Blog Map of Experiences: Extracting and Geographically Mapping Visitor Experiences from Urban Blogs

Takeshi Kurashima, Taro Tezuka, and Katsumi Tanaka

Department of Social Informatics, Graduate School of Informatics,
Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
{ktakeshi, tezuka, tanaka}@dl.kuis.kyoto-u.ac.jp
http://www.dl.kuis.kyoto-u.ac.jp/

Abstract. The prevalence of weblogs (blogs) has enabled people to share the personal experiences of tourists at specific locations and times. Such information was traditionally unavailable, except indirectly through local newspapers and periodicals. This paper describes a method of spatially and temporally obtaining specific experiences by extracting association rules from the content of blog articles. For example, we can read about visitors' activities and evaluations of sightseeing spots. By geographically mapping their experiences, the proposed system enables observation of tourist activities and impressions of specific locations, which can often be more diverse than local guidebooks and more trustworthy than advertisements.

1 Introduction

The prevalence of weblogs enables the observation of personal experiences, specific to location and time. Such information was traditionally unavailable, except indirectly through local newspapers and periodicals. One characteristic of blogs is that the moment they are written, they are stored as attributes. This enables the extraction of the writers' experiences during a specific time period. When combined with the extraction of geographic keywords (geo-coding of blogs) from the articles, tourist experiences, related to a specific place and time can be obtained. For example, in spring many people go out and see the flower blossoms. At famous sight-seeing spots, they are likely to try out local specialties. These are actual experiences, unlike commercially prepared tourist guides or media reports. This kind of information is particularly valuable to potential tourists and marketing analysts, interested in local trends. Recently, some services have allowed users to map their blog articles to spatial locations. However, such systems are not yet widespread, and most of the location-specific personal experiences are stored separately on individual blog sites. Existing search engines do not provide satisfactory results because the search results for specific location names are often a vast collection of blog articles, and it is unrealistic to read them all. In addition, many of the articles contain only generic information about the location,

M. Kitsuregawa et al. (Eds.): WISE 2005, LNCS 3806, pp. 496–503, 2005.

and not the writer's personal experience. In this paper, we describe a system for extracting local and temporal experiences by mining spatially and temporally specific association rules from blog articles. The system enables people to learn about the experiences of others, specific to location and time period.

2 Related Work

2.1 Blog Mining

Kumar et al. discussed blog link structures and proposed *time graphs* and *time-dense community tracking* to represent the growth of blog communities and a way to track changes within communities [1][2]. Bar-Ilan examined links between blogs and their postings and obtained statistics [3]. However, these analyses focused only on the relationships between blogs, and were not targeted at the content of the blog articles. Bar-Ilan's survey also pointed out that personal information and self-expression are becoming important aspects in blogs. Okumura et al. proposed a system that collects blog articles and presents aggregated results [4]. Unlike our system, *blogWatcher* does not perform spatial aggregation. Avesani et al. proposed a system which aggregates blog articles on specified topics [5]. Their work was mainly focused on aggregating user reviews on products.

2.2 Spatial Blogs

There have been various services that unify geographic information with blogs. DC Metro Blogmap and nyc bloggers provide services that link one's personal blogs to metro maps. Users of these services can find bloggers related to specific area in the city [6][7]. Uematsu et al. proposed *Ba-log*, in which users upload blog contents using cellular phones equipped with cameras and GPS extensions [8]. WorldKit is a toolkit for creating map based applications on the Web, and it has been applied to a blog mapping service also [9]. However, these services require manual registrations, and the automatic extraction of knowledge from blog articles is not performed.

2.3 Association Rule Mining

Association rules are patterns in relational data, which can be expressed as $X \Rightarrow Y$, where X and Y indicates sets of expressions with a certain attribute having a specific value [10]. Two factors, *support* and *confidence*, determine the value of each association rule. *Support* is the ratio where one tuple contains both X and Y of all the tuples in a data set and is expressed as, $sup(X \Rightarrow Y)$. *Confidence* is the ratio of tuples containing Y in tuples containing X. *Confidence* is expressed as $conf(X \Rightarrow Y)$.

In extracting association rules, the minimum confidence (MCV) and support values (MSV) are set as thresholds. Association rules with a higher confidence rate than MCV and a higher support rate than MSV are extracted as important rules. The APRIORI algorithm described by Agrawal et al. enabled the extraction of the association rule from large data sets in practical time [11].

