ISSN : 2165-4069(Online)
ISSN : 2165-4050(Print)

IJARAI

International Journal of
Advanced Research in Artificial Intelligence

Volume 5 Issue 1

www.ijarai.thesai.org

A Publication of
The Science and Information Organization

Editorial Preface

From the Desk of Managing Editor…

Artificial Intelligence is hardly a new idea. Human likenesses, with the ability to act as human, dates back to Geek mythology with Pygmalion's ivory statue or the bronze robot of Hephaestus. However, with innovations in the technological world, AI is undergoing a renaissance that is giving way to new channels of creativity.

The study and pursuit of creating artificial intelligence is more than designing a system that can beat grand masters at chess or win endless rounds of Jeopardy!. Instead, the journey of discovery has more real-life applications than could be expected. While it may seem like it is out of a science fiction novel, work in the field of AI can be used to perfect face recognition software or be used to design a fully functioning neural network.

At the International Journal of Advanced Research in Artificial Intelligence, we strive to disseminate proposals for new ways of looking at problems related to AI. This includes being able to provide demonstrations of effectiveness in this field. We also look for papers that have real-life applications complete with descriptions of scenarios, solutions, and in-depth evaluations of the techniques being utilized.

Our mission is to be one of the most respected publications in the field and engage in the ubiquitous spread of knowledge with effectiveness to a wide audience. It is why all of articles are open access and available view at any time.

IJARAI strives to include articles of both research and innovative applications of AI from all over the world. It is our goal to bring together researchers, professors, and students to share ideas, problems, and solution relating to artificial intelligence and application with its convergence strategies. We would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations.

We hope that this journal will inspire and educate. For those who may be enticed to submit papers, thank you for sharing your wisdom.

Editor-in-Chief
IJARAI
Volume 5 Issue 1 January 2016
ISSN: 2165-4069(Online)
ISSN: 2165-4050(Print)

Editorial Board

Reviewer Board Members

CONTENTS

Comparative Study of Optimization Methods for Estimation of Sea Surface Temperature and Ocean Wind with Microwave Radiometer Data

Kohei Arai[1]
Graduate School of Science and Engineering
Saga University
Saga City, Japan

***Abstract*—Comparative study of optimization methods for estimation sea surface temperature and ocean wind with microwave radiometer data is conducted. The well known mesh method (Grid Search Method: GSM), regressive method, and simulated annealing method are compared. Surface emissivity is estimated with the simulated annealing and compared to the well known Thomas T. Wilheit model based emissivity. On the other hand, brightness temperature of microwave radiometer as a function of observation angle is estimated by the simulated annealing method and compares it to the actual microwave radiometer data. Also, simultaneous estimation of sea surface temperature and ocean wind speed is carried out by the simulated annealing and compared it to the estimated those by the GSM method. The experimental results show the simulated annealing which allows estimation of global optimum is superior to the other method in some extent.**

Keywords*—Microwave radiometer; remote sensing; sea surface temperature; nonlinear optimization theory; simulated annealing

I. INTRODUCTION

Microwave scanning radiometer allows estimation of geophysical parameters such as soil moisture, salinity, ocean wind, sea surface temperature, water vapor, cloud liquid, and so on with all weather conditions and in day and night basis [1]-[24]. Several microwave radiometers are carried on the several satellites and used for weather prediction and climate change research. One of the major concerns on the microwave radiometer is estimation accuracy of the geophysical parameters. Most of the methods for estimation of geophysical parameters are based on statistical models, regressive analysis. The estimation accuracy is not good enough because the regressive coefficients are determined with some observation conditions, areas of concerns, specific seasons. Therefore, the estimation accuracy is not good enough when the actual conditions are not matched to the conditions for the determination of regressive coefficients. Other than this, there is physical model based approaches. Through minimization processes between the actual acquired brightness temperature and the estimated brightness temperature derived from the model based method.

Microwave radiometer allows estimation of geophysical parameters such as water vapor, rainfall rate, ocean wind speed, salinity, soil moisture, air-temperature, sea surface temperature, cloud liquid, etc. based on least square method. Due to the fact that relation between microwave radiometer data (at sensor brightness temperature at the specified frequency) and geophysical parameters is non-linear, non-linear least square method is required for the estimations. Although there are some methods which allow estimation optimum solutions, Simulated Annealing: SA method [25] is just one method for finding global optimum solution.

Other methods, such as steepest descending method, conjugate gradient method, etc. gives one of local minima, not the global optimum solution. SA, on the other hand, requires huge computer resources for convergence. In order to accelerate the convergence process, not the conventional exponential function with the temperature control, but osculated decreasing function is employed for cool down function. Geophysical parameter estimation based on simulated annealing is proposed previously [6]. It takes relatively long computational time for convergence. Moreover, optimization with constraints makes much accurate estimation of geophysical parameters. Some of the constraints is relation among the geophysical parameters.

Geophysical parameters have relations each other. For instance, sea surface temperature and water vapor has a positive relation, in general. Therefore, it is better to estimate several geophysical parameters simultaneously rather than the estimation for single parameter. The proposed method is based on modified SA algorithm and is for simultaneous estimation for several geophysical parameters at once. Some experiments are conducted with Advanced Microwave Scanning Radiometer: AMSR [2] onboard AQUA satellite. Then it is confirmed that the proposed method surely works for improvement of estimation accuracy for all the geophysical parameters.

The related research works is described the following section. Then the proposed method is described followed by experiments. The experimental results are validated in the following section followed by conclusion with some discussions.

II. RELATED RESEARCH WORKS

A. Geophysical Parameter Estimation by Regressive Analysis

There are some atmospheric and ocean surface models in the microwave wavelength region. Therefore, it is possible to

estimate at sensor brightness temperature (microwave radiometer) with the geophysical parameters. The real and the imaginary part of dielectric constant of the calm ocean surface is modeled with the SST, salinity (conductivity). From the dielectric constant, reflectance of the ocean surface is estimated together with the emissivity (Debue, 1929 [26]; Cole and Cole,1941 [27]). There are some geometric optics ocean surface models (Cox and Munk, 1954 [28]; Wilheit and Chang, 1980 [29]). According to the Wilheit model, the slant angle against the averaged ocean surface is expressed by Gaussian distribution function.

There is a relation between ocean wind speed and the variance of the Gaussian distribution function as a function of the observation frequency. Meanwhile the influence due to foams, white caps on the emissivity estimation is expressed with the wind speed and the observation frequency so that the emissivity of the ocean surface and wind speed is estimated with the observation frequency simultaneously. Meanwhile, the atmospheric absorptions due to oxygen, water vapor and liquid water were well modeled (Waters, 1976 [30]). Then atmospheric attenuation and the radiation from the atmosphere can be estimated using the models. Thus the at-sensor-brightness temperature is estimated with the assumed geophysical parameters.

Sea surface temperature estimation methods with AMSR data are proposed and published [31] while ocean wind retrieval methods with AMSR data are also proposed and investigated [32]. Furthermore, water vapor and cloud liquid estimation methods with AMSR data are proposed and studied [33]. The conventional geophysical parameter estimation method is based on regressive analysis with a plenty of truth data and the corresponding microwave radiometer data [34].

The brightness temperature which acquired with microwave radiometer depends on geophysical parameters, (1) Sea Surface Temperature: SST, (2) ocean Wind Speed: WS, (3) Cloud Liquid: CL, (4) Water Vapor: WV in the atmosphere, (5) Salinity: SAL, etc. Also, the brightness temperature depends on observation frequency and abservation angle.

There are physical model based approach and statistical model based approach. The most typical statistical model is proposed by Frank Wentz [33]. His model is expressed with the following second order of equation,

$$\text{Geophysical}(x)=c_0 + \Sigma\, a_i\, T_{Bi} + \Sigma\, b_i\, T_{Bi}^2 \quad (1)$$

where Geophysical(x) denotes geophysical parameter of (x) while a_i, b_i denotes regressive coefficients while T_{Bi} denotes observed brightness temperature with microwave radiometer, respectively. When truth data of the geophysical parameter are given, then regressive coefficients are derived through regressive analysis.

Once the regressive coefficients, geophysical parameter can be estimated with the regressive equation and the observed brightness temperature. Example of the regressive coefficients for geophysical parameter of SST for Advanced Microwave Scanning Radiometer: AMSR of the 10GHz frequency band which is carried by AQUA, etc. is shown in Table 1.

