2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Mutual Interaction Model between the Number of People in Real Space and the Number of Tweets in Virtual Space

Masaki Onishi
National Institute of Advanced Industrial Science and Technology (AIST)
onishi@ni.aist.go.jp

Shinnosuke Nakashima
Japan Advanced Institute of Science and Technology (JAIST)

Abstract

The relationship between human behavior and SNS (social networking service) activity is growing stronger every year as SNS websites such as Twitter and Facebook develop and expand. Accordingly, a significant number of studies related to SNS have also been reported. However, the majority of this research, such as that which examines the transmission of information via SNS websites, focuses entirely on the virtual world of SNS and few studies have established connections between virtual and real space. This paper focuses on one SNS website, Twitter, and proposes a method for analyzing the statistical relationship between the number of tweets posted on Twitter and the number of people visiting a real world location in the form of a mutual interaction model consisting of nodes and links. We conducted an experiment to verify the proposed method, in which we compared measurements of the number of passers-by in Akihabara with the number of tweets posted.

(from http://twitpic.com/1az6wc)

Figure 1. Crowds of people flocking to Takeshita Street

1. Introduction

Recently, SNS (social networking service) websites, such as Twitter and Facebook, have gained popularity as tools for transmitting new information. According to a report published by Japanese government, as of January 2012, there were 14 million active Twitter users in Japan while the number of active Facebook users was set to reach 15 million. The major difference between SNS and conventional methods of information diffusion is that SNS websites enable users to access services at any time or place using smartphones. This enables a large number of people to transmit real-time information and obtain and diffuse information posted by others.

One example of an incident in which the characteristics of SNS became apparent was the trouble that occurred around Takeshita Street in Harajuku Japan on March 26, 2010. The incident began when someone announced around 4:20 pm that a certain artist would perform a surprise concert on Takeshita Street. In fact, the artist had no plan of holding a surprise concert and the information was no more than a false rumour. However, this false rumour was disseminated via Twitter and other SNS sites and, as shown in Figure 1, crowds of people thronged to the street. At around 4:30 pm, broadcasters dispatched information helicopters as ambulances rushed to the scene. This example also demonstrates that, in modern society, the relationship between SNS activity and human behavior is growing stronger. Elucidating the relationship between SNS activity and human behavior is not only important from the perspective of disaster prevention, in terms of predicting and preventing the occurrence of abnormal situations due to the explosive spread of information online, but also from an industrial perspective, since SNS can be used to create effective marketing.

As SNS has developed and expanded, a significant number of related studies have also been reported (see [1, 2, 3, 4, 5]). Nevertheless, the majority of this research focuses entirely on phenomena in the virtual world of SNS and there are few studies that establish connections between real and virtual space. Therefore, this paper focuses on one SNS website, Twitter, and proposes a mutual interaction model for connecting the number of tweets in virtual space to the number of people in real space. We measured the number of tweets and people and applied these values to the mutual interaction model, and by analyzing the results, verified the statistical relationship between the two variables.

2. Measuring the Number of People and the Number of Tweets

Since we have already measured the number of people and tweets in previous research, an outline of the measurement method used is given below. See [6] for further details.

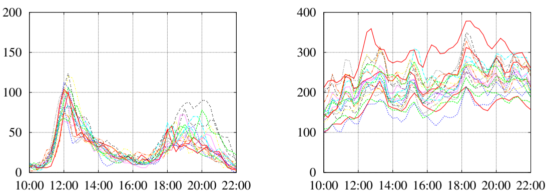
2.1. Measuring the Number of People

We positioned stereo cameras in a complex facility near Akihabara Station and measured the number of people passing in front of the camera. The number of people $\boldsymbol{F}$ was stored as time-series data F_i in 15-minute intervals for each specific direction and the total number of people passing in front of the camera during the 15-minute period was used as the measured value for each time step.

$$\boldsymbol{F} = \{F_1, F_2, F_3, \cdots, F_i, \cdots, F_n\}. \tag{1}$$

Here, $\boldsymbol{F}$ denotes the measured data on the number of people and F_i denotes the total number of people in the 15-minute period.

The complex facility where these measurements were taken consists of several floors with different amenities, including a restaurant floor, an event floor, and an office floor. Therefore, one of the characteristics of the location is that the attributes of the people included in the measurements are different on weekdays and weekends.



