2013 4th International Conference on Intelligent Systems, Modelling and Simulation

Seasonal Infectious Disease Spread Prediction Using Matrix Decomposition Method

Hideo Hirose, Toru Nakazono, Masakazu Tokunaga, Takenori Sakumura, Sirajummonira Sumi, Junaida Sulaiman
Department of Systems Design and Informatics
Kyushu Institute of Technology
Fukuoka, 820-8502 Japan
Email: hirose@ces.kyutech.ac.jp

***Abstract*—The matrix decomposition is one of the most powerful methods in recommendation systems. In the recommendation system, we can assume an incomplete matrix consisted of observed evaluation values by users and items, then we predict the vacant elements of the matrix using the observed values. This method is applied to a variety of the fields, e.g., for movie recommendations, music recommendations, book recommendations, etc. In this paper, we apply the matrix decomposition to predict the seasonal infectious disease spread. Applying the method to the case of infectious gastroenteritis caused by Norovirus in Japan, we have found that the early detection and prediction for the prevalence of the disease spread can be expected accurately. The infectious disease spread prediction using the matrix decomposition is new. To demonstrate the advantageous point and effectiveness of the matrix decomposition method, we applied the method to the influenza spread prediction in Japan, where missing observations are admitted for computation unlike other prediction methods.**

***Keywords*-matrix decomposition; recommendation system; disease spread; Norovirus; influenza; early detection; artificial neural networks; ensemble;**

I. Introduction

In late November 2012 in Japan, an intriguing news was announced; the infectious gastroenteritis caused by Norovirus will be widely spread similarly to the case in 2006, as shown in Figure 1. As seen in the figure on the top, Norovirus spread is seasonal, we can predict the trend by eyes to some extent by overlapping the yearly trends (see on the bottom in the figure). Looking at the 2012 trend, we can find out the similar trend to that in 2006, and we may expect a wide spread of the disease in 2012. It is now disclosed that the reason of this?prevalence?prediction is that the virus has been changed by mutation and the emerging phenomenon will be observed soon. However, without the assistance of numerical prediction methods, we cannot predict how large the spread is.

For infectious disease spread prediction, the SEIR method is well known as a classical mathematical model, where S, E, I, and R denote *s*usceptible, *e*xposed, *i*nfected and *r*emoved populations respectively [1], [6], [16]. This model computes the number of people infected with a contagious disease in a closed population over time, and can quickly deal with simulations of infectious disease spread among homogeneous populations using simple simultaneous ordinary differential equations and a few parameters. Many researchers successfully predicted the infectious disease spread, e.g., [24]. Other models are 1) agent-based models [2], [7], [8], [14], [15], which mimic the real city and human behavior in a computer, and simulates the disease spread by human contacts, 2) internet-based models [4], [5], [9], [12], which use the keywords related to infectious disease in early stages of emerging, 3) conventional statistical truncated model [10], [11], and 4) other models [13]. These models do not assume the seasonality.

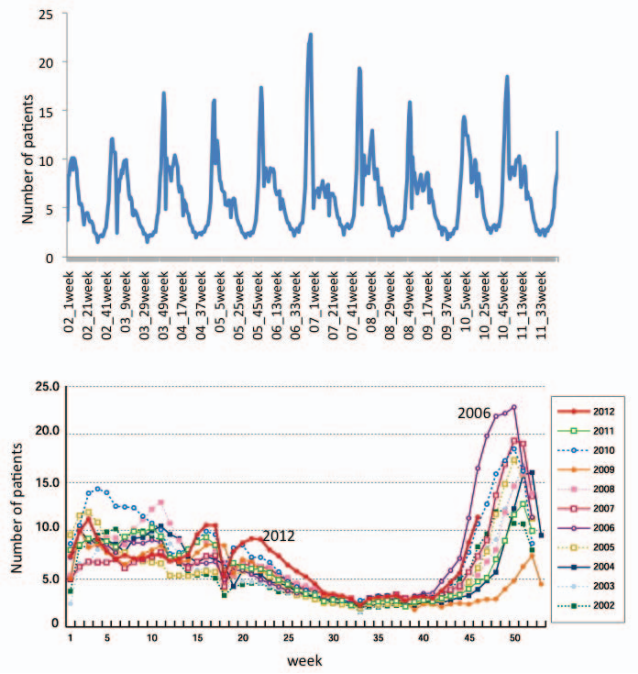

Figure 1. Infectious gastroenteritis spread caused by Norovirus in Japan in 2012. (Data from NIID Japan provided by designated institutions across the nation)

As seen in Figure 1, gastroenteritis spread caused by Norovirus can assume the seasonal effect, where we may use past data for many years. In such a case, time series analysis and other methods, e.g., neural network, have been used for a variety of cases; rainfall forecasting is one example [21],

[22]. However, the recommendation system has not been used for the infectious disease spread prediction. In this paper, we use the matrix decomposition method in seasonal infectious disease spread prediction. This is new.