TABLE I. EXAMPLE OF THE REGRESSIVE COEFFICIENTS FOR GEOPHYSICAL PARAMETER OF SEA SURFACE TEMPERATURE

	Coefficient
c_o	122.317
a_1	2.1117
a_2	0.9079
a_3	0.4618
a_4	-0.6192
a_5	-1.0579
a_6	0.6242
a_7	-8.915
a_8	25.6123
a_9	-0.4318
a_{10}	0.2244
b_1	0.0335
b_2	0.00468
b_3	-0.0293
b_4	0.003914
b_5	-0.4718
b_6	0.000753
b_7	-5.9235
b_8	5.4932
b_9	0.001703
b_{10}	0.0001107

Although this regressive approach is convenient and ensures a marginal accuracy, it is not enough SST estimation accuracy. It depends on the ocean areas, seasons, etc. Therefore, the regressive equation with only one set of coefficients cannot cover these dependencies which results in not so good estimation accuracy.

B. Physical Model Based Approach

Minimizing the difference between a geophysical model based Brightness Temperature: T_m and an acquired actual Brightness Temperature: T_a, input parameter of geophysical parameter can be estimated. T_a, depends on the observation frequency, observation angle, and the geophysical parameters as mentioned above. The observation frequency and angle is known. Therefore, the geophysical parameters can be estimated through minimization of the difference between both of T_m and T_a,. The important thing for this approach is accurate geophysical model. There is the well known sea surface model which is proposed by Thomas T. Wilheit [28].

III. PROPOSED MODEL

A. Basic Idea

The brightness temperatures of the several observation frequency bands can be acquired in both horizontal and vertical polarizations. If the users focus water vapor and cloud liquid, then 23 GHz and 31 GHz of observation frequency bands are needed. It is totally up to frequency dependency of brightness temperature of frequency. There is strong absorption of water vapor at the 23.235 GHz while dual frequency channels allow simultaneous estimation of water vapor and cloud liquid. Therefore, 23 GHz and 31 GHz of frequency bands are effective for water vapor and cloud liquid estimations. And if we focus SST and wind speed, only 6.925 and 10.69 GHz of observation frequency bands are taken into account. In this paper, targeted geophysical parameters are SST and Wind Speed.

The observed brightness temperature at the certain frequency band in horizontal and vertical polarizations are expressed as follows,

$$T_{bh}=\varepsilon_h(T,W)T + n_h \quad (2)$$

$$T_{bv}=\varepsilon_v(T,W)T + n_v \quad (3)$$

where T_{bh}, ε_h, T, W, n_h denotes brightness temperature, emissivity of the sea surface, Planck function of surface temperature, ocean wind speed, and observation noise for horizontal polarization while these for suffix of *v* denotes those for vertical polarization. Cost function of optimization processes is defined as follows,

$$\| T_{bh} - \varepsilon_h(T,W)T\|^2 + \| T_{bv} - \varepsilon_v(T,W)T\|^2 \quad (4)$$

Minimizing the cost function of equation (4) with the changing the input parameter of T and W, T and W can be estimated by using the observed brightness temperature. The most important thing for this method is how to estimate sea surface emissivity. In accordance with the Wilheit model, emissivity in horizontal and vertical polarizations is estimated. Fig.1 shows the example of the calculated emissivity.

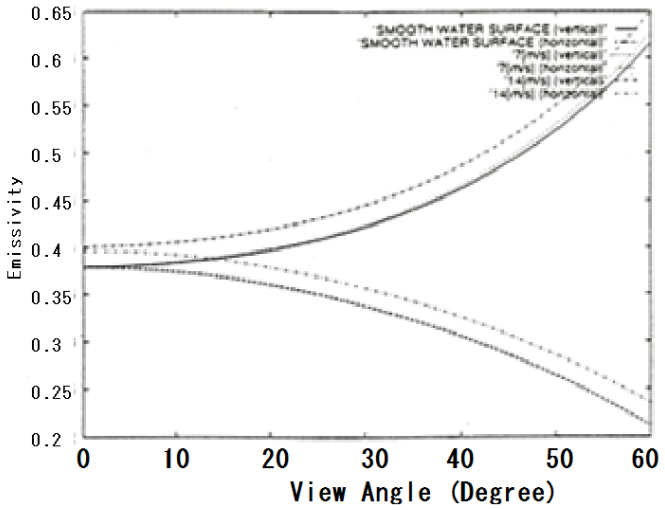


Fig. 1. Emissivity model originated from the Wilheit model

B. Simulated Annealing

The proposed geophysical parameter estimation here is based on the physical model based approach. Minimization of the difference between T_m and T_a, is total identical to optimization model. The problem situated here is how to find the global optimum. Only the solution for that is Simulated Annealing: SA. It, however, takes huge computational resources. Therefore, the proposed model here is modified SA model which has a limitation of iteration. Namely, iterations is stopped at the previously designated upper limit. Therefore, the proposed modified SA is not real SA essentially because the solution does not reach to a global optimum. In the case of the estimation of geophysical parameter with microwave radiometer data, residual error is gradually reduced when the current solution is approaching to a global optimum (the solution does not jump in this stage). Therefore, we may stop the iteration at the certain number of iterations or elapsed computation time.

IV. Experiments

A. Validation of Emissivity Model

As an example of brightness temperature, the brightness temperature of Microwave Imager: TMI onboard Tropical Rainfall Measuring Mission: TRMM satellite of 10.65 GHz for horizontal and vertical polarizations is shown in Fig.2. The actual brightness temperature as a function of observation angle is plotted in Fig.2. The location of intensive study area is the following,

Longitude and latitude: 31.6 North, 109.1 East

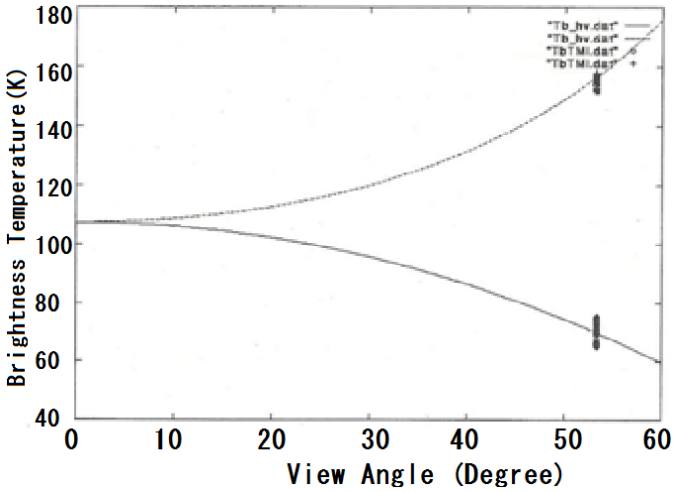


Fig. 2. Brightness temperature for both horizontal and vertical polarizations derived from the proposed physical model based method and actual received brightness temperature with TRMM/TMI of 10.65 GHz of frequency channel acquired on June 2 1998

The actual brightness temperature data are situated at the observation angle of 53 degree because the brightness temperature for horizontal polarization does not depend on ocean wind speed at the observation angle of 53 degree. The estimated brightness temperature is coincident to the actual brightness temperature. This is the same thing for the different observation frequency and both of horizontal and vertical polarizations. Therefore, emissivity model originated from the Wilheit model is validated.

The actual TMI data of the location (Longitude and latitude: 31.6 North, 109.1 East) which is acquired on June 2 1998 is used for the experiment. From the measured data at the site, it is found that SST=294 K, WS=7 m/s, Salinity=36ppm, respectively. The truth geophysical parameters of SST are set at 292 K, 294 K, and 296 K while that of wind speed is set at 7 m/s. The brightness temperature estimated by the proposed physical model based method. The results are as follows,

1) Theoretical brightness temperature: 70.549
The mean of observed brightness temperature: 100.589

The standard deviation of the actual brightness temperature: 9.634

2) Theoretical brightness temperature: 156.574
The mean of observed brightness temperature: 173.814

The standard deviation of the actual brightness temperature: 2.906

3) Theoretical brightness temperature: 70.3
The mean of observed brightness temperature: 100.589

The standard deviation of the actual brightness temperature: 9.635

4) Theoretical brightness temperature: 155.905
The mean of observed brightness temperature: 173.814

The standard deviation of the actual brightness temperature: 2.906

5) Theoretical brightness temperature: 70.081
The mean of observed brightness temperature: 100.589

