Users Sleeping Time Analysis based on Micro-blogging Data

Haoran Yu
Key Lab. on High Performance Computing, Anhui Province
School of Computer Science
University of Science and Technology of China
haoran_yu@hotmail.com

Guangzhong Sun
Key Lab. on High Performance Computing, Anhui Province
School of Computer Science
University of Science and Technology of China
gzsun@ustc.edu.cn

Min Lv
Key Lab. on High Performance Computing, Anhui Province
School of Computer Science
University of Science and Technology of China
lvmin05@ustc.edu.cn

ABSTRACT

The emergence of new social network services, often labeled as Web 2.0, has permitted an amazingly increase of user generated content. In particular, Sina Weibo, a popular Chinese micro-blogging service is designed as platforms allowing users to generate contents that open to the public. From analyzing activates of submitting posts to Sina Weibo, some features of users can be estimated. This paper aims to contribute to this growing body of literature by studying how users' frequent activities reflect their sleeping time and living time zones. By mining a large set of users' activates data from Sina Weibo, we demonstrate its possible role to detect the sleeping time of users and find a new method for judging users' time zone.

Author Keywords

social networking services, micro-blogging, time series, sleeping time, time zone.

ACM Classification Keywords

H.2.8 [Database Management]: Database Applications—Data mining; H.3.3 [Information Storage and Retrieval]:Information Search and Retrieval—Relevance feedback.

General Terms

Measurement, Documentation, Experimentation, Human Factors, Verification.

1. INTRODUCTION

With the development of social network and related technologies, users, as the core part of the service, communicate with each other and generate a lot of content. Twitter, Facebook, Sina Weibo and other services became the most popular social communication platforms. A growing number of researchers are focusing on related topics. Since considerable features of users are reflected by their activities, we can estimate unknown features of users by analyzing the data of users. Within all the features, the time zone and the corresponding location are both important features, which are related with users' sleeping time.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

This paper aims to contribute to this growing body of literature by studying how users' frequent activities reflect their sleeping time and living time zones. By mining a large set of users' activates data on Sina Weibo, we demonstrate its possible role to detect the sleeping time of users and find a new method to judge users' time zone. Specifically, the analysis not only estimates the sleeping time pattern of Weibo users but also gives out the authenticity of user's location in their profile and the result of movement between time zones (e.g. from China to U. S.) detection. Although our study involves only simple methods and the general conclusion, it presents a kind of interesting orientation for analyzing users' sleeping activities and time zone detecting.

The main contributions of this paper are as below:

- A new open problem related to the relationship between sleeping time and micro-blogging activities is presented. Besides, this paper associates the pattern of sleep and the time zone users live in.
- For such a problem above, preliminary simple solutions are presented in this paper. Based on experiment results on real data set, the proposed solutions are proved both efficient and effective.

UbiComp'12, Sep 5 – Sep 8, 2012, Pittsburgh, USA.

- A data set of users' time series of activities on Weibo is collected and open to the public. The URL is: http://www.haoranyu.com/research/weibosleep

Meanwhile, there are some weak points in this study. Some impact factors, including the job of users, the scale of the city and the individual habits, are not considered. Plus, the technical depth is not very deep so far. Text analysis on micro-blogging data, interactions between friends and geo-tagged sources can be used to further improve the accuracy of prediction in the future.

The rest of this paper is organized as follows. In Section 2, we introduce the Sina Weibo platform and review related works. Section 3 motivates our research and provides our method and results for detecting sleeping time of users. Later in Section 4 we analyzed the relationship between time zone and users' sleeping time. Meanwhile we show some interesting cases. Finally, we conclude the paper in Section 5.

2. RELATED WORK

2.1. Sina Weibo

Sina Weibo[6] (http://weibo.com/), as a micro-blogging platform, akin to a hybrid of Twitter and Facebook; it is used by well over 30% of 513 million users (Up to the end of year 2011) in China. It has a similar market penetration that Twitter has established out of China. Sina Weibo has a verification program for famous people and special organizations. Once an account is verified, a 'V' badge and a verification description will be added next to the account name.

