Proposal of Time-crawler which collects an event time by reading Exif data in blogs

Ismail Arai, Kazutoshi Fujikawa and Hideki Sunahara
Graduate School of Information Science, Nara Institute of Science and Technology, Nara 630-0192, Japan
ismail-a@is.naist.jp, fujikawa@itc.naist.jp, suna@wide.ad.jp

Abstract— **To enable an outdoor user to acquire mobile specific information, an information retrieval system should provide contents that reflect user's circumstance such as user's position and time. There are a lot of information services that can provide contents based on geographical location. But, about temporal information, there is few works, because there is no useful indexing method for time information. In this paper, we propose a crawler which collects temporal expressions by reading Exif data of photos on blogs. Because the Exif data has shooting time of photo, the proposed system is expected to have good precision with small granularity. As a result of evaluation, the precision of the proposed method reached 0.72 even though the proposed method adopts a simple algorithm.**

I. Introduction

With the progress of mobile computers, users want mobile specific information such as restaurant information near their position at that time, weather forecast of business destination, tourist information, and so on. In some portal sites (Ex. The weather channel[13] and Zagat Survey[16], etc...), users can search mobile specific information by clicking their menu buttons. But, users want to search their information more rapidly and simply. To solve this problem, a desired system must provide users information which reflects their scheduled/current time and position. It means the information service should have positional and temporal index of contents.

Meanwhile, personal web sites, especially, blogs are increasing. Users expect to be able to get suitable mobile specific information from them. Conventional sites, especially commercial sites cannot have aggressive or negative comments (Ex. "a food in the restaurant A is bad taste." or "the waiter behaved poorly."). Users cannot judge whether a content suits their tastes or not by watching middle-of-the-road comments. Against that, users can judge contents by watching writing styles of blogs more easily. Also, a lot of bloggers writes their feelings straightforwardly. When a user who prefers strongly-flavored tomato sauces watches a description such as "I went to Pizza-House-A. The tomato sauce of the pizza tastes nice with strongly-flavored unlike that of Pizza-House-B," he/she may come to the restaurant. Or, a user who like Pizza-House-B would not come Pizza-House-A according to the information. A straightforward comment in a blog often influences a user action as follows.

- Gourmet information
  There are a huge amount of blogs that include reputations of restaurants bloggers came. When a user is influenced from their comments, he/she may take notes of the address of the restaurant, and may visit there.
- Tourist information
  After trips, most of the bloggers write their travels with the photos. When a user has a plan to go to the tourist spot, he/she watches the blogs. He/She may go to the place that bloggers introduced at the same situation.

Anyhow, a time and position information in blogs are important to give users behavioral support information. Therefore, blogs must be indexed by temporal and positional meanings according to its description.

To index content position by latitude and longitude is efficient. When a user detects his/her position by GPS, and sends the position to a location aware service, he/she could know the distance between his/her position and the position of a content. Then, he/she finds out nearby contents. Recently, there are a lot of APIs, as typified by Google Maps API[9], which handles location information. Because contents distributors put location information on web pages by utilizing those APIs, useful location search services are increasing[5]. Additionally, there are some works that collect address description automatically[1].

However, there are few temporal indexing method. Even if a user can get information of nice wine bar, there is no point in getting the information at noon. A location information service may have no effectiveness. Even though RSS (RDF site summary)[11] can describe published time of contents and provide the latest information, these information may be useless for outdoor users. To achieve higher precision with low granularity, we propose a method that collects the event occurred time of a content. It reads the shooting time of a photo which is written in Exif (Exchangeable Image File Format)[7] data of the photo on blogs.

We explain the problems to collect event time from blogs, a proposal of Time-Crawler, the results of evaluations, future work, and conclusion as follows.

II. Extracting temporal expressions from web contents

The update time of contents written in RSS is popular to inform users about latest news or blogs. However, when a user searches restaurant information near his/her location, he/she needs not results of matching with the update time of contents but results of matching with the event occurred time of contents. If a temporal information service provides RSS