The standard deviation of the actual brightness temperature: 9.635

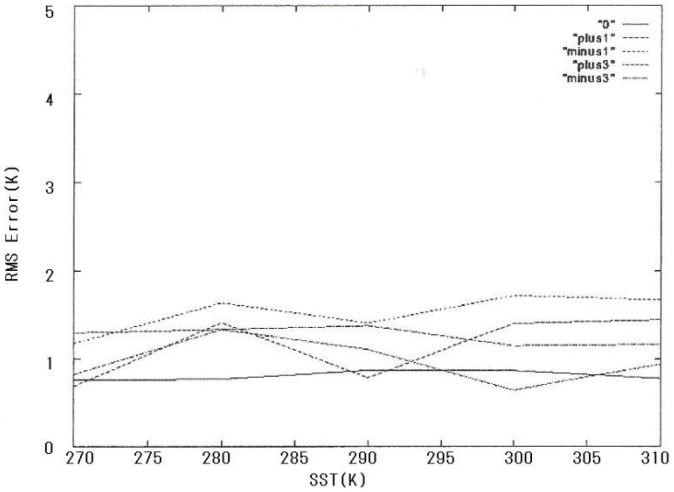
6) Theoretical brightness temperature: 155.284
The mean of observed brightness temperature: 173.814

The standard deviation of the actual brightness temperature: 2.906

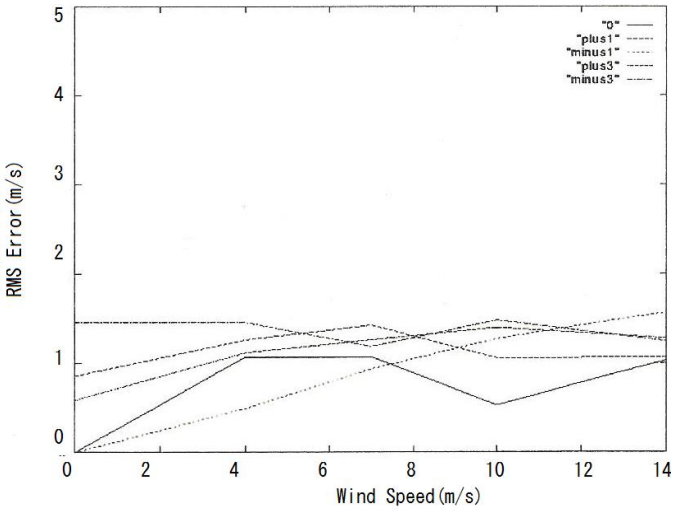
Thus the proposed model is validated with some extent of estimation errors.

B. Comparison of Estimated Sea Surface Temperature

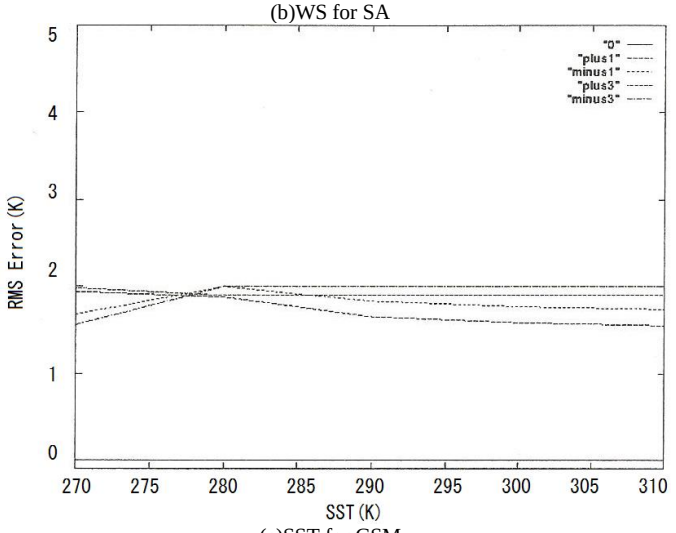
In order to show the advantage of the proposed method, the estimated SST and WS with the proposed method is compared to those with the statistical model based method, conventional GSM method. Fig.3 shows the results from the comparative study. In the experiment, observation frequency channels are set at 6.925 GHz and 10.69 GHz. Fig.3 shows RMS error of SST and WS with the designated biases of plus minus 1(K), 3(K) for SST and plus minus 1(m/s), 3(m/s) for WS as well as without any bias for the proposed SA based method and the conventional GSM method.



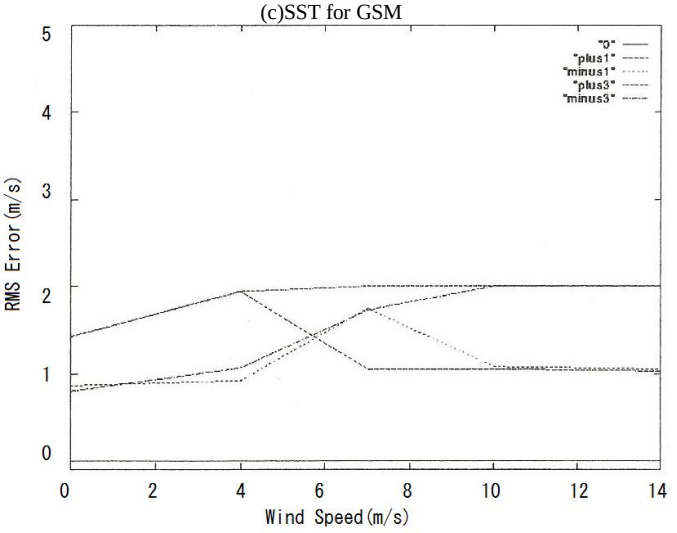
(a)SST for SA



(b)WS for SA



(c)SST for GSM



(d)WS for GSM

Fig. 3. RMS error of SA and GSM for the estimation of SST and WS with the designated bias of plus minus 1(K), 5(K) for SST and plus minus 1(m/s), 3(m/s) for WS as well as without any bias

As the results, it is found that RMS error of the proposed SA based method is superior to the conventional GSM method by approximately 50 (%) for both of SST and WS. Also, it is found that the RMS error is getting large in accordance with increasing of additive biases.

Root Mean Square: RMS error is evaluated and compared. Table 2 shows the results of RMS errors for the statistical model based method, GSM method and the proposed SA method.

TABLE II. RMS ERROR COMPARISONS AMONG THE STATISTICAL MODEL BASED METHOD, GSM METHOD AND THE PROPOSED SA METHOD

Method	SST(K)	WS(m/s)
Statistical Approach	0.46	0.66
GSM	0.274	0.327
SA	0.492	0.435

If the biases are added to the theoretical SST and WS intentionally, then the RMS errors are varied as shown in Table 3 for GSM method while those for SA method is shown in Table 4.

TABLE III. RMS ERRORS OF SST AND WS FOR GSM METHOD AS A FUNCTION OF DEVIATIONS

Biases	SST(K)	WS(m/s)
0	0	0
+1	1.722	1.302
-1	1.805	1.135
+3	1.916	1.874
-3	1.912	1.520

TABLE IV. RMS ERRORS OF SST AND WS FOR SA METHOD AS A FUNCTION OF DEVIATIONS

Deviation	SST(K)	WS(m/s)
0	0.809	0.739
+1	1.146	1.132
-1	1.520	0.853
+3	1.064	1.127
-3	1.169	1.363

By using the actual brightness temperature data of TMI, SST and WS estimation errors are evaluated. Table 5 shows the estimated SST and WS as well as RMS errors for the cases of SST are set at 292, 294 and 296(K). In these cases, the estimated SST and WS are compared to the actual TMI data derived SST and WS. RMS error of SST shows around 4.5(K) while that of WS is approximately 3.7(m/s) respectively.

TABLE V. ESTIMATED SST AND WS AS WELL AS RMS ERRORS FOR THE CASES OF SST ARE SET AT 292, 294 AND 296(K)

Case	SST(K)	RMSE(SST)	WS(m/s)	RMSE(WS)
296(K)	291.8	4.297(K)	6.708	4.036(m/s)
294(K)	289.498	4.619(K)	7.313	2.997(m/s)
292(K)	287.348	4.753(K)	6.604	4.156(m/s)

As the results from the experiments, it is found that the proposed SA based method is superior to the statistical model based method and the GSM method.

V. CONCLUSION

Comparative study of optimization methods for estimation sea surface temperature and ocean wind with microwave radiometer data is conducted. The well known mesh method (Grid Search Method: GSM), regressive method, and simulated annealing method are compared. Surface emissivity is estimated with the simulated annealing and compared to the well known Thomas T. Wilheit model based emissivity. On the other hand, brightness temperature of microwave radiometer as a function of observation angle is estimated by the simulated annealing method and compares it to the actual microwave radiometer data. Also, simultaneous estimation of sea surface temperature and ocean wind speed is carried out by the simulated annealing and compared it to the estimated those by the GSM method. The experimental results show the simulated annealing which allows estimation of global optimum is superior to the other method in some extent.

