together in a Web browser. The second reason is that the time to understand the contents of the extracted keywords is longer than the time to understand the contents of pictures [6, 10]. Therefore, it is considered that it costs much time by showing the extracted keywords instead of pictures. On using pictures for the system output, if the contents are not described simply in pictures, it takes much time to understand the contents of them. Though some of the participants answered that it took much time to understand the contents of the pictures, they pointed that to a part of the pictures, not all of the pictures. The result indicates that showing the pictures supports users for obtaining texts about personal experience.

4.4 Efficiency of Pictures Visualizing Place Keyword

We asked the participants of the proposed system which combination of pictures they used for extracting texts about personal experience. Table 8 shows the result. The number of the participants who used pictures corresponding to place keywords was the highest. It is considered that when users judge whether events are written in blogs or not, pictures of place keywords (e.g. mountain, sea, and river) are more useful than pictures of action keywords (e.g. walk, run, and watch) and pictures of object keywords (e.g. book, bicycle, and dish). Therefore, most of the participants used pictures of place keywords. The result indicates that pictures of place keywords are most useful in obtaining personal experiences.

In Table 8, though the number of the participants using pictures of place keywords was the highest (19 participants), 17 participants used another combinations of pictures. It is considered that pictures of object keywords and pictures of action keywords are also useful for obtaining personal experiences. Therefore, it is necessary to visualize the three pictures to obtain personal experiences.

5 Conclusion

This paper has proposed a support system for obtaining personal experiences from blogs using event visualizing pictures. The system visualizes an event with three pictures,

place, object, and *action*. Users judge whether texts about personal experience are written in blog or not by watching the output pictures. Experimental results showed that the proposed system can support users efficiently to obtaining personal experiences.

References

1. Blog search engine, http://blogs.yahoo.co.jp/
2. Chang, C.H., Kayed, M., Girgis, M.R., Shaalan, K.: A Survey of Web Information Extraction Systems. IEEE Transactions on Knowledge and Data Engineering 18(10), 1411–1428 (2006)
3. Dave, K., Lawrence, S., Pennock, D.M.: Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In: Proc. of the 12th International World Wide Web Conference, pp. 519–528 (2003)
4. Fujiki, T., Nanno, T., Suzuki, Y., Okumura, M.: Identification of Bursts in a Document Stream. In: Workshop on Knowledge Discovery in Data Streams (2004)
5. Glance, N.S., Hurst, M., Tomokiyo, T.: BlogPulse: Automated Trend Discovery for Weblogs. In: Proc. of the WWW 2004 Workshop on the Weblogging Ecosystem (2004)
6. Hulbert, S., Beers, J., Fowler, P.: Motorists' Understanding of Traffic Control Devices. AAA Foundation for Traffic Safety (1979)
7. Image search engine, http://search.yahoo.co.jp/images
8. Kleinberg, J.: Bursty and Hierarchical Structure in Streams. In: Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1–25 (2002)
9. Noro, T., Inui, T., Takamura, H., Okumura, M.: Time Period Identification of Events in Text. In: Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006), pp. 1153–1160 (2006)
10. Pietrucha, M., Knoblauch, R.: Motorists' Comprehension of Regulatory, Warning and Symbol Signs, vol.2, Technical Report Contract DTFH61-83-C-00136, FHWA, U.S. Department of Transportation (1985)
11. Turney, P.D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 417–424 (2002)
12. Web page, http://search.biglobe.ne.jp/ranking/
13. Web page, http://blog360.jp/
14. Web page, http://blogsphere.biz/
15. Web page, http://opinion.labs.goo.ne.jp/cgi-bin/index.cgi
16. Web page, http://kizasi.jp/
17. Web page, http://shopping.nifty.com/
18. Web page, http://shooti.jp