(a) Number of people (b) Number of tweets

Figure 2. Number of people measured at the office complex in Akihabara and tweets containing the keyword Akihabara

2.2. Measuring the Number of Tweets

We used Twitter API to retrieve tweets. We determined keywords in advance and continuously retrieved tweets and retweets containing these keywords. We measured the number of tweets $\boldsymbol{T}$ in a similar manner to the number of people by storing this number T_i as time-series data in 15-minute intervals and using the total number of tweets and retweets as the measured value for each time step.

$$\boldsymbol{T} = \{T_1, T_2, T_3, \cdots, T_i, \cdots, T_n\}. \tag{2}$$

Here, $\boldsymbol{T}$ denotes the measured data on the number of tweets, and T_i denotes the total number of tweets in the 15-minute period.

2.3. Relationship between the Number of People and the Number of Tweets

In order to discuss the relationship between the number of people and the number of tweets, we will focus on the correlation coefficient of these two variables. The data used in the study was collected in March and April 2014 and, considering shop opening times, the data focused on will be that collected between 10 am and 10 pm. The average number of people measured in the complex facility and the number of tweets are shown in Figure 2. The correlation coefficient was calculated using the total number of people and tweets during the measurement periods for each day of the experiment. Since the number of people was measured in Akihabara, we set the following six keywords (city name in Japan) for retrieving tweets: Akihabara; Tsukuba; Asakusa; Ueno; Shinjyuku; and Shinagawa. The correlation coefficient and correlation diagram between the number of people and the number of tweets/retweets containing each keyword is shown in Table 1 and Figure 3.

In the case of tweets, the results confirm that the correlation coefficient with the number of tweets contain-

Table 1. Correlation coefficients between the number of passers by and the number of tweets/retweets.

	Tweets	ReTweets
Akihabara	**0.804**	0.171
Tsukuba	0.722	0.182
Asakusa	0.700	0.093
Ueno	0.691	0.094
Shinjyuku	0.650	−0.020
Shinagawa	0.422	−0.080

ing the keyword "Akihabara" is high at 0.804, indicating a strong correlation. On the other hand, the highest correlation coefficient with the number of tweets for other keywords is low at 0.422, indicating a weaker correlation. It can be deduced that high correlation with tweets containing the keyword "Akihabara" occurred because the measurements were taken at a complex facility in Akihabara.

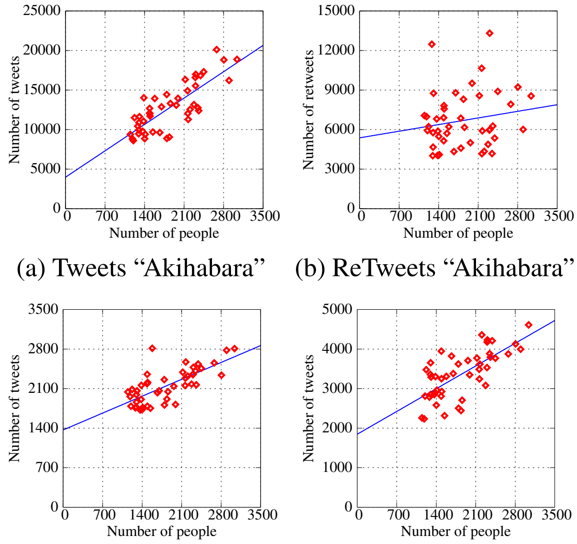
Next, we calculated the correlation coefficient for retweets. The correlation coefficient with retweets containing the keyword "Akihabara" was 0.171, a significantly weaker correlation than that with the number of tweets. Furthermore, we found that, in general, the correlation coefficient with the number of retweets was smaller than that with the number of tweets. This can be attributed to different nature of tweets and retweets. Since tweets are generally used by individuals to transmit information by themselves, they are closely linked to the behavior of posters and the situation they face. On the other hand, retweets constitute an action of diffusing information posted by someone else; thus, their content is not necessarily connected to the poster and location of the retweet. Therefore, correlation with the number of tweets is stronger than correlation with the number of retweets.

3. Mutual Interaction Model

In **2.3**, we clarified a strong correlation between the number of tweets containing specific keywords and the number of people in real space. In this section, we focus on the statistical relationship between the number of people and the number of tweets based on this result.

3.1. Outline of the Mutual Interaction Model

In this section, we propose a mutual interaction model as a model for expressing the statistical relationship between the number of people and number of tweets. The structure of the mutual interaction model is shown in Figure 4. The mutual interaction model is a graph model consisting of nodes and links. The nodes correspond to the number of people $\hat{F}_i$ and the number of tweets $\hat{T}_i$ at each time step. These values are calculated based on the formulas presented below. The nodes are connected by weighted links and the statistical relationship is expressed by setting the weight of the links on basis of the strength of influence between the nodes. The parameters contained in the model have the following meanings:



(a) Tweets "Akihabara" (b) ReTweets "Akihabara"

(c) Tweets "Tsukuba" (d) Tweets "Asakusa"

Figure 3. Correlation diagram between the number of people in Akihabara and the number of tweets/retweets.