As the results, it is confirmed that the well known Wilheit sea surface model is appropriate for estimation of geophysical parameters. Also, it is confirmed that the statistical model based method for geophysical parameter estimation shows marginal estimation accuracies of SST and WS (0.46(K) and 0.66(m/s), respectively). It is found that the estimated SST and WS are compared to the actual TMI data derived SST and WS. RMS error of SST for the proposed SA based method shows around 4.5(K) while that of WS is approximately 3.7(m/s) respectively.

ACKNOWLEDGMENT

The author would like to thank Ms. Emi Shimomura of Saga University for her effort to conduct the experiments.

REFERENCES

[1] Kohei Arai, Preliminary Assessment of Radiometric Accuracy for MOS-1 Sensors, International Journal of Remote Sensing, Vol.9, No.1, pp.5-12, Apr.1988.

[2] K.Tachi, Kohei Arai and Y.Satoh, Advanced Microwave Scanning Radiometer -Requirements and Preliminary Design Study-, IEEE Trans. on Geoscience and Remote Sensing, Vol.27, No.2, pp.177-183, Jan.1989.

[3] Kenbu Teramoto, Kohei Arai, Toshio Imatani, Antenna Pattern Correction of Microwave Radiometer Based on 凸 Projection Method, Journal of Remote Sensing Society of Japan, Vol.15, No.4, pp.38-49, Sep.1994.

[4] Kohei Arai, E.Ishiyama and Y.Terayama, A method for ice concentration estimation with microwave radiometer data by means of inversion techniques, Advances in Space Research, Vol.16, No.10, pp.129-132, A31-32, Jul.1994.

[5] Kohei Arai, Kenbu Teramoto, Algorithm of Multi-Frequency Microwave Radiometer Data Correction, Notes of Meteorological Research, No.187, pp.169-176, Aug.1996.

[6] Kohei Arai and J.Sakakibara, Estimation of SST, wind speed and water vapor with microwave radiometer data based on simulated annealing, Advances in Space Research, 37, 12, 2202-2207, 2006.

[7] Kohei Arai, Kenta Azuma, Ocean wind speed estimation accuracy improvement using microwave radiometer data taking into account wind direction dependency of the observed brightness temperature, Journal of Remote Sensing Society of Japan, 27, 5, 465-473, 2007

[8] Kohei Arai, Kenta Azuma, Rainfall rate estimation with AMSR data on board AQUA satellite taking into account geometric relation between the microwave radiometer and observation targets, Journal of Japan Society of Photogrammetry and Remote Sensing, 49, 1, 32-40, 2010

[9] K.Arai, Simultaneous estimation of geophysical parameters with microwave radiometer data based on accelerated Simulated Annealing: SA, International Journal of Advanced Computer Science and Applications, 3, 7, 90-95, 2012.

[10] K.Arai, Nonlinear Optimization Based Sea Surface Temperature: SST Estimation Methods with Remote Sensing Satellite Based Microwave

Scanning Radiometer: MSR Data, International Journal of Research and Reviews in Computer Science (IJRRCS) Vol. 3, No. 6, 1881-1886, December 2012, ISSN: 2079-2557

[11] Kohei Arai, Data fusion between microwave and thermal infrared radiometer data and its application to skin sea surface temperature, wind speed and salinity retrievals, International Journal of Advanced Computer Science and Applications, 4, 2, 239-244, 2013.

[12] K.Arai, T.Igarashi and C.Ishida, Evaluation of MOS-1 Microwave Scanning Radiometer (MSR) data in field experiments, Proc. of the 18th International Symposium on Remote Sensing of Environment, 1-8, 1984.

[13] K.Arai, T.Igarashi and Y.Takagi, Emissivity model of snowpack for passive microwave observations, Proc. of the 36th International Astronautics Federation (IAF) Congress, IAF-85-98, 1-8, 1985.

[14] K.Arai and T.Suzuki, Beam compressed microwave scanning radiometer, Proc. of the IGARSS'89, I1-2, 268-270, 1989.

[15] Y.Itoh, K.Tachi, Y.Sato and K.Arai, Advanced Microwave Scanning Radiometer: AMSR, Preliminary study, Proc. of the IGARSS'89, I1-4, 273-276, 1989.

[16] K.Arai, K.Teramoto and T.Imatani, Influence due to antenna pattern changes in brightness temperature estimation for a space based microwave radiometer, Proc. of the European ISY Conference, 311-316, 1992.

[17] K.Arai, E.Ishiyama and Y.Terayama, Method for ice concentration estimation with microwave scanning radiometer data by means of inversion, Proc. of the 30th COSPAR Congress, A3.1-032, 1993.

[18] Arai,K., E.Ishiyama and Y.Terayama, A method for ice concentration estimation with microwave radiometer data by means of inversion techniques, Proceedings of the 30th COSPAR Symposium,., A31-032, 1994

[19] Arai,K., New algorithms for ice concentration estimation with passive microwave data, Proceedings of the 1st ADEOS-II Science Symposium, Nov., 1994.

[20] K.Teramoto, K.Arai and T.Imatani, Antenna Pattern Correction for Microwave Radiometry Using A Prior Knowledge Based on Projection Convex Sets Method, Proceeding of the International Geoscience and Remote Sensing Symposium, IGARSS'95, Florence,July 1995.

[21] K.Teramoto and K.Arai, POCS Based Array Processing in Incoherent Microwave Radiometric Image Reconstruction, Proceedings of the ICASSP'96, SSAP#615, Atlanta, May 1996.

[22] Kohei Arai, Sea Surface Temperature (SST) estimation with microwave radiometers by means of simulated annealing based on an ocean surface model, Proceedings of the NASA Oceanography Scientific Conference, Florida, USA, 2001.

[23] Kohei Arai, Sea Surface Temperature (SST) retrieval with microwave radiometer data based on simulated annealing, Proc. of the Kyushu Brunch Symposium of the electronics related Society of Japan, Asian Session, 2001.

[24] Kohei Arai and Jun Sakakibara, Estimation of SST, wind speed and water vapor with microwave radiometer data based on simulated annealing, Abstracts of the 35th Congress of the Committee on Space Research of the ICSU, A1.1-0130-04, (2004)

[25] S. Kirkpatrick, C.D. Gelett, M.P. Cecchi, Optimization by simulated annealing, Science, 220, 621-630, 1983.

[26] Debue, R. Polar Molecules, Chemical Catalog, New York, 1929.

[27] Cole, K.S., Cole, R.H. Dispersion and absorption in dielectrics. J. Chem. Phys. 9, 341–351, 1941.

[28] Cox, C.S., Munk, W.H. Measurement of the roughness of the sea surface from photographs of the sun_s glitter. J. Opt. Sci. Am. 44, 838–850, 1954.

[29] Wilheit, T.T., Chang, A.T.C. An algorithm for retrieval of ocean surface and atmospheric parameters from the observations of the Scanning Multichannel Microwave Radiometer (SMMR). Radio Sci. 15, 525–544, 1980.

[30] Waters, J.R. Absorption and emission by atmospheric gasses. in: Meeks, M.L. (Ed.), Methods of Experimental Physics, vol. 12B.Academic, Orland, 1976 (Chapter 2.3).

[31] Dong, SF; Sprintall, J; Gille, ST, Location of the antarctic polar front from AMSR-E satellite sea surface temperature measurements, *JOURNAL OF PHYSICAL OCEANOGRAPHY*, Nov 2006, 2075-2089.

[32] Konda, M., A. Shibata, N. Ebuchi, and K. Arai, An evaluation of the effect of the relative wind direction on the measurement of the wind and the instantaneous latent heat flux by Advanced Microwave Scanning Radiometer, *J. Oceanogr*., vol. 62, no. 3, pp. 395-404, 2006.

[33] Cosh, M. H., T. J. Jackson, R. Bindlish, J. Famiglietti, and D. Ryu, A comparison of an impedance probe for estimation of surface soil water content over large region, *Journal of Hydrology*, vol. 311, pp. 49-58, 2005.

[34] Wentz, F. AMSR Ocean Algorithm, second version of ATBD, NASA/GSFC, 2000.

AUTHORS PROFILE

Kohei Aarai He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 and also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission-A of ICSU/COSPAR since 2008. He wrote 33 books and published 510 journal papers. He is Editor-in-Chief of International Journal of Advanced Computer Science and Applications as well as I ternational Journal of Intelligent Systsems and Applications.