978-1-4244-2156-5/08/$25.00 ©2008 IEEE

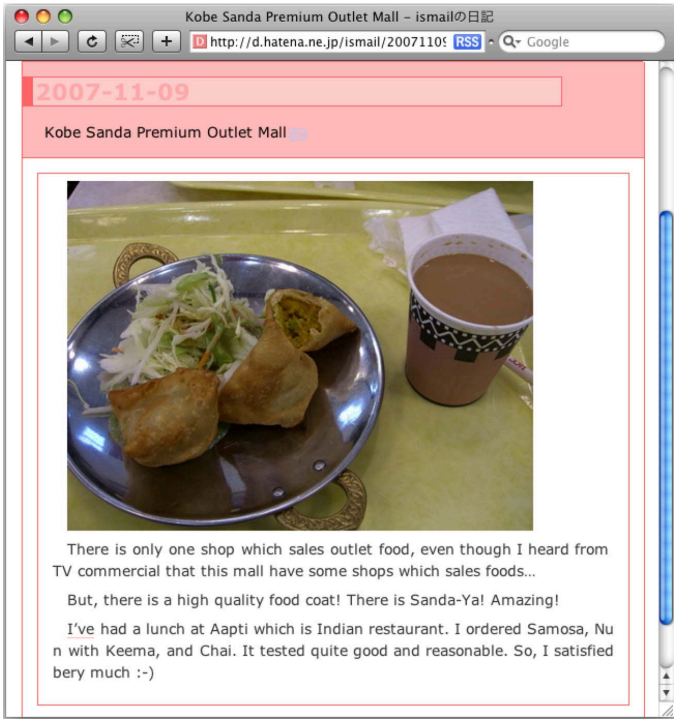

Fig. 1. A sample of a blog

TABLE I

A SAMPLE OF AN EXIF DATA

Item name	Sample value
Camera Model Name	EX-Z750
Shooting Date/Time	11/04/2007 15:00:20
F-number	2.8
Shutter Speed	1/125
Image Size	450x338

information to an outdoor user, the content may not reflect user's circumstance.

TimeML[12] which is a framework for temporal structurization of contents would solve that problem. But, it is out of touch with reality that contents distributors create TimeML information manually. There are some works[2], [3] that create such TimeML information automatically from news sites. Precision, recall and F-measure of the works are approximately 80%. For blogs, a natural language processing is efficient to collect a period of time[4]. But it's difficult to collect an event occurred time because of scarce descriptions of event occurred time in blogs. Even though a user read article, he/she often cannot find the event occurred time. Therefore, we must discuss other methods different from natural language processings.

III. A PROPOSAL OF TIME-CRAWLER

A contents distributor writes a blog about the place where he/she went or the foods which he/she ate; He/she often attaches some photos which is related in the manuscript to the blog. As shown in Fig.1, the photo is attached at top of the blog. Then the content distributor writes his/her diary

TABLE II

A SPECIFICATION OF BLOGS

	Yahoo! blog	ameba blog	rakuten blog	total
having exif	35	34	22	91
not having exif	55	57	33	145
no photos	110	109	145	364
rates of exif	17.5%	17.0%	11.0%	15.2%

TABLE III

RELATIONS OF PHOTOS AND SENTENCES IN BLOGS

	having relations	no relation	Rates of relations
having exif	78	13	85.7%
no exif	109	36	75.2%

below the photo. Ordinary, he/she takes this photo by using a digital camera. Most of photos that are taken by a digital camera have Exif data which is the specification to add some specific metadata to an image file. Table I shows a sample of an Exif data. This data has a lot of information such as date and time information, camera settings and etc. The data is utilized for print optimization normally. We are just interested in the Shooting Date/Time. We can obtain an event time by the second. If most of the manuscript is related to the photos, we expect that the precision of this method will get higher than that of related work.

So, we propose a Time-crawler which collects event time of contents by reading Exif in blogs. We explain the sequence of the system as follows.

1) Collects each article in blogs.
2) Judge whether the article contains photos that have Exif data or not.
3) Extract the Shooting Date/Time from the Exif data; Write it as metadata of the event time.
4) Write up the metadata which includes the event time and the URL of the article; Add the metadata to the database.

We expect that the location information service provides more suitable information to outdoor users by utilizing the stored metadata in the database.

IV. EVALUATION

A performance of the proposed system depends on the relation between the photo and sentences in a blog. At first, we classify blogs by visually checking their relations. Then, we evaluate precisions of the proposed method.

A. Assortment survey of blogs

We assume that a photo which contains Exif has relation to the sentences in the blog. To proof that, we visually check the relation on random sample blogs.

We collects the latest 600 pages from Yahoo! blogs[14], ameba blog[6], and Rakuten blog[10] (all of them are written in Japanese). We visually check them and classify them in three classifications as follows.

- Blogs which contain Shooting Date/Time in Exif

TABLE IV

PRECISIONS OF EACH GOURMET KEYWORDS

Keyword	System correct	System incorrect	Precision
Chinese noodle	20	2	0.91
Yakitori	7	6	0.54
Japanese wheat noodle	11	6	0.65
Pasta	24	6	0.80
Sukiyaki	11	6	0.65
Jial-zi	12	4	0.75
Curry	11	6	0.65
Grilled meat	11	2	0.85
Cake	29	4	0.88
Total	109	42	0.72

If a photo in the blog is generated from a digital camera, the photo would have Shooting Date/Time in its Exif data. The shooting time may be considered as the event occurred time.

- Blogs which contain no Shooting Date/Time
  Some photos don't have Shooting Date/Time despite of having Exif. The disappearance may occur at somewhere around retouching photos. Illustrations are also out of scope for the proposed method.
- Blogs which contain no photo/illustration
  A blogger doesn't print a photo/illustration in his/her blog in the case of that writing only sentences is enough to report some events.