II. Matrix Decomposition Method

The idea for the recommendation system by using the matrix decomposition is simple [23]. We consider an incomplete matrix consisted of observed evaluation values by users and items, then we predict the vacant elements of the matrix using the observed values. That is, the unobserved value for (i, j) element in the matrix is estimated (to $\hat{x}(i, j)$) by using the observed values of $x(i, j)$; see Figure 2.

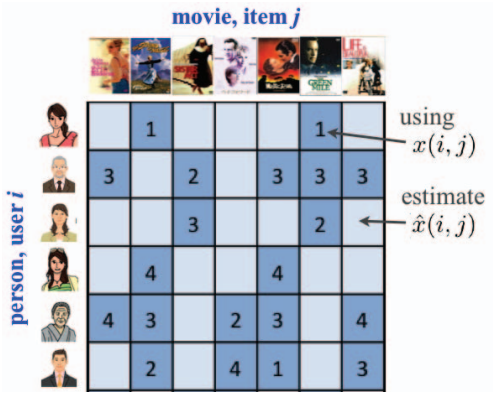

Figure 2. The idea for the recommendation system using the matrix decomposition.

A variety of methods have been proposed to solve this kind of problem. The use of the similarities such as the correlation coefficient or cosine is the primary method to estimate the vacant elements; k-nearest neighbor approach using the similarity is the next step; a promising methodology is the use of matrix decomposition (matrix factorization) [20]. Although the winner of the Netflix competition [17] used the ensemble technique combining all the known methods together [3], one of the most effective and major methods is still the matrix decomposition.

The singular-value decomposition, abbreviated as SVD, is one of the factorization algorithms for various applications which include computing the pseudo-inverse, least squares fitting of data, matrix approximation, and determining the rank, range and null space of a matrix. Suppose $P \in R^{m \times n}$, $U \in R^{f \times m}$, and $M \in R^{f \times n}$ are matrices. A simple idea that a matrix factorization $P = U^T M$ produces the missing data of score matrix V leads us to the use of the collaborative filtering. Thus, the matrix decomposition, which is also used for recommendation systems (see references [18], [19], [25]), is used for the least square method here. That is, we want to find the matrix U and M by minimizing the target function E such that sum of the squares of the difference between the observed score $V(i, j)$ and the predicted score $P(U_i, M_j)$,

$$E = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} I(i, j)(V(i, j) - P(U_i, M_j))^2, \quad (1)$$

where $P(U_i, M_j)$ denotes the (i, j) element of $U^T M$. This idea of the matrix decomposition is derived by the usual SVD formulation such that $A = U \Sigma V^*$ where U and V are orthonormal and Σ provides the singular values in the diagonal elements. If Σ is absorbed by either or both U and V, we can accomplish the matrix decomposition of A.

Suppose $V \in R^{m \times n}$ is the score matrix of m users and n items, and $I \in 0, 1^{m \times n}$ is its indicator. The SVD algorithm finds two matrices U and M as the feature matrix of users and items. That is, each user or item has an f-dimension feature vector and f is called the dimension of the SVD. A prediction function p is used to predict the values in V. The value of a score $V(i, j)$ is estimated by $p(U_i, M_j)$, where U_i and M_j represent the feature vector of user i and item j, respectively. Once U and M are found, the missing scores in V can be predicted by the prediction function.

For stable and robust computing, the optimization of U and M is actually performed by minimizing the sum of squared errors between the existing scores and their prediction values with penalty factors:

$$\begin{aligned} E &= \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} I(i, j)(V(i, j) - p(U_i, M_j))^2 \\ &+ \frac{k_u}{2} \sum_{i=1}^{m} || U_i ||^2 + \frac{k_m}{2} \sum_{j=1}^{n} || M_j ||^2, \end{aligned} \quad (2)$$

where k_u and k_m are regularization coefficients to prevent overfitting; $|| \cdot ||$ means the Frobenius norm (l^2 norm). This formulation is a kind of the ridge regressions. The most common prediction function is the dot product of feature vectors. That is, $p(U, M) = U^T M$. The optimization of U and M thus becomes a matrix factorization problem where $V \approx U^T M$.