Rescue System with Health Condition Monitoring Together with Location and Attitude Monitoring as Well as the Other Data Acquired with Mobile Devices

Kohei Arai[1]
1 Graduate School of Science and Engineering
Saga University
Saga City, Japan

Taka Eguchi[1]
1 Graduate School of Science and Engineering
Saga University
Saga City, Japan

***Abstract*—Rescue system with health condition monitoring together with location and attitude monitoring as well as the other data acquired with mobile devices is proposed. Backup system for location estimation is also proposed. On behalf of GPS receivers and WiFi beacon receivers, ZigBee is used as a backup system. Attitude can be monitored with acceleration-meters equipped in the commercially available smart phones and i-phones. Also, the number of steps and calorie consumptions can be monitored with the commercially available smart phones and i-phones. By using these body attached sensors, health condition of the persons who need a help for rescue when the emergency situations can be monitored and used for rescue planning and triage. Overall system configuration is proposed together with the detailed system descriptions with some of the experimental data.**

Keywords*—Rescue system; Location estimation; Attitude estimation; Health monitoring; Mobile applications; Triage; Rescue planning

I. INTRODUCTION

There are previously proposed methods and systems which allow physical health monitoring [1]-[5]. Most of previous methods and systems are not wearable and do not allow psychological status monitoring. The proposed physical and psychological health monitoring system is intended to monitor these five major vital signs. Instead of direct blood pressure measurement, indirect blood pressure measurement is proposed by using a created regressive equation with the measured body temperature, heart rate and the number of steps because it is hard to measure the blood pressure directly. Also, consciousness can be monitored by using acquired eye images and its surroundings on behalf of using EEG sensors, because EEG signals are used to be suffered from noises.

There are previously proposed evacuation and rescue methods and systems [6]-[8]. It may be possible to find that multi agent-based simulation makes it possible to simulate the human activities in rescue and evacuation process [9],[10]. A multi agent-based model is composed of individual units, situated in an explicit space, and provided with their own attributes and rules [11]. This model is particularly suitable for modeling human behaviors, as human characteristics can be presented as agent behaviors. Therefore, the multi agent-based model is widely used for rescue and evacuation simulation [9]-[13].

In this study, GIS map is used to model objects such as road, building, human, fire with various properties to describe the objects condition. With the help of GIS data, it enables the disaster space to be closer to a real situation [13]-[16].

A rescue model for people with disabilities in large scale environment is proposed. The proposed rescue model provides some specific functions to help disabled people effectively when emergency situation occurs. Important components of an evacuation plan are the ability to receive critical information about an emergency, how to respond to an emergency, and where to go to receive assistance. Triage is a key for rescue procedure. Triage can be done with the gathered physical and psychological data which are measured with a sensor network for vital sign monitoring. Through a comparison between with and without consideration of triage, it may be possible to find that the time required for evacuation from disaster areas with consideration triage is less than that without triage. The following section describes the proposed rescue system with triage followed by examples of the monitored data of health conditions together with the location of attitude monitoring. Then alternative location determination with ZigBee receiver and transmitter is described with some experimental data. Finally, conclusion is described together with some discussions.

II. PROPOSED RESCUE SYSTEM

A. Basic Idea

Fig.1 shows the concept of the proposed rescue system. There are three major components, persons who need a help for evacuation, Information Collection Center: ICC for health, traffic, and the other conditions together with the location and attitude information of the persons who need a help and the rescue peoples. Body attached sensors allow measurements of health conditions and the location and attitude of the persons who need a help. The measured data can be transmitted to the ICC through smart-phone, or i-phone, or tablet terminals of which the persons who need a help are carrying. By using the collected health condition and the location/attitude as well as traffic condition information, most appropriate rescue peoples are determined by the person by the person. It is better to consider a triage in the emergency rescue stages. Therefore, health condition monitor is necessary. Fig.2 shows the proposed health condition monitoring system together with the acquired data transmission system.

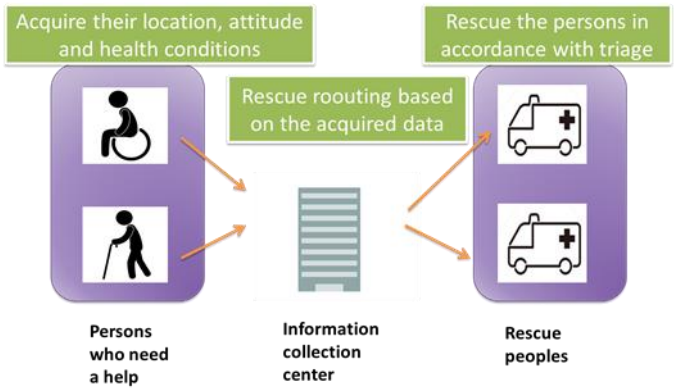


Fig. 1. Concept of the proposed rescue system

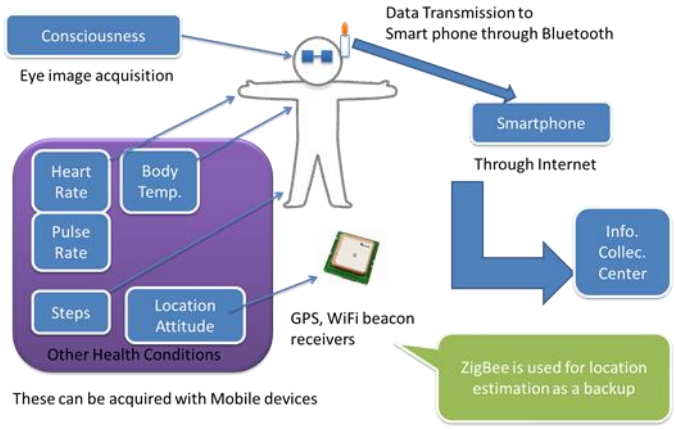


Fig. 2. Proposed health condition monitoring system together with the acquired data transmission system

B. Sensors

There are four major items of the vital signs for triage. Those are Body temperature, Blood pressure, Pulse rate, Number of blesses. Therefore, these four items are mandatory for triage. Other than these, Number of steps, calorie consumptions would be better to monitor together with the location and attitude as well as types of movement (walking, standing, sitting, laying, running, jumping, etc.). These data can be transmitted to the ICC through WiFi networks (Bluetooth for transmission from sensors to smart-phone and WiFi network for transmission from the smart-phone to ICC). Location measurement can be done with GPS receiver in the smart-phone and with WiFi beacon receiver is the same smart-phone. The GPS receiver does not work indoor situation. Also, both GPS receiver and WiFi receiver based location determination accuracy is not good enough. Therefore, some of alternative method would be better to add to the rescue system. In this paper, ZigBee of transmitter and receiver is used for location determination as a backup as shown in Fig.2.

Outlooks of body attached sensors are shown n Fig.3. (a)Pulse Rate (b)Heart Rate (c)Body Temperature (d)Blood Pressure (e)Step, Calorie Consumption, Location and Attitude can be measured with these sensors.

(a)Pulse Rate (b)Heart Rate (c)Body Temperature

(d)Blood Pressure (e)Step, Calorie Consumption, Location and Attitude

Fig. 3. Used body attached sensors for health monitoring

C. Triage

In the triage stage, the types of disabilities which are shown in Table 1 are taken into account. Through a consideration of these types of disabilities, 10 grades of disabilities are taken into account in the triage.

TABLE I. TYPES OF DISABILITY

Types of Disability
Cognitive Disorder
Neuropathy
Movement Disorder
Elderly Condition
Hearing Loss
Language, Visual Impairment

D. Examples of Measured Data

Other than smart-phone based step monitoring, there are some body attached step monitoring sensors. Fig.4 shows an example of the step monitoring sensor. The sensor allows measurement not only the number of steps but also calorie consumption can be measured. Also, these measured data are archived and referred through Bluetooth communications. One of example of the archived steps and calorie consumption is shown in Fig.5.

Fig. 4. Alternative step monitoring sensor

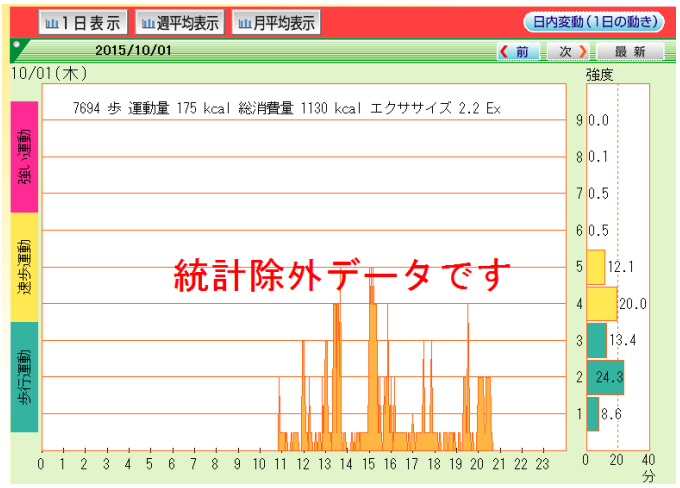


Fig. 5. Example of the acquired step data together with calorie consumption

Moreover, the number of steps can be measured with i-phone application software tools. Fig.6 shows an example of the measures steps in a month with iOS8 of i-phone.