$\tilde{F}_i$: Stationary number of people at each time step

$\tilde{T}_i$: Stationary number of tweets at each time step

$\alpha_{j,k}$: Strength of influence of T_j on F_k $(k \geq j)$

$\beta_{j,k}$: Strength of influence of F_j on T_k $(k \geq j)$

$\gamma_{j,k}$: Strength of influence of F_j on F_k $(k \geq j)$

$\delta_{j,k}$: Strength of influence of T_j on T_k $(k \geq j)$

In other words, α represents the influence of the number of tweets on the number of people and β represents the influence of the number of people on the number of tweets. In the proposed mutual interaction model, the number of people $\hat{F}_i$ and the number of tweets $\hat{T}_i$ at

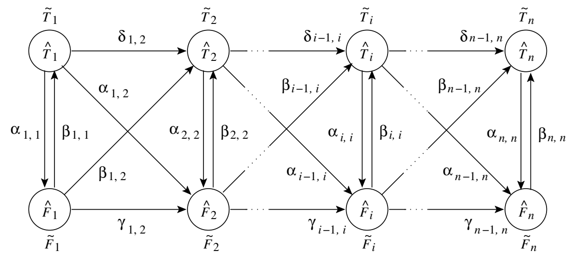


Figure 4. Structure of the mutual interaction Model

each time step are influenced by previous time steps but not by later time steps. In the present study, the model is simplified by applying the following constraint to j.

$$k \geq j \geq k - 1. \tag{3}$$

The constraint expressed by Eq.(3) presupposes that the number of people $\hat{F}_i$ and the number of tweets $\hat{T}_i$ at a certain time step are only influenced by that time step and the previous one time step. $\tilde{F}_i$ and $\tilde{T}_i$ represent the stationary number of people and tweets, respectively, and their values vary according to time. The stationary values are not affected by the number of people or tweets in other times steps and denote the number of people and tweets that are presumed to exist at each time step to begin with.

The number of people $\hat{F}_i$ and number of tweets $\tilde{T}_i$ at an arbitrary time is calculated based on Eq.(4) and (5), respectively.

$$\hat{F}_i = \tilde{F}_i + \gamma_{i-1,i} \cdot F_{i-1} + \alpha_{i-1,i} \cdot T_{i-1} + \alpha_{i,i} \cdot T_i, \tag{4}$$
$$\hat{T}_i = \tilde{T}_i + \delta_{i-1,i} \cdot T_{i-1} + \beta_{i-1,i} \cdot F_{i-1} + \beta_{i,i} \cdot F_i. \tag{5}$$

3.2. Method for Estimating Parameters

Each parameter in the mutual interaction model is estimated by regression analysis of the number of passers-by actually measured F_i and the number of tweets T_i. We defined Eq.(6) and (7) as objective functions and conducted analysis using the least-squares method.

$$f_F = \sum_{i=1}^{n} \left(F_i - \hat{F}_i\right)^2, \tag{6}$$

$$f_T = \sum_{i=1}^{n} \left(T_i - \hat{T}_i\right)^2. \tag{7}$$

The square error for each time step was calculated and the parameters determined so that the sum of the squares would be minimized.

4. Experiments

To verify the effectiveness of the mutual interaction model, we conducted an experiment using actual measurements of the number of people and the number of tweets.

4.1. Experimental Environment

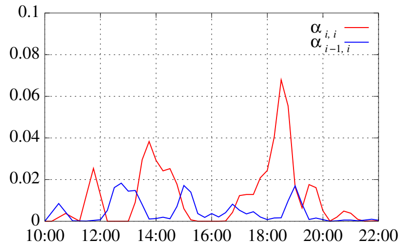
In the verification experiment, we utilized the data from March and April 2014 and, considering shop opening times, focused on the data collected between 10 am and 10 pm. Since the attributes of the people visiting the complex facility in Akihabara where the measurements were taken would likely vary between weekdays and weekends, we analyzed the data for weekdays and weekends separately. Having discarded the measured data that was incomplete due to system failure, we were left with 26 days of weekday data and 14 days of weekend data. In this paper, we discuss the results using the weekday and weekend measurement data. The model is evaluated by cross-validation. That is, we estimated the parameters using the data remaining after one day of data was excluded from the measured data set and performed the verification using the measured data from the excluded day. In other words, we estimated the parameters using measured data for 25 weekdays and 13 weekend days.