3 Experience Extraction

The aim of this paper is to describe the extraction of tourists' real life experiences. These experiences are written while the memories are still fresh and with an honesty often lacking in commercially written pieces, although not all sentences in blog articles describe real-life experiences. In order to extract actual experiences, we extract sentences that refer to actions. After that, we are mining spatially and temporally specific association rules from blog articles because tourist's real life experiences is local and temporal. The *antecedent* of the association rules is place and time, and the *consequent* is verb and noun. The following subsections describe each of steps of the extraction.

3.1 Blog Collection

Blog articles are collected using generic blog search engines, provided by blog hosting services. These hosting services provide the search results in RSS metadata. RSS is an RDF rich site summary that consists of titles, summaries, date, and other attributes. The system collects articles by following these steps:

1. Set location names as search queries.
2. Send queries to generic blog search engine
3. Retrieve search results in RSS.
4. Extract title, content, links, and date from RSS.
5. Store in the blog database.
6. Wait for set amount of time, and repeat from (2).

3.2 Morphological Analysis and Database Insertion

From the collected articles, transaction sets are extracted. The surrounding text of a location name is first extracted from the blog content. The extracted text is divided into sentences and then into morphemes. This process is language dependent and is discussed in more detail in the implementation section. Then the sentences are converted into a *transaction*. The scheme is as follows.

$T = (date, geo, noun_1, ..., noun_n, verb_1, ..., verb_m)$
T: Transaction.
$date$: Date attribute of the blog article containing the sentence.
geo: Location name found in the sentence.
$noun_i$: ith noun in the sentence.
$verb_i$: ith verb in the sentence.

3.3 Refinement 1: Extraction of Verbs Referring to Actions

In order to extract only the rules that are related to user experiences, the transaction database must be refined. In general, sentences can be divided into the three groups listed below.

1. Do statements (ex. *I saw autumn leaves.*)
2. Become statements (ex. *The autumn leaves turned yellow.*)
3. Be statements (ex. *Autumn leaves are beautiful.*)

Sentences are categorized into these groups, based on their verbs. Experiences are most closely related with the do statements, because these sentences indicate user action. Therefore, we only use transactions that contain verbs that are used in do statements.

3.4 Refinement 2: Elimination of Verbs Indicating Movements

Actions that do not take place in the specified location are eliminated. For example, verbs such as "go" or "come" do not indicate actions that take place at the specified location. Instead, these indicate movement toward the location. Therefore, transactions containing these verbs are eliminated from the transaction database.

3.5 Association Rule Mining

The association rules are extracted using the APRIORI algorithm. There are three types of rules that we can extract.

Type 1: [Time, Location name] ⇒ [Verb]
Type 2: [Time, Location name, Verb] ⇒ [Noun]
Type 3: [Time, Location name] ⇒ [Verb, Noun]

Other rules, such as those between nouns, are eliminated because they do not match our purpose of extracting spatially and temporally specific experiences.

3.6 User Interaction

The extracted association rules are presented as a *summary* to the user's request. The user query consists of a combination of four attributes: location name, time, action, and object. Some typical combinations are described below.

Case 1: Query = Space, Time Search Result = Action

In this case, the user wants to know about typical activities taking place at a specific place during a specific time period. For example, the user can input "Kyoto city,April".

Case 2: Query = Space, Time, Action Search Result = Object

In this case, the user wants to know the object of specified activities taking place at a specific place and time. For example, the user can input terms, such as "Kyoto city, April, Eat".

Case 3: Query = Space, Time Search Result = Action, Object

In this case, the user wants to know about typical activities taking place at a specific place and time period, as well as the object of the action.

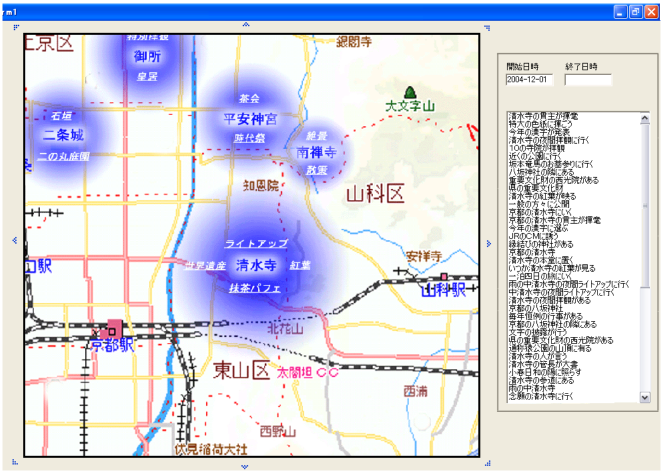
Then the user can view the set of original articles and read about the writer's experiences in more detail by making a click on a link. The user can also access to the extracted living experiences through a map interface. A mock-up image of the visual interface is shown in Figure 1.