There are several conventions used by Weibo users to convey information within the limit of 140 characters.

- Hashtags (#) userd in the "#Topic#" format are used to add discussion topic and group related posts together.
- Weibo users may talk to or quote other people by using an @username in post.
- Weibo users can re-post with "//@UserName". Similar to Twitter's retweet function.
- URLs posted by user are automatically shortened using the domain name 't.cn'.
- Comments to a post can be shown as a list right below the post.

2.2. Previous Research

There are more and more published studies related user activities in micro-blogging and social network services. The authors of [3] collected data from digg and reddit and get an understanding of how content is generated and how the popularity of a post evolves overtime. Bertrand De Longueville, improve the understanding on how LBSN can be used as a reliable source of spatio-temporal information, by analysing the temporal, spatial and social dynamics of Twitter activity during a major forest fire event in the South of France in July 2009. [2] Kate Ehrlich and N. Sadat Shami (2010) studied users of BlueTwit and Twitter over the same time period. Deepen the understanding of the workplace benefit of micro-blogging and examined how people appropriate social technologies for public and private use.[4] Tony S.M. Tse and Elaine Yulan Zhang (2012) analyzes blog and microblog contents created by mainland Chinese visitors sharing their Hong Kong experiences and find out that a generally positive image of Hong Kong as a destination among the mainland Chinese bloggers.[7] Alan Mislove, Bimal Viswanath, ect discovered that certain user attributes can be inferred with high accuracy when given information on as little as 20% of the users[1].

SHI Xuemin and ZHANG Chuang (2011) studied users' activities on Sina Weibo and found that the peak time that users update posts is different in every time zone. [5] The data set of their research is not properly pretreated. Besides, in their experiment the locations labeled by users are 100% trusted. This study, however, may just be scratching the surface of how time series of micro-blogging can be used for research.

3. DETECTING SLEEPING TIME

Although, more and more researchers are now paying attention to micro-blogging and other similar social network services, most of them have focused on their social dimension by studying users' motivations, interactions or collaboration. This paper pays attention to the posts time series of on Sina Weibo.

3.1. When do users sleep?

According to the common sense, excluding few people who works on the night shift, people are active at day time and inactive while sleeping time at night. However, it is not an easy task to accurately figure out when do users of Sina Weibo sleep. This question will be discussed in this paper.

Sina Weibo is obviously a reliable network service that bundles time and location together. Each of the posts on Sina Weibo has been published with a level of accuracy of 1 minute and the list of posts is organized in reverse order by a natural time. Thus, it is necessary to obtain a sequence of data points of time from Sina Weibo.

3.2. Find the longest inactive time span to discover sleeping time from dense data

This method is used to estimate the span of users' sleeping time under ideal conditions. Under this condition, people sleep for a long time once a day and keep active in the rest of time. Namely, we can't find out two or more long inactive span, equal or greater than 5 hours, within any 1440-minutes-period (as long as a day, that is 24 hours × 60 minutes /hour). Plus, any of the users must stays in the same time zone and keeps stable activity habits

As depicted in Figure 1, time series data is a sequence of time points, $T = \{t_1, t_2, \dots ,t_n\}$. Each time point $t_i \in T$ corresponds to a time of a post on Sina Weibo.

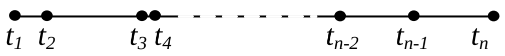

Figure 1: Time Series

From the time series data, the lengths of spans $D = \{d_1, d_2, \ldots, d_{n-1}\}$ determined by every two points can be defined as:

$$d_i.length = t_{i+1} - t_i$$

All the long inactive spans ($d_i.length \geq$ 5h) can be easily detected from a time series data. Each of the result spans is marked as $d'_j = [t_{j0}, t_{j1})$, $j \in (1, k)$. Therefore, a final span (d')¯representing the average sleeping time of a user can be figured out by calculating the average lower bound and upper bound of all result spans.

$$\overline{d'} = \left[\frac{\sum_{j=1}^{k} t_{j0}}{k}, \frac{\sum_{j=1}^{k} t_{j1}}{k}\right)$$

3.3. Statistics method in discovering sleeping time pattern from sparse data

As a matter of fact, the majority of Sina Weibo users are not always keeping active during every whole day period. Namely, the method in section 3.2 is not practical or useful in real application. In the real world, the users' data is much sparse. Less than 1% of Weibo users can satisfy the requirement. In this section, we will present a statistics method in discovering sleeping time pattern from sparse data. In this new method, we still assume that users keep a daily routine for their life, going to bed and waking up on time.