Fig. 6. Step monitoring with iOS8 of i-phone

There is health condition monitoring application software tool so called HealthKit under the iOS8. The menu of the HealthKit is shown in Fig.7.

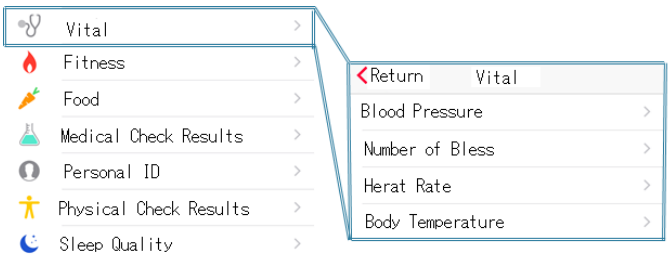


Fig. 7. Health condition monitoring with HealthKit with iOS8

Other than the vital records, Fitness, Food, Medical check results, physical check results, sleep quality and personal ID can be referred. Process flow of the proposed health monitoring system together with the acquired data transmission system is shown in Fig.8.

Fig. 8. Process flow of the proposed health monitoring system together with the acquired data transmission system

On the other hand, Android OS of smart-phone which is shown in Fig.9 provides API which allows the step count and step detector as shown in Fig.10. That is Android4.4kit-kat.

Fig. 9. Example of Android OS of smart-phone

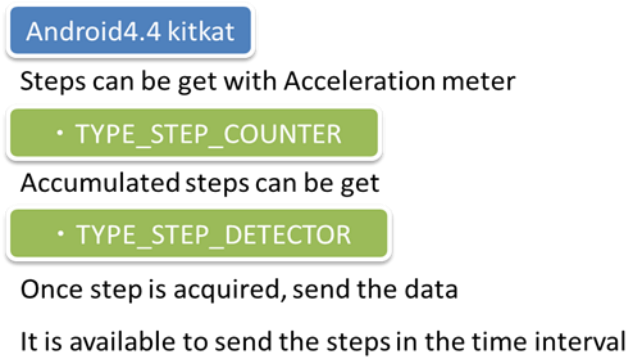


Fig. 10. Additionally available APIs by using Android4.4kitkat

Using the API of Android4.4kit-kat together with acceleration meter, types of movements can be determined as shown in Fig.11. Also, identification of attitude type by using the difference between actual and reference power spectrum derived from acceleration meter is available as shown in Fig.12.

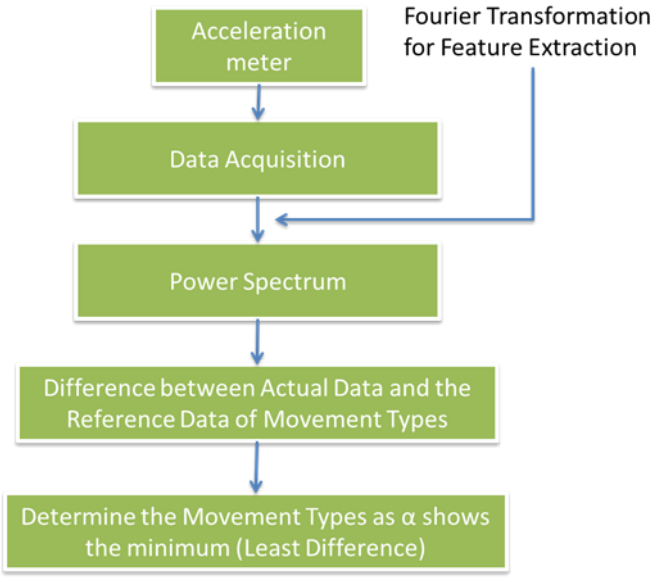


Fig. 11. Method for attitude detection

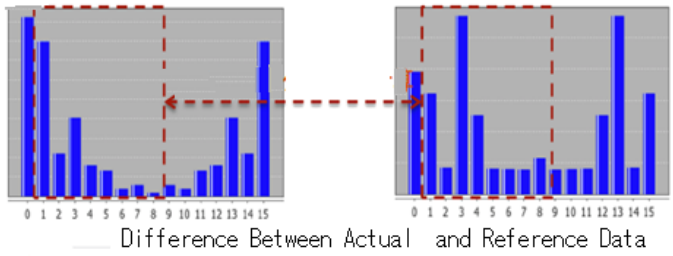


Fig. 12. Identification of attitude type by using the difference between actual and reference power spectrum derived from acceleration meter

Acquired acceleration meter data is compared to the previously acquired reference data of the designated several movement types in the frequency domain (Power spectrum). The frequency components between *a* and *b* are compared followed by the summation of the different between both, *a* and *b* calculation as shown in equation (1).

$$\alpha = |a2 - b2| + |a3 - b3| + \cdots + |a9 - b9| \quad (1)$$

Thus movement types are discriminated. Also, movement types can be discriminated with ZigBee. Fig.13 shows outlook of the ZigBee used for the experiments. Movement types, in this case, can be identified as "Stop", "Begin to move" and "Freely falling down" as shown in Fig.14. One of the examples of the receiving signal for the movement type of the "Freely falling down" is shown in Fig.15 while that of the "Stop" on the floor is shown in Fig.16, respectively. Thus the movement types can be identified with receiving signal strength of ZigBee. Meanwhile, one of examples of the measured pulse rate with the Pulse Coach which is shown in Fig.3 (a) is shown in Fig.17. Pulse per minute can be measured every one second.

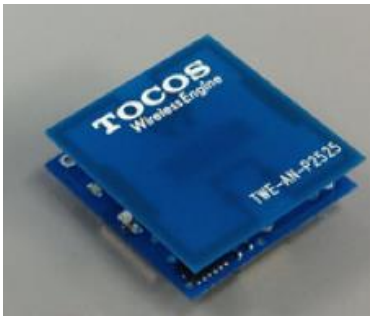


Fig. 13. Outlook of ZigBee

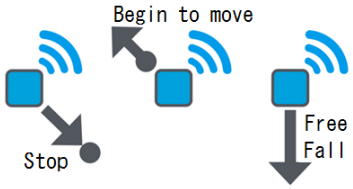


Fig. 14. Movement type identification with ZigBee

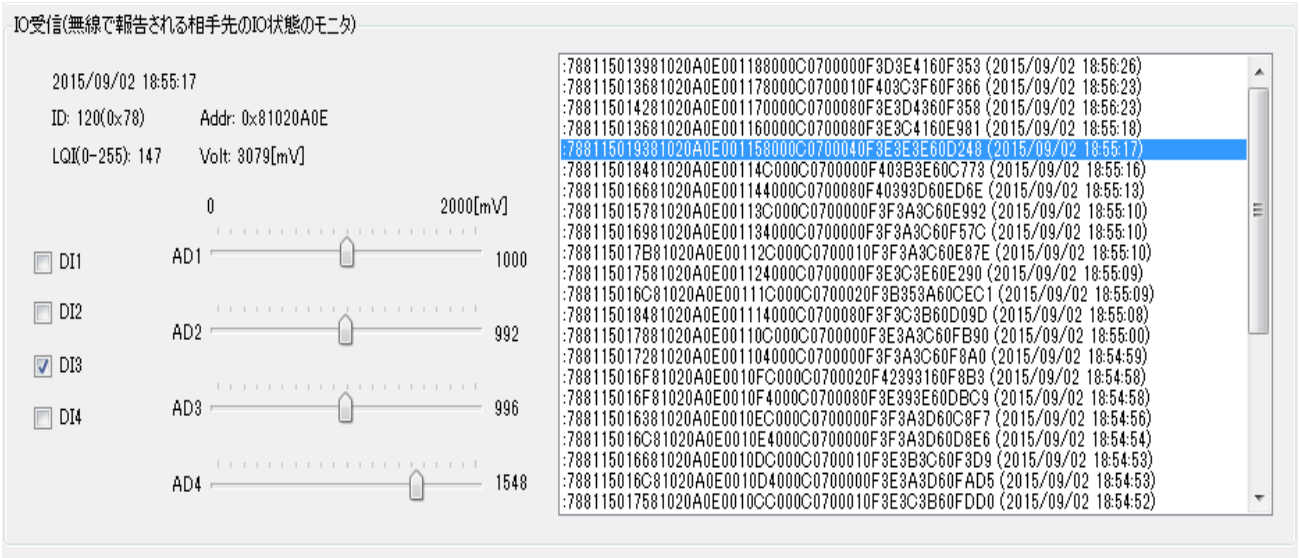


Fig. 15. Example of the receiving signal from the ZigBee when it is falling freely (DI3)