Table 1. Case3 summary

Location name: Date:
[rank] Verb1:conf([Location name,Date] ⇒ [Verb])

Object1:conf([Location name,Date,Verb] ⇒ [Noun])
Object2:conf([Location name,Date,Verb] ⇒ [Noun])

[rank] Verb2:conf([Location name,Date] ⇒ [Verb])
Object1]:conf([Location name,Date,Verb] ⇒ [Noun]) Object2:conf([Location name,Date,Verb] ⇒ [Noun])



開始日時 2004-12-01
終了日時

清水寺の貫主が揮毫
特大の色紙に揮ごう
今年の漢字が発表
清水寺の夜間拝観に行く
10の寺院が拝観
近くの公園に行く
坂本竜馬のお墓参りに行く
八坂神社の隣にある
重要文化財の西光院がある
県の重要文化財
清水寺の紅葉が映る
一般の方々に公開
京都の清水寺にいく
京都の清水寺の貫主が揮毫
今年の漢字に選ぶ
JRのCMに誘う
縁結びの神社がある
京都の清水寺
清水寺の本堂に置く
いつか清水寺の紅葉が見る
一泊四日の旅にいく
雨の中清水寺の夜間ライトアップに行く
中清水寺の夜間ライトアップに行く
清水寺の夜間拝観がある
京都の八坂神社
毎年恒例の行事がある
京都の八坂神社の隣にある
文字の披露が行う
県の重要文化財の西光院がある
通称猿公園の山頂に有る
清水寺の人が言う
清水寺の管長が大書
小春日和の陽に照らす
清水寺の参道にある
雨の中清水寺
念願の清水寺に行く

Fig. 1. Visual user interface for browsing actual experiences

4 Implementation

Based on the algorithm described in the previous section, we implemented a prototype system, the Blog Map of Experiences, which extracts and summarize actual experiences from blog articles. The next section describes the implementation details. A system configuration is shown in Figure 2.

4.1 Blog Collection and Morphological Analysis

Our implementation collects blog articles from "goo blog"[12] and "livedoor blog"[13]. Both are typical blog hosting services in Japan. Location names used as search queries were taken from digitized residential maps provided by Zenrin Ltd.[14] The collected blogs were stored into the MySQL database [15].

Morphological analysis of the blog articles was performed, using the Chasen morphological analyzer [16]. Chasen divides sentences into words and estimates their parts of speech. The results are then stored in the transaction database.

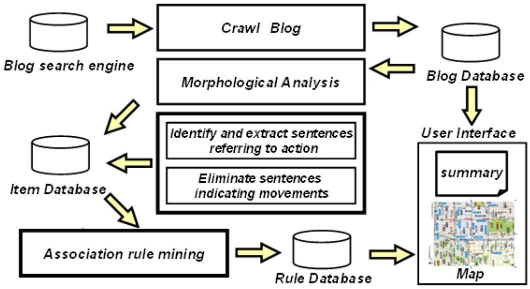


Fig. 2. system configuration

4.2 Refinement of Transactions and Association Rule Mining

The collection of verbs referring to actions was obtained from the lexical database of the Japanese Vocabulary System [17]. It categorizes verbs into tree structure, and verbs referring to actions are grouped into one top-level category.

We then manually listed verbs that indicate movements. For example, verbs such as “go”, “come”, and their synonyms were among the list. Transactions containing these verbs were eliminated, since their original sentences are likely be describing the motions toward the location, rather than the actions at the location. The latter is what we want to extract and present to the users. As a preprocessing of the association rule mining, time attribute of the transactions were grouped by month, so that the resulting association rules have higher support values. Finally, APRIORI algorithm was used for the association rule mining.

In our implementation, we extracted Type 1 and Type 2 rules discussed in the previous section. The result of a preliminary experiment showed that Type 3 rules contain too much noise and could not be used without further improvements in the refinement methods. The precisions of these rules were around 10 percent.

The extraction of Type 2 rules is performed in two steps. First, Type 1 rules are extracted and a set of typical verbs for the given place name is obtained. Second, Type 2 rules are extracted for the pairs of the given place name and the extracted typical verbs. The pseudocode for this process is the following.

```
Define a set of place names $P = p_1, p_2, ..., p_n$.
For $i = 1$ to $n$ do
    Obtain $m$ association rules $p_i \rightarrow v$,
      in the decreasing order of the support value.
    Obtain the set of verbs $V_i = v_{i1}, v_{i2}, ..., v_{im}$
    For $j = 1$ to $m$ do
        Obtain $k$ association rules $p_i, v_{ij} \rightarrow n$,
          in the decreasing order of the support value.
        Store $k$ rules.
    Done
Done
```

The result was stored into the rule database.

