An analysis of CGM contents pageview using SIR Model and GBM

Noguchi, Kazuhisa
Graduate School of Integrated Frontier Sciences, Kyushu University

Iida, Tomoya
Department of Electrical Engineering and Computer Science, Kyushu University

Ito, Eisuke
Research Institute for Information Technology, Kyushu University

出版情報：Proceedings of the International Conference on Compute and Data Analysis, pp.19-22, 2017-05-19
バージョン：
権利関係：

An analysis of CGM contents pageview using SIR Model and GBM

Kazuhisa Noguchi
Graduate School of Integrated Frontier Sciences
Kyushu University
Fukuoka, Japan
noguchi.k.567@s.kyushu-u.ac.jp

Tomoya Iida
Department of Electrical Engineering and Computer Science, Kyushu University
Fukuoka, Japan
t.iida.630@s.kyushu-u.ac.jp

Eisuke Ito
Research Institute for Information Technology
Kyushu University
Fukuoka, Japan
ito.eisuke.523@cc.kyushu-u.ac.jp

***Abstract*—In consumer generated media (CGM) site, such as YouTube and nicovideo, only few contents are viewed very much, but most contents are only viewed few times. Our research target CGM sites are nicovideo.jp and syosetu.com. Nicovideo.jp is a popular movie CGM site in Japan and syosetu.com is the largest novel CGM site in Japan. We already found that pageview distribution of contents in both CGM sites follow a lognormal distribution. In this paper, we consider user's content selection model which will lead lognormal distribution. We apply Geometric Brownian Motion model into SIR model. SIR model is used for simulation of population transition process or epidemic process of infection disease. In this paper, we report the results of some simulation.**

Keywords-component; CGM; contents pageview; SIR Model; Stochastic Differential Equation; Geometric Brownian Motion

I. INTRODUCTION

In Consumer Generated Media (CGM) site, such as YouTube and nicovideo, only few contents are viewed very much, but most contents are only viewed few times. Recently, some article mentioned that diversity contents and genre are decreasing in some CGM sites. We found the diversity of contents quantitatively, for actual two CGM sites, nicovideo.jp and syosetu.com [1, 2].

We also found that pageview distribution of contents in both CGM sites follow a lognormal distribution [3]. Geometric Brownian Motion is a stochastic process that follows a lognormal distribution. It is SDE that has stochastic process as solution.

In this paper, to analyze the transition of GCM, we use SIR Model to predict epidemic process due to population fluctuation. An ordinary SIR model is given by a simultaneous ordinary differential equation [5]. When an initial value is determined, a solution is determined definitively. In general, however, it is more natural to assume that the prediction result is a random motion. In order to reproduce random motion, we introduce stochastic differential equations (SDE) into the SIR model. Relationship studies of SIR Model and stochastic differential equations [6, 7, 8] are as follows TABLE I. SDE type SIR model is mainly used for epidemic prediction of infection disease.

TABLE I. RELATED RESEARCH OF SDE TYPE SIR MODEL

Authors	SDE Model	year
Skiadas	Sigmoid Curve	2009
Maki et al.	SIR Model	2013
Cai et al.	SIRS Model	2013

We introduce Geometric Brownian Motion into SIR model. We regard the infection spread model as a model of spread of epidemics. We aim to probabilistically estimate the transition of the number of contents pageviews.

The rest of this paper is organized as follows. Section II describes Lognormal Distribution and Geometric Brownian Motion (GBM). In section 3, we describe SIR Model and its parameter estimation. In section 4, we propose SIR model in geometric Brownian motion. In section 5 we simulate the proposed model. Finally, we conclude this paper in section 6.