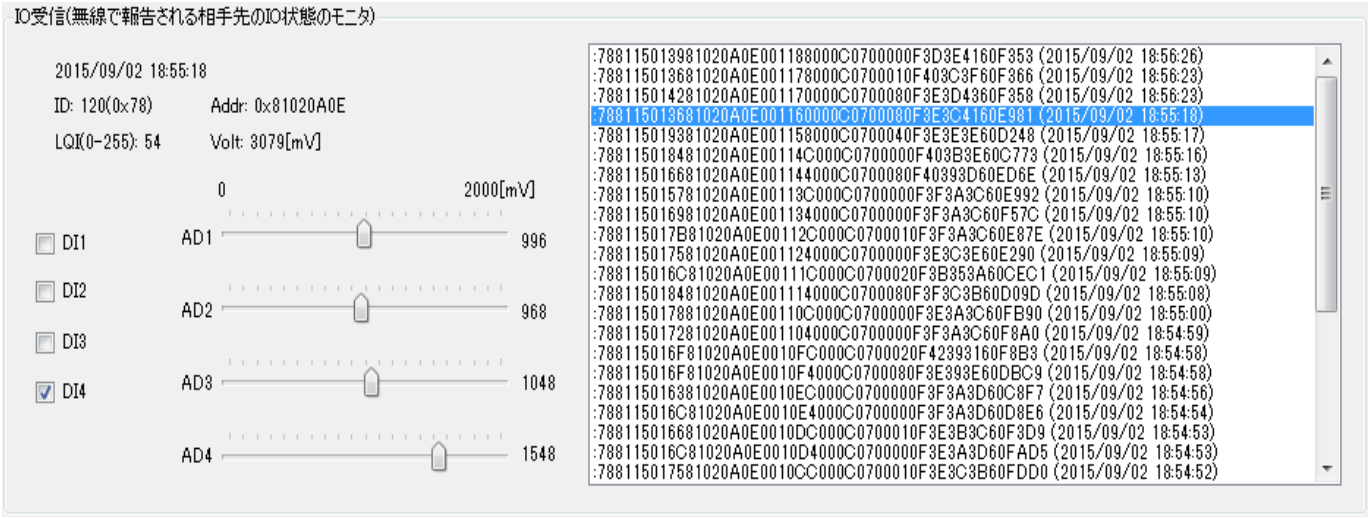


Fig. 16. Example of the receiving signal from the ZigBee when it is on the floor (DI4)

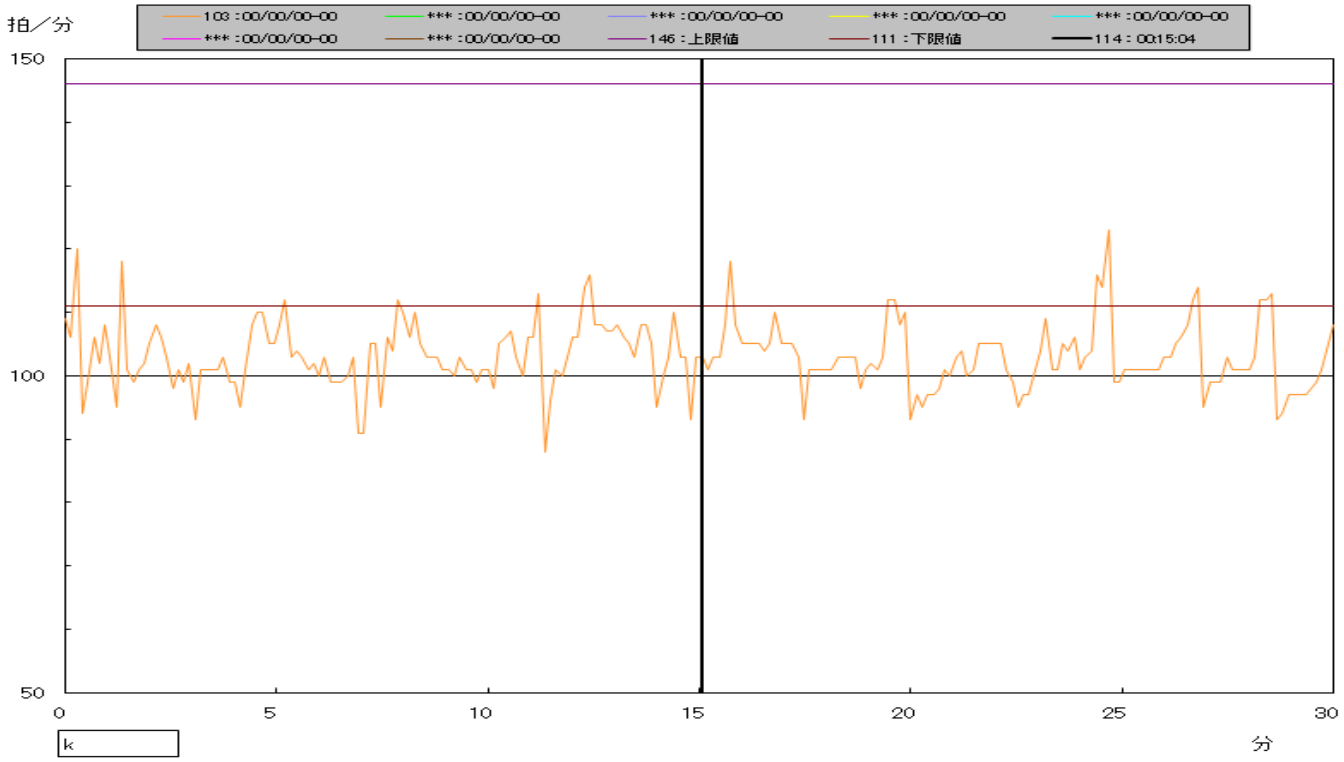


Fig. 17. Example of Pulse Rate measured data (1/minute)

III. Distance Measurements with ZigBee

A. Basic Idea

ZigBee coverage is limited up to around 100 m. Therefore, the location of the ZigBee receiver can be identified from the surrounding three ZigBee transmitters as shown in Fig.18. Also, one pair of ZigBee transmitter and receiver makes distance measurements with signal strength. Also, ZigBee coverage can be expanded with through repeaters as shown in Fig.19.

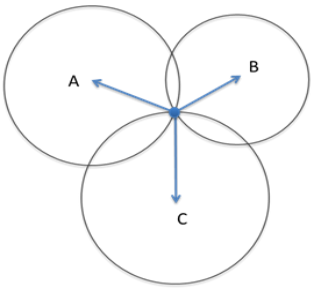
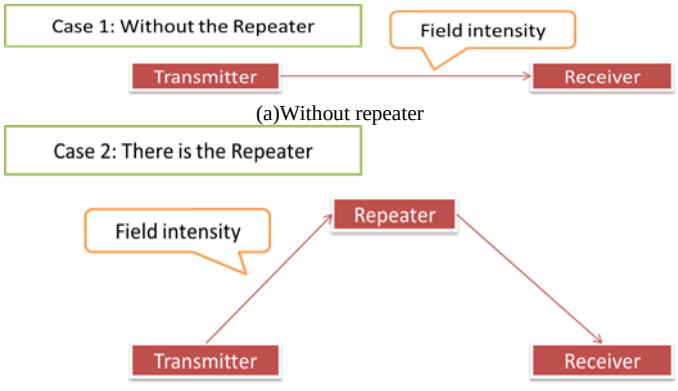


Fig. 18. Location determination concept with three ZigBee stations



(a)Without repeater

(b)With repeater

Fig. 19. Distance measurements

B. Measurement Data

Fig.20 shows example of the received signal of the repeater. The data are aligned as shown n Fig.20 (b). The signal includes not only signal strength but also the repeater ID. Therefore, it can be identified that the data is received through which repeaters.

```
;178;
;179;
;179;00000000;135;006;1002717;3340;0742;7998;1333;0742;S;
;179;0100269B;120;006;1002717;3340;0742;7998;1333;0742;S;
;180;
;181;
;182;
;183;
;183;00000000;138;007;1002717;3340;0740;7986;1331;0740;S;
;183;0100269B;153;007;1002717;3340;0740;7986;1331;0740;S;
;184;
;185;
```