When using the prediction function of $p(U, M) = U^T M$, the objective function and its negative gradients have the following forms:

$$-\frac{\partial E}{\partial U_i} = \sum_{j=1}^{n} I(i, j) \left((V(i, j) - p(U_i, M_j)) \frac{\partial p(U_i, M_j)}{\partial U_i} \right) - k_u U_i$$

$$-\frac{\partial E}{\partial M_j} = \sum_{i=1}^{m} I(i, j) \left((V(i, j) - p(U_i, M_j)) \frac{\partial p(U_i, M_j)}{\partial M_j} \right) - k_m M_j$$

One can then perform the optimization of U and M by the descent gradient method or stochastic descent gradient

method by using the algorithm,

$$U^{(t+1)} \leftarrow U^{(t)} + \mu \frac{\partial E}{\partial U}$$
$$M^{(t+1)} \leftarrow M^{(t)} + \mu \frac{\partial E}{\partial M},$$

where μ is the learning rate.

III. Infectious Disease Spread Prediction Using the Matrix Decomposition Method

The seasonal trend of the infectious disease spread to each year can be described in a matrix; weekly for column and year for row. For example in the case of the infectious gastroenteritis caused by Norovirus in Japan, we can collect data yearly from the first week of 2002 to the 46th week of 2012 as shown in Figure 3 which is the same as in Figure 1. The number of weekly data in a year is 52. This incomplete matrix is used to the matrix decomposition, and we can predict the unobserved elements in the matrix.

	week 1	week 2	week 3	···	week 46	week 47	···	week 52
2002	4.99	6.22	6.75	···	8.3	9.64	···	7.99
···	···	···	···	···	···	···	···	···
2009	4.98	10.23	8.24	···	2.68	2.86	···	7.39
2010	8.63	10.53	13.87	···	10.68	12.74	···	8.65
2011	7.98	8.49	9.16	···	4.7	5.09	···	9.97
2012	7.33	9.87	11.21	···	11.39	not yet available and predict		

Figure 3. The matrix data of the infectious gastroenteritis caused by Norovirus in Japan.

In Figure 4, on the top, we show the predicted result for week 6 to week 52 using the yearly data from 2002 to 2011 and from week 1 to week 5, and the similar predicted result for week 46 to week 52 on the bottom of the figure. We can see that a wide disease spread similar to the 2006 case can be predicted. Here, we have used the parameters such that $\mu = 5 \times 10^{-6}$, and $k_m = k_u = 2 \times 10^{-5}$, and $f = 10$.

We may think that the prediction range is restricted to the range of past history. Thus, we assumed here a case of artificial trend mimicked to year 2012 with the twice scale of the number of patients. We show this result in Figure 5. We can see that the prediction is not restricted to the past history. The parameters used are the same as above.

IV. Accuracy of the Prediction

To check if the proposed method provides a good accuracy or not, we have compared the $RMSE$ (root mean squared error) obtained by the matrix decomposition method with that by using the other methods; we have used the artificial neural networks (ANN) method in time series analysis and its extension with ensemble methods for comparison. The