4.2. Results of the Analysis of Weekday Data

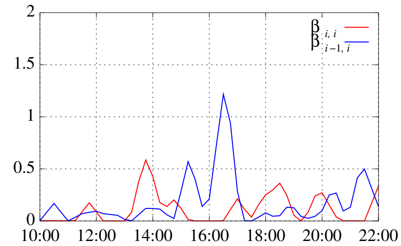
In this section, we discuss the results of the verification experiment using the weekday measurement data. The results described here refer to the average values for the results of 26 groups of data obtained using cross-validation.

We calculated the average error rate between the values obtained from Eq.(4) and (5) and the measured values. The error rate for the number of people was approximately 11.95% and the error rate for the number of tweets was approximately 4.37%. The mutual interaction model proposed in this paper can be seen to capture the statistical relationship between the number of people and number of tweets on weekdays.

The strengths of the influence of the number of tweets on the number of people (α) and the number of people on the number of tweets (β) are shown in Figure 5 (a) and (b), respectively. It is important to note that the data sets used in both Figure 5 (a) and (b) were smoothed by filtering. On each graph, the solid line represents the strength of influence from the current time step and the dotted line represents that from the previous time step. We will now focus on the size of the values.

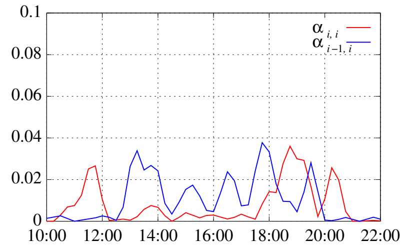


(a) Effect of tweets on people (weekdays)

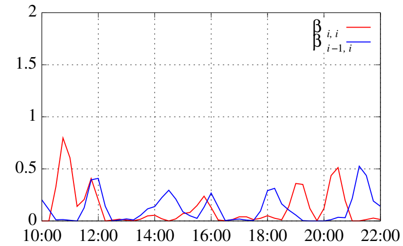


(b) Effect of people on tweets (weekdays)

Figure 5. Strength of the effect between the number of people and the number of tweets (weekdays).



(a) Effect of tweets on people (weekends)



(b) Effect of people on tweets (weekends)

Figure 6. Strength of the effect between the number of people and the number of tweets (weekends).

Several peaks can be seen in Figure 5. As mentioned in **2.3**, the relationship between the number of people and the number of tweets is statistical and we cannot assert a direct relationship between the two variables. Therefore, we will now examine the factors underlying this statistical relationship in light of the structural characteristics of the complex facility.

First, we will focus on the peak of $\alpha_{i,i}$ that occurs around 11:45 am. It can be assumed that one factor causing this peak is the movement of people visiting the restaurants in the complex facility. Generally, the number of people visiting restaurants in the complex facility increases as more people visit Akihabara. If we assume that the number of tweets containing the keyword "Akihabara" corresponds to the level of liveliness in the Akihabara district, the finding that the number of people at lunch time increases in proportion to the number of tweets is reasonable.

Next, the peaks of $\alpha_{i,i}$ and $\beta_{i,i}$ that occur at around 2:00 pm are different in character to the above-mentioned peak in that they both occur around the same time. One factor causing such peaks concerns the structural characteristics of the complex facility, which contains an event floor. When events take place in the complex facility, a considerable amount of information is disseminated by event participants in the form of tweets. Therefore it may be assumed that, in this case, the number of people influences the number of tweets. The dissemination of tweets serves to notify more people about the event, and as a result, the number of tweets ends up influencing back the number of people.

Similarly to the 11:45 am peak, one of the factors causing the peak of $\alpha_{i,i}$ at 6:30 pm is people visiting the complex facility to eat dinner. Furthermore, since a peak of $\beta_{i,i}$ also occurs at around the same time, it is reasonable to assume that, like the 2:00 pm peak, this peak is also influenced by events taking place at the complex facility.

The peak of $\beta_{i-1,i}$ at 9:15 pm can also be attributed to people using the restaurants in the complex facility; however, this peak is different from the other peaks caused by people dining earlier in the day in that there is no corresponding peak of $\alpha_{i,i}$ or $\alpha_{i-1,i}$. In other words, at this time of the day, people do not move as a result of viewing tweets. This is only natural considering that the movement of people subsides later in the evening.