II. GEOMETRIC BROWNIAN MOTION (GBM)

A. Lognormal Distribution

When a random variable Y is distributed normally with Y's expectation μ and standard variance σ^2, Normal Distribution Probability Density Function (PDF) $g(y)$ is

$$g(y;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right).$$

Then $\ln x$, Lognormal Distribution PDF $f(x)$ is

$$f(x;\mu,\sigma^2)=g(\ln x)\frac{dy}{dx}=\frac{1}{\sqrt{2\pi}\sigma x}\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right).$$

B. Geometric Brownian Motion (GBM)

Stochastic process $\mathrm{X}(\mathrm{t}_i)$ follows Geometric Brownian Motion (GBM) if SDE is described as

$$d\mathrm{X}(\mathrm{t}_i) = \mu_X(\mathrm{t}_i)\mathrm{X}(\mathrm{t}_i)dt_i + \sigma_X(\mathrm{t}_i)\mathrm{X}(\mathrm{t}_i)dB(\mathrm{t}_i)$$

where $\mu_X(\mathrm{t}_i)$ (drift term), $\sigma_X(\mathrm{t}_i)$ (diffusion term) are parameter and $dB(\mathrm{t}_i)$ is Standard Brownian Process.

We rewrite the previous equation.

$$\frac{d\mathrm{X}(\mathrm{t}_i)}{\mathrm{X}(\mathrm{t}_i)} = \mu_X(\mathrm{t}_i)dt_i + \sigma_X(\mathrm{t}_i)dB(\mathrm{t}_i)$$
$$d(\ln \mathrm{X}(\mathrm{t}_i)) = \mu_X(\mathrm{t}_i)dt_i + \sigma_X(\mathrm{t}_i)dB(\mathrm{t}_i)$$

This equation shows the PDF of $\mathrm{X}(\mathrm{t}_i)$ is Lognormal Distribution.

We can regard the GBM's Stochastic process $\mathrm{X}(\mathrm{t}_i)$ as $\mathrm{X}(\mathrm{t}_i)$ is Lognormally distributed expectation $\mu_X(\mathrm{t}_i)$ and variance $\sigma_X(\mathrm{t}_i)$.

III. SIR Model

A. Simultanious differntial equations Model

SIR Model has three groups, $\mathrm{S}(\mathrm{t})$, $\mathrm{I}(\mathrm{t})$, $\mathrm{R}(\mathrm{t})$ at time t. $\mathrm{S}(\mathrm{t})$ is the number of susceptible, who are not yet infected but could become infected. $\mathrm{I}(\mathrm{t})$ is the number of infective, who have the disease and can transmit them to susceptible. $\mathrm{R}(\mathrm{t})$ is the number of removed or recovered people from the disease.

SIR Model is the following Simultaneous differential equations.

$$\begin{cases} \dfrac{d\mathrm{S}(\mathrm{t})}{dt} = -\beta \mathrm{S}(\mathrm{t})\mathrm{I}(\mathrm{t}) \\ \dfrac{d\mathrm{I}(\mathrm{t})}{dt} = \beta \mathrm{S}(\mathrm{t})\mathrm{I}(\mathrm{t}) - \gamma \mathrm{I}(\mathrm{t}) \\ \dfrac{d\mathrm{R}(\mathrm{t})}{dt} = \gamma \mathrm{I}(\mathrm{t}) \end{cases}$$

β is infection rate, which is the contact rate of the disease. γ is removal rate, which shows the mean death(removed) or recovery rate. The sum of S, I, R is total population N.

$$\mathrm{S}(\mathrm{t}) + \mathrm{I}(\mathrm{t}) + \mathrm{R}(\mathrm{t}) = N$$

B. Parameter estimation β, γ

SIR model parameters, β and γ are estimated by two methods.

The first method is forward difference as follows.

$$\beta(\mathrm{t}_i) = \frac{\mathrm{S}(\mathrm{t}_i) - \mathrm{S}(\mathrm{t}_{i+1})}{\mathrm{S}(\mathrm{t}_i)\mathrm{I}(\mathrm{t}_i)},\ \gamma(\mathrm{t}) = \frac{R(\mathrm{t}_{i+1}) - \mathrm{R}(\mathrm{t}_i)}{\mathrm{I}(\mathrm{t}_i)}$$

Second is backward difference as follows.

$$\beta(\mathrm{t}_i) = \frac{\mathrm{S}(\mathrm{t}_{i-1}) - \mathrm{S}(\mathrm{t}_i)}{\mathrm{S}(\mathrm{t}_i)\mathrm{I}(\mathrm{t}_i)},\ \gamma(\mathrm{t}) = \frac{R(\mathrm{t}_i) - \mathrm{R}(\mathrm{t}_{i-1})}{\mathrm{I}(\mathrm{t}_i)}$$

$\widehat{\beta}$ and $\widehat{\gamma}$ are expectation of β and γ.

$$\widehat{\beta} = \mathrm{E}[\beta(\mathrm{t}_i)], \widehat{\gamma} = \mathrm{E}[\gamma(\mathrm{t}_i)]$$

IV. SIR model into Geometric Brownian Motion

We apply SIR model into Geometric Brownian Motion. We try to probabilistically estimate the transition of the number of contents pageviews.