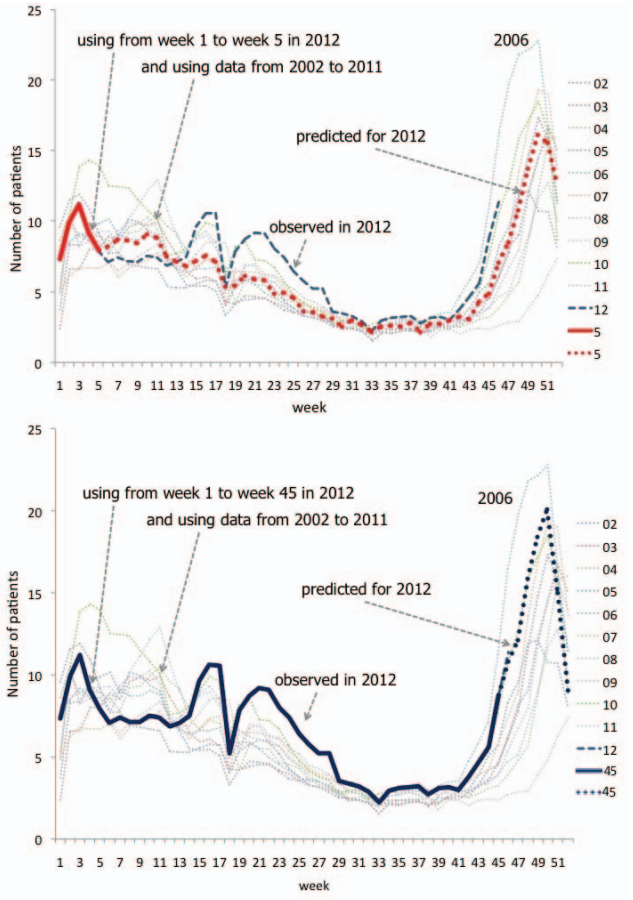

Figure 4. The predicted result for 2012 using the yearly data from 2002 to 2011.

$RMSE$ is defined by

$$RMSE = \sqrt{\frac{1}{|T|} \sum_{i,j} I(i,j)(\hat{x}(i,j) - x(i,j))^2},$$
$$(|T| = \sum_{i,j} I(i,j),\ I(i,j) \text{ for test data}). \quad (3)$$

We will use the data of 2002-2010 as training, and data of 2011 as test. However, in the case of the matrix decomposition method, we have to include at least one observation to 2011 row in the matrix. The cases of adding 1-5 weeks, 1-10 weeks, ... , 1-45 weeks data are considered here. The $RMSE$ computed in such a manner are shown in Figure 6. In the figure, we have shown four cases such that dimensions for SVD, f, are 2, 3, 4, 5. The larger the number of dimension, the smaller the value of the $RMSE$. The values of $RMSE$ are located in $[1.5, 2.5]$.

When we apply the method of ANN to predict the same case as in the case that 1-5 weekly data in 2011 and 2002-2010 are used as training and 6-52 weeks as test, the $RMSE$ becomes 2.3 which is larger than those by using

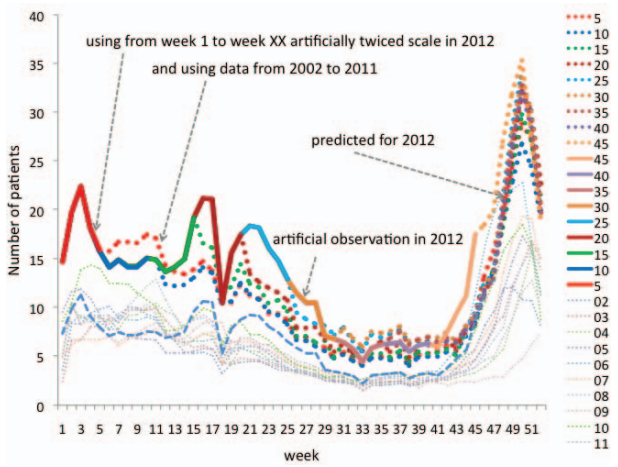

Figure 5. The predicted result for 2012 using the artificial trend mimicked to year 2012 with the twice scale of the number of patients.

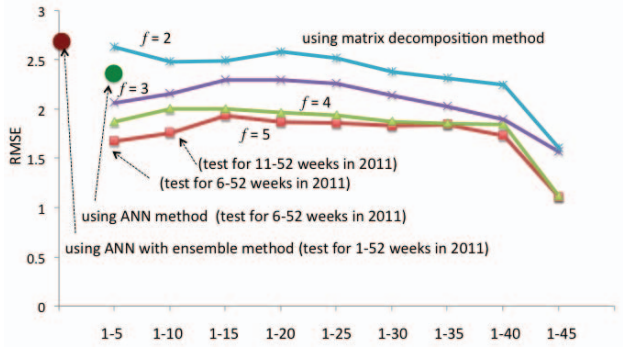

Figure 6. The $RMSE$ for accuracy of prediction by the matrix decomposition method. (1-5 means that we used 1-5 weekly data in 2011 as well as 2002-2010 as the observed data)

the matrix decomposition method for dimensions for SVD f are 3, 4, 5; see Figure 6. Here, the specification for the ANN detail [21] is, 1) arrange data such that x_1 and x_2 (as the inputs) are the previous one year and two years data for the week correspond to y (as the response), 2) build ANN model with 2-7-1 (input-hidden-output) nodes, 3) train ANN model with Lavenberg-Marquardt algorithm, while $\tanh$ activation function is adopted for hidden nodes whereas linear activation function is used for output nodes.

This shows the superiority of the proposed method over the conventional methods using the ANN. The predicted curve for 6-52 weeks as well as the observed curve in 2011 are shown in Figure 7.