Table II shows the classification of blogs. In total, approximately 15% of them have a Shooting Date/Time in their Exif data. It seems low rate. But, as shown in Table III 85.7% of the blogs having each Shooting Date/Time have relation of the sentences. So, the proposed method is expected to have high precision.

B. Evaluation of precision for a gourmet searching

We evaluate a precision by searching gourmet information as a real scenario of searching user behavioral support information.

To implement proposed method simply, we utilize Google Blog Search (for Japanese)[8]. We input keywords to the Google Blog Search 9 times and download each latest 100 results. The 9 typical keywords of foods are chosen from main categories in Yahoo! gourmet (for Japanese)[15]. Then, we evaluate the precision by visually checking the downloaded blogs. As shown in Table IV, The result of precision was 0.72 in all. The reasons of errors are classified as follows.

- No relation between the photos and sentences. (30 blogs)
  The rate is approximately same as the rate of Table III. However, these errors comprise a large parcentage of the all errors. We should consider to utilize image indexing technologies to solve this problem.
- The keywords hit the title of blogs. (8 blogs)
  This error occurred because we utilize Google Blog Search simply. Setting a search range to only articles would decrease this error.
- The intention was mismatched between the keywords and the results of searching. (6 blogs)
  We have no idea to solve this problem as long as we utilize keyword pattern matching.

The precision of the proposed method is not much higher than that of conventional work that utilize natural language processing. But the granularity becomes smaller enough as it's punctual to the second. So, we confirmed that the proposed method can collect event occurred time with good precision.

V. FUTURE WORK

The precision of the proposed method was 0.72. The first future work is raising this precision by considering the reasons of errors as shown in IV-B.

Furthermore, we have plans to confirm the admissibility of a user behavioral support system which utilizes the proposed method and other positional indexing technologies as follows.

- A gourmet or tourist information search services reflecting a user's circumstance such as his/her schedules and position
  To enable users temporal search, the proposed method is efficient, because this method generates an index which includes event occurred times and the blogs' URL. To enable users positional search, the conventional technologies such as Google Maps API is useful. When their technologies are combined, a user will be able to get his/her behavioral support information. We have a plan to generate queries for the searching automatically by utilizing GPS, personal schedule data, and the clock in a computer to get current time, and so on.
- A support tool to input temporal metadata
  The first motivation of the proposal is making an index of blogs according to their event occurred time. But, their precision is not perfect. The managing cost to keep the index correct is still higher. If each blogger helps to manage the temporal index, the cost gets lower. When he/she uploads his/her blog, a support tool utilizing the proposed method can suggest the event occurred time of the article. If the suggestion is correct, user would confirm and upload it with his/her blog. In opposite case, he/she would correct the suggestion.

VI. CONCLUSION

To provide outdoor users' behavior support information, indexing web contents by time and location is important. We proposed a Time-crawler which indexes articles of blogs by temporal information in Exif data. We expect that this information enhances the effectiveness of conventional location information services. As a result of evaluation, we confirm that 85.7% of the blogs having each Shooting Date/Time have relation of the sentences, and the precision was 0.72 totally. In the future, we will improve the precision, and implement some applications to confirm an admissibility of users.

REFERENCES

[1] Lin Can, Zhang Qian, Meng Xiaofeng, and Liu Wenyin. Postal Address Detection from Web Documents. *Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration*, pages 40–45, April 2005.

[2] Benjamin Han, Donna Gates, and Lori Levin. Understanding Temporal Expressions in Emails. *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 136–143, June 2006.
[3] Inderjeet Mani and George Wilson. Robust Temporal Processing of News. *In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, pages 69–76, 2000.
[4] Taichi Noro, Takashi Inui, Hiroya Takamura, and Manabu Okumura. Time period identification of events in text. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 1153–1160, July 2006.
[5] Yuuichi Teranichi, Junzo Kamahara, and Shinji Shimojo. MapWiki: A Map-based Content Sharing System for Distributed Location-dependent Information. *Journal of computers*, 1(3):13–19, June 2006.
[6] ameba blog. http://ameblo.jp/.
[7] Exif. http://it.jeita.or.jp/document/publica/standard/exif/english/jeida49e.htm.
[8] Google Blog Search. http://blogsearch.google.jp.
[9] Google Maps API. http://www.google.com/apis/maps/.
[10] Rakuten blog. http://plaza.rakuten.co.jp/.
[11] RDF Site Summary (RSS) 1.0. http://web.resource.org/rss/1.0/.
[12] TimeML. http://timeml.org/site/.
[13] The weather channel. http://www.weather.com/.
[14] Yahoo! blogs. http://blogs.yahoo.co.jp/.
[15] Yahoo! gourmet. http://gourmet.yahoo.co.jp/.
[16] Zagat survey. http://www.zagat.com/.