A. Simultanious differntial equations Model

We suggest SIR Model's $\mathrm{I}(\mathrm{t})$ term into diffusion term.

$$\begin{cases} d\mathrm{S}(\mathrm{t}) = -\beta \mathrm{S}(\mathrm{t})\mathrm{I}(\mathrm{t})dt \\ d\mathrm{I}(\mathrm{t}) = \{\beta \mathrm{S}(\mathrm{t}) - \gamma\}\mathrm{I}(\mathrm{t})dt + \sigma_I \mathrm{I}(\mathrm{t})dB(\mathrm{t}_i) \\ d\mathrm{R}(\mathrm{t}) = \gamma \mathrm{I}(\mathrm{t})dt \end{cases}$$

σ_I is diffusion term of $\mathrm{I}(\mathrm{t}).\mathrm{I}(\mathrm{t})$ term follows GBM. $\mathrm{S}(\mathrm{t}), \mathrm{I}(\mathrm{t}), \mathrm{R}(\mathrm{t})$ are Stochastic process.

B. Parameter estimation σ

Parameters in stochastic model are estimated as follows.

$$\mathrm{dI}(\mathrm{t}_i) = \{\beta \mathrm{S}(\mathrm{t}_\mathrm{i}) - \gamma\}\mathrm{I}(\mathrm{t}_\mathrm{i})d\mathrm{t}_\mathrm{i} + \sigma_\mathrm{I}\, \mathrm{I}(\mathrm{t}_i)\mathrm{d}B(\mathrm{t}_i)$$
$$\{\mathrm{dI}(\mathrm{t}_i)\}^2 = \{\{\beta \mathrm{S}(\mathrm{t}_\mathrm{i}) - \gamma\}\mathrm{I}(\mathrm{t}_\mathrm{i})d\mathrm{t}_\mathrm{i} + \sigma_\mathrm{I}\, \mathrm{I}(\mathrm{t}_i)dB(\mathrm{t}_i)\}^2$$

Then by Itô's lemma,

$$dB(\mathrm{t}_i)d\mathrm{t}_\mathrm{i} = d\mathrm{t}_\mathrm{i}dB(\mathrm{t}_i) = (d\mathrm{t}_\mathrm{i})^2 = 0,$$
$$dB(\mathrm{t}_i) = \sqrt{d\mathrm{t}_\mathrm{i}},$$

$\{\mathrm{dI}(\mathrm{t}_i)\}^2$ can be transformed to

$$\{\mathrm{dI}(\mathrm{t}_i)\}^2 = \{\sigma_\mathrm{I}\mathrm{I}(\mathrm{t}_i)\mathrm{d}B(\mathrm{t}_i)\}^2$$
$$= \sigma_\mathrm{I}^2\{\mathrm{I}(\mathrm{t}_i)\}^2\mathrm{d}\mathrm{t}_i.$$

If we assume

$$\mathrm{dI}(\mathrm{t}_i) \cong \mathrm{I}(\mathrm{t}_{i+1}) - \mathrm{I}(\mathrm{t}_i), \mathrm{dt}_i = \mathrm{t}_{i+1} - \mathrm{t}_i,$$

then diffusion term σ_I is

$$\sigma_{\mathrm{I}} = \frac{\mathrm{dI}(\mathrm{t}_i)}{\mathrm{I}(\mathrm{t}_i)\sqrt{\mathrm{dt}_i}} \cong \frac{\mathrm{I}(\mathrm{t}_{i+1}) - \mathrm{I}(\mathrm{t}_i)}{\mathrm{I}(\mathrm{t}_i)\sqrt{\mathrm{t}_{i+1} - \mathrm{t}_i}}.$$