5 Evaluation

In this section, we evaluate two refinement algorithms proposed in the previous section, as listed below.

Refinement 1: Identify and extract sentences refering to actions.
Refinement 2: Eliminate sentences indicating movements.

We first apply the conventional association rule mining method (APRIORI algorithm) and obtain results. We then calculate the precision for different sizes of the extracted rules. We then apply two refinement algorithms, and observe if they improve the precision.

We performed experiments using two major blog hosting services in Japan, goo and livedoor [12][13]. The target locations were 20 popular sightseeing spots in Japan, and we collected 500 blog articles for each. First, Type1 rules are extracted, and top j rules are obtained in the decreasing order of the *support* value. Second, Type2 rules are also extracted each of the verbs which are the *consequent* of the Type1 rules. We obtained top 10 nouns which are the *consequent* of the Type2 rules. We evaluated the extracted pairs of verbs and nouns. Table 2 is the average results of precisions. The results show that the combination of Refinement 1 and 2 improves the resulting rule set.

Table 2. The average Precision of the extracted association rules

top j result	Size of experiences	Assoc.rules:unrefined	Refine.1	Refine.1+Refine.2
3	30	0.007	0.083	0.216
5	50	0.058	0.111	0.221
10	100	0.087	0.131	0.182

6 Existing Problems

There are still several remaining problems in our system. One is that the system cannot handle synonyms yet. Another is that it does not consider dependency between terms. There is still some ambiguity regarding whether the noun extracted using the association rule is really the object of the action. Our future plans include applying dependency analysis to extract the objects of the action more precisely.

7 Conclusion

In this paper, we described a system for extracting actual experiences related to a specific location and time period. Association rules between place, time, action, and objects aggregate the user experiences expressed in blog articles. We have implemented a system that allows users to search actions or objects using specified places and times. The results verified that the system could successfully extract actions and objects.

Acknowledgments

This work was supported in part by the Japanese Ministry of Education, Culture, Sports, Science and Technology under a Grant-in-Aid for Software Technologies for Search and Integration across Heterogeneous-Media Archives, a Special Research Area Grant-In-Aid For Scientific Research (2) for the year 2005 under a project titled Research for New Search Service Methods Based on the Web's Semantic Structure (Project No, 16016247; Representative, Katsumi Tanaka), and the Informatics Research Center for Development of Knowledge Society Infrastructure (COE program by Japan's Ministry of Education, Culture, Sports, Science and Technology).

References

1. R. Kumar, J. Novak, P. Raghavan and A. Tomkins, On the bursty evolution of blogspace, Proceedings of the 12th International World Wide Web Conference, pp. 568-576, 2003
2. R. Kumar, J. Novak, P. Raghavan and A. Tomkins, Structure and evolution of blogspace, Communications of the ACM, 47(12) pp. 35-39, 2004
3. J. Bar-Ilan, An outsider's view on 'topic-oriented' blogging, Proceedings of the Alternate Papers Track of the 13th International World Wide Web Conference, pp. 28-34, 2004
4. M. Okumura, T. Nanno, T. Fujiki and Y. Suzuki, Text mining based on automatic collection and monitoring of Japanese weblogs, The 6th Web and Ontology Workshop, The Japanese Society for Artificial Intelligence, 2004
5. P. Avesani, M. Cova, C. Hayes and P. Massa, Proceedings of the WWW2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, Chiba, Japan, 2005
6. nyc bloggers, http://www.nycbloggers.com/
7. DC Metro Blogmap, http://www.reenhead.com/map/metroblogmap.html
8. D. Uematsu, K. Numa, T. Tokunaga, I. Ohmukai and H. Takeda, Ba-log: a proposal for the use of locational information in blog environment, The 6th Web and Ontology Workshop, The Japanese Society for Artificial Intelligence, 2004
9. worldKit, http://www.brainoff.com/worldkit/index.php
10. D. J. Hand, H. Mannila and P. Smyth, Principles of Data Mining (Adaptive Computation and Machine Learning), 425p, MIT Press, 2001
11. R. Agrawal and R. Srikant, Fast algorithms for mining association rules in large databases, Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487-499, 1994
12. goo blog, http://blog.goo.ne.jp
13. livedoor blog, http://blog.livedoor.com/
14. Zenrin Co.,Ltd, http://www.zenrin.co.jp/
15. MySql, http://www.mysql.com/
16. Chasen, http://chasen.aist-nara.ac.jp/index.html
17. Japanese Vocabulary System, http://www.ntt-tec.jp/technology/C404.html
18. Geolink Kyoto, http://www.digitalcity.gr.jp