Within any 1440-minutes-period, we may able to find two or more long inactive span, equal or greater than 5 hours. Thus, the statistics method is supposed to be an effective way. Daily data of time series are united together by a section as it is showed in Figure 2.

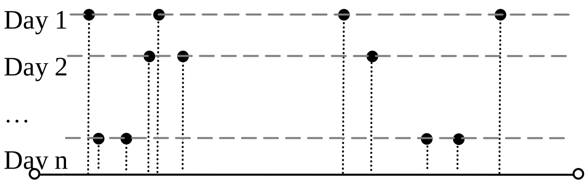

Figure 2: United Time Series

3.3.1. United time series by month

In order to solve this problem in this case, daily data of the time series are united by month for example, as new sequences Tm = {tm1, tm2 … tmn}. Each time point tmi∈Tm corresponds to a time point of a Sina Weibo post with the month m. (See Figure 3)

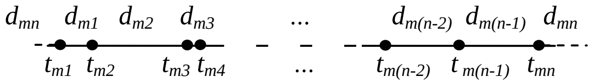

Figure 3: Time Series by Month

In each of the time series data of month m, the lengths of spans Dm = {dm1, dm2, … ,dmn} determined by every two points can be defined as:

$$d_{mi}.length = \begin{cases} t_{m(i+1)} - t_{mi} \ (1 \leq i < n) \\ t_{m1} - t_{mn} + 1440 \ (i = n) \end{cases}$$

Each of the long inactive spans can be easily obtained from the time series data of the corresponding month. Accordingly, the lower bound and upper bound of user sleeping time in a specific month will be approximately showed.

3.3.2. Subsequence statistics

Yet the method in section 3.3.1 also has defects in practical use. The sparsity of time series data is different from different users. The chosen length of the section for doing statistics cannot match all the circumstances. Some users may frequently submit posts on Sina Weibo and the length of the section for statistics may better be set as a month. At the same time, some other users seldom have a post on Sina Weibo and the length may better be set as long as a quarter (or even a year).

In addition, time points that users start traveling are not always at the first date or last date of the section we chose. In other words, people may have different daily sleeping time during the section of time chosen, which will lead to a meaningless result.

Using the algorithm shown in the following figure, these spans of sleeping time can be estimated automatically from a time series data $T = \{t_1, t_2, \ldots, t_n\}$ of a user.

Algorithm SleepTime_estimation

```
Input: A time series data of a user $T$
Output: A set of sleeping time $ST$
$S$=∅, i=1, pointNum = |$T$| // the number of time point
$S$.insert($t_i$);
while $i$ <= pointNum do,
    $j$ := $i$ + 1;
    while !inactSp(S) do,//find no inactive span in $S$
        S.insert($t_j$);
        $j$:=$j$+1;
    $ST$.insert(inactSp($S$));
    $i$:=$i$+1;
    S=∅;
return $ST$;
```

The reasons for why we estimate sleeping time span in such a way lie in two aspects. On the one hand, it automatically determines the start point and the shortest length of section needed for finding out an inactive span of time. On the other hand, the sections are different, but overlapped. If we need to figure out the sleeping time with a time span, we can average the result from all sections covered by this time span. Moreover, an empty set or a set with a few elements will be returned if the input data series is too sparse.

However, with the increasing of computing accuracy, the computation efficiency of this method decreases significantly. It will take excessive time for process thousands of time point of each user. Thus, we will not use it for experiment.

3.4. Dataset and Experiment Result

In order to gain information for the present study, a spider program is written to crawl related information from Sina Weibo, obtaining users' location and all the posts they submitted.