4.3. Results of the Analysis of Weekend Data

In this section, we discuss the results of the verification experiment using the weekend measurement data. The results described here refer to the average values for the results of 14 groups of data obtained using cross-validation. We calculated the average error rate between the values obtained from Eq.(4) and (5) and the measured values and found that the rate for the number of people was approximately 10.04% and that for the number of tweets was approximately 3.61%. Similarly to the results of the verification experiment using the weekday data, these results show that the mutual interaction model proposed in this paper captures the statistical relationship between the number of people and number of tweets on weekends. The strengths of the influence of the number of tweets on the number of people (α) and the number of people on the number of tweets (β) are shown in Figure 6 (a) and (b), respectively.

First, a comparatively large peak of $\beta_{i,i}$ occurs around 10:45 am and a corresponding peak of $\alpha_{i,i}$ can also be observed, albeit smaller. The fact that the number of people and the number of tweets are influencing each other at around the same time suggests that these peaks are caused by events taking place at the complex facility. Next, a peak of $\alpha_{i,i}$ can be seen to occur at 11:45 am. As was the case on weekdays, one factor causing this peak can be thought to be movement of people visiting the restaurants in the complex facility. Furthermore, peaks of $\beta_{i,i}$ and $\beta_{i-1,i}$ can also be observed at around the same time and it is reasonable to assume that these peaks are also caused by people visiting restaurants to eat lunch. The content of tweets containing the keyword "Akihabara" circulating in virtual space is diverse: some people tweet about events taking place in Akihabara and some people in Akihabara tweet about themselves. Since tweets of the latter nature are considered to increase at lunch time in proportion to the number of people, the result that the influence of the number of people on the number of tweets is comparatively large is reasonable.

Peaks of $\alpha_{i-1,i}$ and $\beta_{i-1,i}$ occur at around 6:00 pm and peaks of $\alpha_{i,i}$ and $\beta_{i,i}$ at around 7:00 pm and 8:30 pm. As with the 11:45 am peak, one of the factors causing the above peaks can be considered to be the number of people eating dinner or participating in drinking parties. The peak of $\beta_{i-1,i}$ at 9:15 pm can also be attributed to the number of people using restaurants for the same reason as that given for weekdays.

5. Conclusion

In this paper, we proposed a mutual interaction model consisting of nodes and links with the aim of analyzing the statistical relationship between the number of people and the number of tweets. The mutual interaction model represents the number of people and the number of tweets as time-series data and expresses the statistical relationship between the two variables at each time step through weighted links connecting nodes to which the number of people and number of tweets are assigned. We conducted an experiment that aimed to verify the effectiveness of this mutual interaction model and analyze the statistical relationship between actual measurements of the number of people and number of tweets. In the verification experiment, we used the weekday data for the number of people measured in March and April 2014 at a complex facility near Akihabara Station and the number of tweets containing the keyword "Akihabara" posted during the same period. Considering the possibility that the attributes of visitors to the complex facility in Akihabara might vary between weekdays and weekends, we analyzed the data for weekdays and weekends separately. With the aim of verifying the effectiveness of the mutual interaction model, we began by calculating the average error rate between the values calculated using the model and the measured values. The average error rate for the number of people was approximately 11.95% for weekdays and approximately 10.04% for weekends and the rate for the number of tweets was approximately 4.37% for weekdays and approximately 3.61% for weekends, confirming that the mutual interaction model captures the statistical relationship between the number of people and the number of tweets.

References

[1] Haji Mohammad Saleem, Yishi Xu and Derek Ruths: "Effects of Disaster Characteristics on Twitter Event Signature," Vol.78, pp.165–172 (2014).

[2] A.J. Morales, J. Borondo, J.C. Losada and R.M. Benito: "Efficiency of human activity on information spreading on Twitter," Social Networks, Vol.39, pp.1–11 (2014).

[3] Tatsuro Kawamoto: "A stochastic model of tweet diffusion on the Twitter network," Physica A: Statistical Mechanics and its Applications, Vol.392, No.16, pp.3470–3475 (2013).

[4] J. Ko, H.W. Kwon, H.S. Kim, K. Lee and M.Y. Choi: "Model for Twitter dynamics: Public attention and time series of tweeting," Physica A: Statistical Mechanics and its Applications, Vol.404, pp.142–149 (2014).

[5] Luca Cagliero, Tania Cerquitelli, Paolo Garza and Luigi Grimaudo: "Twitter data analysis by means of Strong Flipping Generalized Itemsets," Journal of Systems and Software, Vol.94, pp.16–19 (2014).

[6] M. Onishi: "Analysis and Visualization of Large-Scale Pedestrian Flow in Normal and Disaster Situations," ITE Trans. on Media Technology and Applications, Vol.3, No.3, pp.170–183, (2015).