σ_{I}'s estimator, $\hat{\sigma}_{\mathrm{I}}$ is given by

$$\hat{\sigma}_{\mathrm{I}} \cong \frac{1}{\mathrm{n}-1}\sum_{i=1}^{\mathrm{n}-1} \frac{\mathrm{I}(\mathrm{t}_{i+1}) - \mathrm{I}(\mathrm{t}_i)}{\mathrm{I}(\mathrm{t}_i)\sqrt{\mathrm{t}_{i+1} - \mathrm{t}_i}}.$$

V. SIMULATION

A. *nicohart dataset*

Nicovideo.jp (*Nico-Nico Douga*, in Japanese) is a popular movie sharing CGM in Japan. Nicovideo started on December 12, 2006. Nicochart is a site which nicovideo data includes Video Ranking, Total number of videos, Total number of views of video, Total comments on video [4].

In this paper, we used the number of videos and their views data in each day by date from June 19, 2007 to December 1, 2016.

B. *Prediction using GBM-SIR Model*

We predict the total number of videos and their views data in each day using SIR Model. After that we simulate their data using GBM-SIR Model.

Since SIR Model has a deterministic solution with simultaneous ordinary differential equations, errors of prediction occur. On the other hand, the SIR model introducing the GBM calculates a random solution by diffusion term. Therefore, since the upper and lower limits of the estimated value calculated by the GBM-SIR model an estimation interval of the measured value is obtained.

To predict the total number of videos and the total number of views of videos data in each day corresponds to the prediction of SIR Model's R. We select the parameters based on observation data.

C. *The number of videos*

We predict the number of videos using SIR model and GBM-SIR model. Initial values of each parameter of SIR and GBM-SIR Model are as shown in TABLE II .

Figure 1 shows the SIR model and Figure 2 shows the simulation results by the GBM-SIR model. The pink line is the observation value of the number of videos, and the blue line is the estimation curve of the number of videos.

TABLE II. SIR'S INITIAL VALUE OF THE NUMBER OF VIDEOS

Parameter	Initial values
N	2.0e+7
S	2.0e+7-1967
I	1967
R	0
β	9.91e-08
γ	1
σ_{I}	0.44

D. *The number of views of videos*

We predict the number of views of videos using SIR model and GBM-SIR model. Initial values of each parameter of SIR and GBM-SIR Model are as shown in TABLE III.

Figure 3 shows the SIR model and Figure 4 shows the simulation results by the GBM-SIR model. The pink line is the observation value of the number of views of videos, and the blue line is the estimation curve of the number of views of videos.

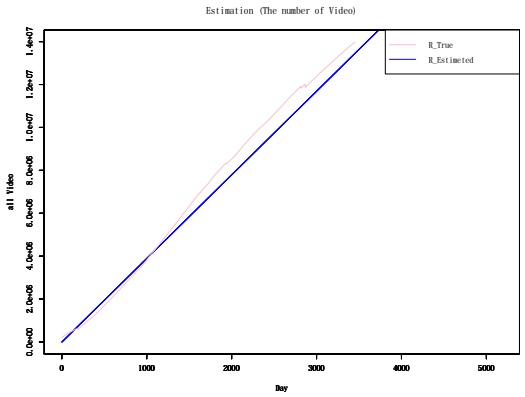

Figure 1. Prediction the number of videos using SIR model

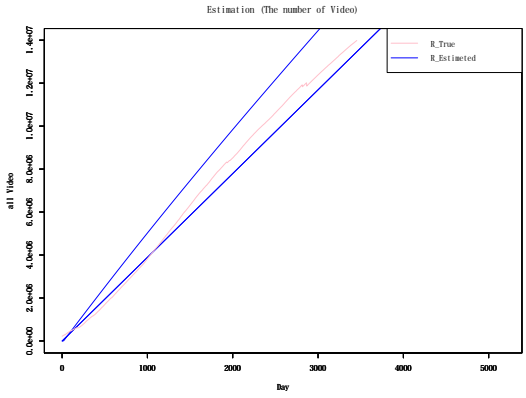

Figure 2. Prediction the number of videos using GBM-SIR model

TABLE III. SIR'S INITIAL VALUE OF THE NUMBER OF VIEWS OF VIDEOS

Parameter	Initial values
N	2.0e+11
S	2.0e+11-7338224
I	7338224
R	0
β	1.54e-11
γ	1
σ_1	0.47