844 users are included in the dataset. These users are randomly selected from all the verified users. All individuals of them have more than 500 fans and 100 posts. This requirement ensures that the users are active for a long time and the time series of them will not be too sparse for statistic.

About 1, 579, 623 posts of these users are grasped (started on the 14th of August, 2009 and ended in April, 2012). From each post, a time stamp can be extracted from it. Time series of a user consists of all the time stamps of this user. (The data set mentioned above can be obtained from section 1)

By using the method we introduced in section 3.3.1, the lower bound and upper bound of user sleeping time in a specific month are obtained. Two sample graphs of result are shown in Figure 5 and Figure 6.

4. DETECT TIME ZONES BY SLEEPING TIME SPAN

Location information of users is crawled from their profiles. We first map the time zone of these users from their labeled location. Then we can find a relationship between the detected sleeping time and users' time zone.

4.1. Detect time zone of Sina Weibo users

Users in the same time zone share the similar sleeping time pattern. Namely, each sleeping time pattern can correspond to a time zone that users are live in. Therefore the pattern of each time zone is discovered.

According to the real dataset, most of Sina Weibo users submit less than 10 posts a day. It is obviously sparse. The following experiments are based on the method in 3.3.1. Sleeping time patterns of GMT/ GMT+1/ GMT+6/ GMT+7/ GMT+8/ GMT+9 time zones are as follows. (Not all the data are included. Some users label their location as a country and can't map to a time zone, for example the United States and Australia).

Time zone	Main City	Start sleeping time (GMT+8)	End sleeping time (GMT+8)
GMT	London	7:00 ~8:00	14:00~15:00
GMT +1	Paris	6:00 ~7:00	13:00~14:00
GMT +6	Dhaka	1:00~2:00	9:00~10:00
GMT +7	Hanoi	0:00~1:00	8:00~9:00
GMT +8	Beijing	23:00~0:00	8:00~9:00
GMT +9	Tokyo	23:00~0:00	7:00~8:00

Table 1. Relationship between time zone and sleeping time.

4.1.1. User with fake or unclear location label

Sina Weibo users label their location in the profile without verification. Some of them label the location as 'unknown' or as fake locations.

Fake location labels do considerable harm to the social network services. For example, recommendation algorithm with location factor will be apparently affected.

In our result, 12 users are judged to have obvious fake location labeled in their profiles. (Fake location within the real time zone can't been detected)

For example, Xiaosong Gao (weibo.com/u/1191220232), a famous musician of mainland China, labels his location as 'The United States, Oversea' (GMT-8~GMT-5). But due to our result, his location is judged to be not true. His sleeping time pattern is shown in Figure 4. Nage

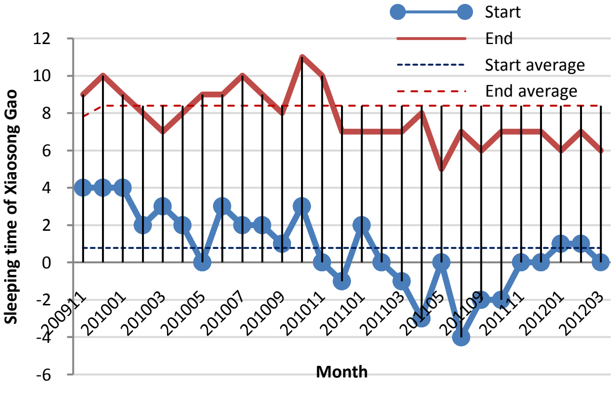

Figure 4: Sleeping time of Xiaosong Gao

The result shows that he sleeps at about 0 a.m. (GMT+8) and wakes up at about 8 a.m. (GMT+8). The corresponding time zone is GMT+7. Therefore, the location Mr. Gao labels himself is fake according to the difference between the location in his profile and the detected time zone.

4.1.2. Detection on movement between different time zones

People are not always located in an area without a movement. There are users who study abroad or have business overseas. They travel far, but they won't change the label of location in the profile page frequently.