Another comparison of the ANN with ensemble methods to the proposed method results a similar tendency. Using the method of the basic ensemble of ANN [22] with random selection of the weights and optimal lag was obtained by using the linear correlation analysis (LCA), the $RMSE$ value became 2.7 when 2002-2010 data were used for training and 2011 data as test.

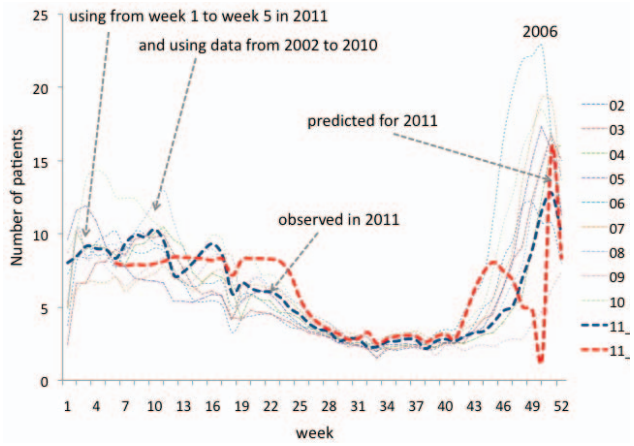

Figure 7. The predicted result for 2011 using the yearly data from 2002 to 2010. (Predicted by using the ANN)

V. Discussion

Unlike the time series analysis with the ANN and its relatives (i.e., ANN with ensemble methods), the matrix decomposition method need not require all the observed values in the matrix. Not only the elements for prediction but also the elements with missing observation can be admitted. Because of this, the matrix decomposition method is applicable to a variety of fields.

Figure 8 shows the influenza case in Japan from 1996 to 2012. There are many unobserved elements in the matrix. However, using the matrix decomposition method, we can predict the future trend of the influenza spread. Figure 9 shows the prediction for weeks in 2012 using the 1996-2011 yearly data and 1-5, 1-10, ... , 1-45 weekly data in 2012.

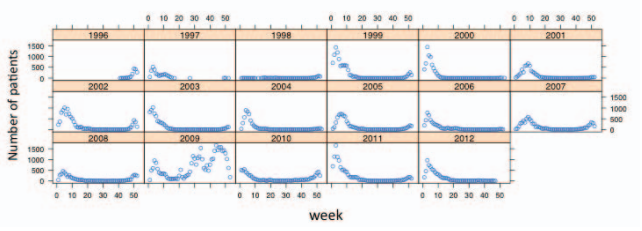

Figure 8. Influenza observation in Japan from 1996 to 2012.

VI. Concluding Remarks

To predict the vacant elements of the incomplete matrix using the observed values in the matrix, the matrix decomposition is one of the most promising methods, which is often used in the recommendation systems. This method is applied to a variety of the fields, e.g., for movie recommendations, music recommendations, book recommendations, etc. In this

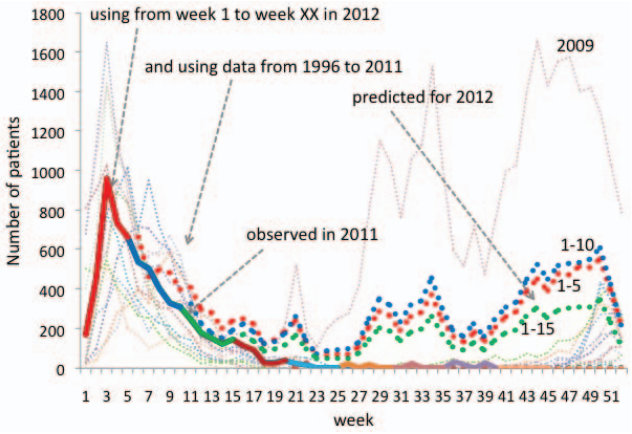

Figure 9. Influenza prediction in Japan 2012.

paper, we apply the matrix decomposition to predict the seasonal infectious disease spread. Applying the method to the case of infectious gastroenteritis caused by Norovirus in Japan, we have found that the early detection and prediction for the prevalence of the disease spread can be expected accurately. Comparing the root mean squared error between the predicted and observed data, we have found that the proposed method shows the superiority over the conventional methods using the method of artificial neural networks.The infectious disease spread prediction using the matrix decomposition is considered to be a new methodology. To demonstrate the advantageous point and effectiveness of the matrix decomposition method, we applied the method to the influenza spread prediction in Japan, where missing observations are admitted for computation unlike other prediction methods.