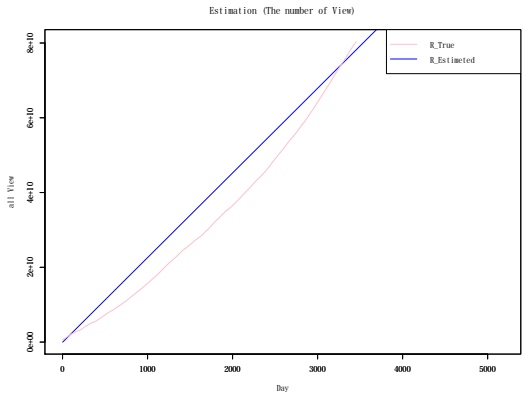

Figure 3. Prediction by SIR model the number of views of videos.

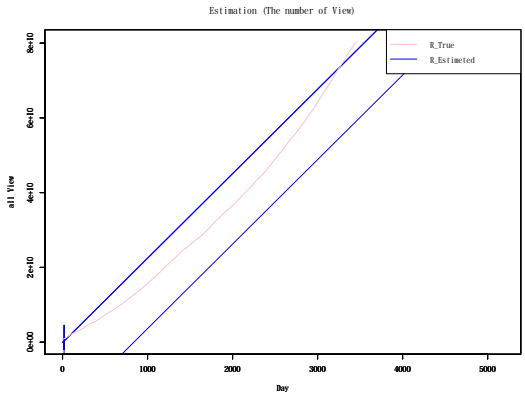

Figure 4. Prediction by GBM-SIR model the number of views of videos

VI. CONCLUSION

In this paper, we analyzed GCM in the SIR Model used to predict epidemic process due to population fluctuation. In order to introduce stochastic movement into SIR Model, we proposed Geometric Brownian Motion into SIR model (GBM-SIR Model). We regard the infection spread model as a model of spread of epidemics. We regard the infection spread model as a model of spread of epidemics. We simulated the transition of the number of video viewers and the total number of videos in Nicovideo.jp with the GBM-SIR Model.

In the future, we would like to predicts the transition of views of videos by analyzing the number of views per day for each video. In order to analyze the diversity of videos, classify the videos for each genre of videos and analyze the number of videos and the number of views of videos. we would like to consider a method of searching for the total population number N expected to view GCM contents and upload GCM contents. Furthermore, we evaluate the validity and estimate of GBM-SIR model.

REFERENCES

[1] Eisuke, I., Kazunori, S.: Frequency and link analysis of online novels toward social contents ranking, Proc. of SCA2012, pp.531-536, 2012.

[2] Kazunori, S., Eisuke, I., Sachio, H.: Predicting Future Ranking of Online Novels based on Collective Intelligence, Proc. of ICDIPC2013, SDIWC, pp.261-272, 2013.

[3] Kyohei, K., Eisuke, I.: A quantitative contents diversity analysis on a consumer generated media site, AROB 2016

[4] nicochart: http://www.nicochart.jp

[5] Kermack, W. O., McKendrick, A. G.: 1933. Contributions to the mathematical theory of epidemics- iii. further studies of the problem of endemicity. Proceedings of the Royal Society 141A, 94-122.

[6] Skiadas, C.H.:Exact Solution of Differential Equation: Gompertzand Generalized Logistic, Methodology and Computing in Applied Probability, Vol.12 (2009), pp.261270.

[7] Yoshihiro, M., Hideo, H.: Infectious Disease Spread Analysis Using Stochastic Differential Equations for SIR Model, " the 4th International Conference on Intelligent Systems, Modelling and Simulation (ISMS 2013), pp.152-156, 2013.

[8] Y. Cai, X. Wang, W. Wang, and M. Zhao: Stochastic Dynamics of an SIRS Epidemic Model with Ratio-Dependent Incidence Rate, Journal of Mathematical Biology Volume 2013 (2013), Article ID 172631, 11 pages.