Detect and understand the obvious location movements are beneficial. Customized advertising related to locations, such as advertising for local business can be accurately sent to users.

In our result, 33 users have obvious movement (move from locations in different time zones). A user, named Xiaoqian Liu (weibo.com/u/1693775877) describes himself as a reporter of CCTV Brazil and labels his location as 'Brazil, Oversea'. In the result, he is judged to have obvious movement. His sleeping time pattern is shown in Figure 5.

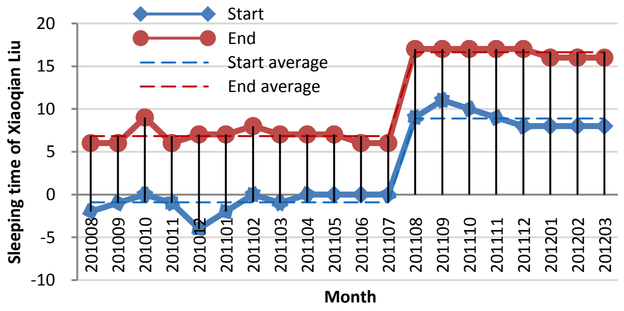

Figure 5: Sleeping time of Xiaoqian Liu

The result indicates that, before Sept. 2011, he goes to bed at about 0 a.m. (GMT+8) and wakes up at about 6 a.m. (GMT+8). The corresponding time zone is GMT+8~GMT+9. After Sept. 2011, the time he starts sleeping change to about 9 a.m. (GMT+8) and the sleep period end at 4 p.m. (GMT+8). The corresponding time zone is GMT-3.5 ~ GMT-2.5. Thus, the obvious movement is detected and the time of this change can be estimated.

5. CONCLUSION

In this paper, three methods for estimating the sleeping time of users are expounded. Broad locations are detected by the linear relationship between sleeping time and time zone of users, which is benefit to study on location-based social networks (LBSN).[8][9] Besides, the study can be extended to other data sets. For example, by mining the web logs on the web server, the real users and robots from specific area can be clearly distinguished. In the future work, some relevant research can be developed to compensating the shortage of this study.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China under the grant No. 61033009 and No. 61103228. It is also supported by the Anhui Natural Science Foundation under the grant No. 1208085QF106.

REFERENCES

1. Alan Mislove, Bimal Viswanath, P. Krishna Gummadi, Peter Druschel. You are who you know: inferring user profiles in online social networks. *WSDM 2010*, (2010), 251-260
2. Bertrand De Longueville, Robin S. Smith, Gianluca Luraschi. "OMG, from here, I can see the flames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires. *GIS-LBSN 2009*, (2009), 73-80
3. Christian Wallenta and Mohamed Ahmed and Ian Brown and Stephen Hailes and Felipe Huici. Analysing and Modelling Traffic of Systems with Highly Dynamic User Generated Content. *UCL Research Note RN_08_10*, (2008), http://web4.cs.ucl.ac.uk/staff/C.Wallenta/research/wallenta_RN_08_10.pdf.
4. Kate Ehrlich, N. Sadat Shami. Microblogging Inside and Outside the Workplace. *ICWSM 2010*, (2010)
5. SHI Xuemin, ZHANG Chuang. Time Zone Prediction Based on User Behavior of Microblogging. *Science paper Online*, (2011), (in Chinese) . http://www.paper.edu.cn/index.php/default/releasepaper/content/201112-467
6. Sina Weibo. http://en.wikipedia.org/wiki/Sina_Weibo
7. Tony S.M. Tse, Elaine Yulan Zhang. Analysis of Blogs and Microblogs: A Case Study of Chinese Bloggers Sharing Their Hong Kong Travel Experiences, *Asia Pacific Journal of Tourism Research*, (2012), DOI:10.1080/10941665.2012.658413
8. Yu Zheng. Location-based social networks: Users. Computing with Spatial Trajectories, Yu Zheng and Xiaofang Zhou Eds. Springer (2011). ISBN: 978-1-4614-1628-9
9. Yu Zheng. Tutorial on Location-Based Social Networks. *WWW2012*, (2012).