REFERENCES

[1] R. Anderson and R. May, Infectious diseases of humans: Dynamics and control, Oxford University Press, 1991.

[2] C.L. Barrett, S.G. Eubank and J.P. Smith, If smallpox strikes Portland, Scientific American, 292, pp. 54-61, 2005.

[3] R.M. Bell, J. Bennett, Y. Koren and C. Volinsky, The Million Dollar Programming Prize. IEEE Spectrum. May 2009.

[4] A. Culotta, Detecting influenza outbreaks by analyzing Twitter messages, Science, 16, May, pp. 1-11, 2010.

[5] A. Culotta, Towards detecting influenza epidemics by analyzing Twitter messages, Proceedings of the First Workshop on Social Media Analytics (SOMA'10), pp. 115-122, 2010.

[6] O. Diekmann and J.A.P. Heesterbeek, Mathematical epidemiology of infectious diseases: model building, analysis and interpretation, New York: Wiley, 2000.

[7] L.R. Elveback, J.P. Fox, E. Ackerman, A. Langworthy, M. Boyd, L. Gatewood, An influenza simulation model for immunization studies, American Journal of Epidemiology, 103, pp. 152-65, 1976.

[8] S. Eubank, Scalable, efficient epidemiological simulation, Proceedings of the 2002 ACM symposium on Applied computing, 139-145, 2002.

[9] J. Ginsberg, et. al., Detecting influenza epidemics using search engine query data, Nature 457, pp. 1012-1014, 2009.

[10] H. Hirose, The mixed trunsored model with applications to SARS, Mathematics and Computers in Simulation, vol. 74, pp. 443-453, 2007.

[11] H. Hirose, Estimation for the size of fragile population in the trunsored and truncated models with application to the confidence interval for the case fatality ratio of SARS, Information, vol. 12, pp. 33-50, 2009.

[12] H. Hirose, L. Wang, Prediction of Infectious Disease Spread using Twitter: A Case of Influenza, the 5th International Symposium on Parallel Architectures, Algorithms and Programming (PAAP'12), 2012. accepted

[13] H. Hirose, Estimation of the number of failures in the Weibull model using the ordinary differential equation, European Journal of Operational Research, Vol. 223, pp. 722-731, 2012.

[14] C.Y. Huang, C.T. Sun, J.L. Hsieh, Y.M.A. Chen and H.L. Lin, A novel small-world model: Using social mirror identities for epidemic simulations, Simulation, 81, pp. 671-699, 2005.

[15] I.M. Longini, Jr., M.E. Halloran, A. Nizam and Y. Yang, Containing pandemic influenza with antiviral agents, American Journal of Epidemiology, 159, pp. 623-633, 2004.

[16] W.O. Kermack and A.G. McKendrick, Contributions to the mathematical theory of epidemics-III: Further studies of the problem of endemicity, Proceedings of the Royal Society, 141A, pp. 94-122, 1933.

[17] Netflix prize, http://www.netflixprize.com/

[18] A. Paterek. Improving regularized Singular Value Decomposition for collaborative fitering. Proceedings of KDD Cup and Workshop, 2007.

[19] R. Salakhutdinov and A. Mnih, Probabilistic Matrix Factorization, Proc. Advances in Neural Information Processing. Systems 20 (NIPS 07), ACM Press, pp. 1257-1264, 2008.

[20] Netflix Update: Try This at Home. http://sifter.org/ simon/journal/20061211.html

[21] J. Sulaiman, H. Hirose, A Method to Predict Heavy Precipitation using the Artificial Neural Networks with an Application, 7th International Conference on Computing and Convergence Technology (ICCIT2012), pp. 687-691, 2012.

[22] S.M. Sumi, M.F. Zaman, H. Hirose, A rainfall forecasting method using machine learning models and its application to Fukuoka city case, International Journal of Applied Mathematics and Computer Science, Vol. 22, 2012. to appear

[23] S. Takimoto and H. Hirose, Recommendation systems and their preference prediction algorithms in a large-scale database, Information, Vol.12, No.5, pp. 1165-1182, 2009.

[24] Y. Toyosaka and H. Hirose, Pandemic simulations by MADE: a combination of multi-agent and differential equations, The 2009 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2009), pp. 331-335, 2009.

[25] S. Zhang, W. Wang, J. Ford, F. Makedon, and J. Pearlman. Using Singular Value Decomposition approximation for collaborative fitering. Seventh IEEE International Conference on E-Commerce Technology (CEC 2005), pp. 257-264, 2005.