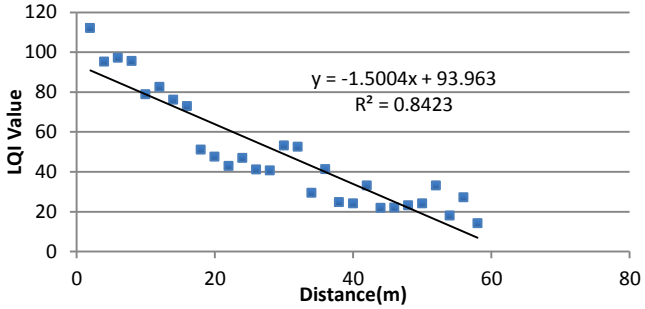
(a)Actual data

The order of the actual data from the top right to left bottom
Time stamp
ID of the repeater
Signal Strength: LQI
Continuous No.
ID of the receiver
Supply voltage of the receiver
AI3(mV)
AI1(three times of voltage)
AI1(mV)
AI3(mV)
Packet ID

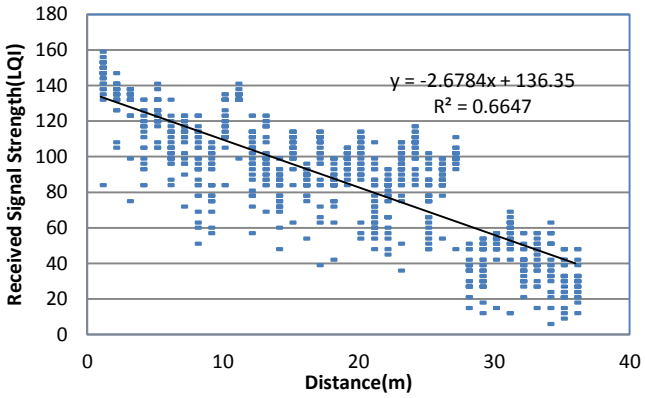
(b)Alignment of the data

Fig. 20. Example of the receiving signal

Fig.21 (a) shows example of the distance measurement results for the case of without repeater (Outdoor) while Fig.21 (b) shows that of with repeater (Indoor), respectively. There is much electro-magnetic interference from the wall, pillar, etc. for the case of "Indoor". Therefore, location estimation accuracy is not so good for the "Indoor" case.



(a)Without Repeater (Outdoor)



(b)With Repeater (Indoor)

Fig. 21. Example of relation between receiving signal strength and the distance

IV. Conclusion

Rescue system with health condition monitoring together with location and attitude monitoring as well as the other data acquired with mobile devices is proposed. Backup system for location estimation is also proposed. On behalf of GPS receivers and WiFi beacon receivers, ZigBee is used as a backup system. Attitude can be monitored with acceleration-meters equipped in the commercially available smart phones and i-phones. Also, the number of steps and calorie consumptions can be monitored with the commercially available smart phones and i-phones. By using these body attached sensors, health condition of the persons who need a help for rescue when the emergency situations can be monitored and used for rescue planning and triage. Overall system configuration is proposed together with the detailed system descriptions with some of the experimental data.

Also, it is found that the distance measurements can be done with ZigBee. Moreover, it is found that the coverage of the ZigBee location identification can be expanded with ZigBee transmitter and receiver (Repeater).

Acknowledgment

The author would like to thank Dr. Trang Xuang Sang of Vinh University in Vietnam for his effort to conduct simulation studies.

REFERENCES

[1] K.Arai,WeIarable healthy monitoring sensor network and its application to evacuation and rescue information server system for disabled and elderly person, International Jmynal of Research and Review on Computer Science, 3, 3, 1633-1639, 2012.

[2] Kohei Arai, Wearable computing system with input output devices based on eye-based Human Computer Interaction: HCI allowing location based Ib services, International Jmynal of Advanced Research in Artificial Intelligence, 2, 8, 34-39, 2013.

[3] Kohei Arai, Vital sign and location/attitude monitoring with sensor networks for the proposed rescue system for disabled and elderly persons who need a help in evacuation from disaster areas, International Jmynal of Advanced Research in Artificial Intelligence, 3, 1, 24-33, 2014.

[4] Kohei Arai, Method and system for human action detection with acceleration sensors for the proposed rescue system for disabled and elderly persons who need a help in evacuation from disaster areas, International Jmynal of Advanced Research in Artificial Intelligence, 3, 1, 34-40, 2014.

[5] Kohei Arai, Frequent physical health monitoring as vital sign with psychological status monitoring for search and rescue of handicapped, disabled and elderly persons, International Jmynal of Advanced Research in Artificial Intelligence, 2, 11, 25-31, 2013

[6] .J. Kaprzy Edt., Kohei Arai, Rescue System for Elderly and Disabled Persons Using Iarable Physical and Psychological Monitoring System, Studies in Computer Intelligence, 542, 45-64, Springer Publishing Co. Ltd., 2014.

[7] Obelbecker G., & Dornhege M., "Realistic cities in simulated environments - an Open Street Map to Robocup Rescue converter", Online-Proceedings of the Fmyth International Workshop on Synthetic Simulation and Robotics to Mitigate Earthquake Disaster, 2009.

[8] Sato, K., & Takahashi, T., "A study of map data influence on disaster and rescue simulation's results", Computational Intelligence Series, vol. 325. Springer Berlin / Heidelberg, 389–402, 2011.

[9] Ren C., Yang C., & Jin S., "Agent-Based Modeling and Simulation on emergency", Complex 2009, Part II, LNICST 5, 1451 – 1461, 2009.

[10] Zaharia M. H., Leon F., Pal C., & Pagu G., "Agent-Based Simulation of Crowd Evacuation Behavior", International Conference on Automatic Control, Modeling and Simulation, 529-533, 2011.

[11] Quang C. T., & Drogoul A., "Agent-based simulation: definition, applications and perspectives", Invited Talk for the biannual Conference of the Faculty of Computer Science, Mathematics and Mechanics, 2008.

[12] Cole J. W., Sabel C. E., Blumenthal E., Finnis K., Dantas A., Barnard S., & Johnston D. M., "GIS-based emergency and evacuation planning for volcanic hazards in New Zealand", Bulletin of the New Zealand society for earthquake engineering, vol. 38, no. 3, 2005.

[13] Batty M., "Agent-Based Technologies and GIS: simulating crowding, panic, and disaster management", Frontiers of geographic information technology, chapter 4, 81-101, 2005.

[14] Patrick T., & Drogoul A., "From GIS Data to GIS Agents Modeling with the GAMA simulation platform", TF SIM 2010.

[15] Quang C. T., Drogoul A., & Boucher A., "Interactive Learning of Independent Experts' Criteria for Rescue Simulations", Jmynal of Universal Computer Science, Vol. 15, No. 13, 2701-2725, 2009.

[16] Taillandier T., Vo D. A., Ammyoux E., & Drogoul A., "GAMA: a simulation platform that integrates geographical information data, agentbased modeling and multi-scale control", In Proceedings of Principles and practice of multi-agent systems, India, 2012.

AUTHORS PROFILE

Kohei Aarai He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 and also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission-A of ICSU/COSPAR since 2008. He wrote 33 books and published 510 journal papers. He is Editor-in-Chief of International Journal of Advanced Computer Science and Applications as well as I ternational Journal of Intelligent Systsems and Applications.

Evaluation of Cirrus Cloud Detection Accuracy of GOSAT/CAI and Landsat-8 with Laser Radar: Lidar and Confirmation with Calipso Data

Kohei Arai[1]
1Graduate School of Science and Engineering
Saga University
Saga City, Japan

Masanori Sakashita[1]
1Graduate School of Science and Engineering
Saga University
Saga City, Japan

***Abstract*—Cirrus cloud detection accuracy of GOSAT/CAI and Landsat-8 is evaluated with a ground based Laser Radar: Lidar data and sky view camera data. Also, the evaluation results are confirmed with Calipso data together with a topographic representation of vertical profile of cloud structure. Furthermore, origin of cirrus clouds is estimated with forward trajectory analysis. The results show that GOSAT/CAI de4rived cirrus clouds is not accurately enough due to missing of cirrus cloud detection spectral channel while Landsat-8 derived cirrus cloud.**

***Keywords*—Cirrus cloud; GOSAT/CAI; Landsat; LiDAR; Sky view camera; Calipso; topogramphic representation of 3D clouds**

I. Introduction

Cloud detection is one of though issues in satellite remote sensing in particular for cirrus clouds [1]-[16]. It is not so easy to detect cirrus clouds in particular for remote sensing satellite onboard instruments. In order to detect cirrus clouds, 1.38 micrometer wavelength channel is adopted for Moderate resolution of Imaging Spectrometer: MODIS[1] and Landsat-8 Operational Land Imager: OLI[2], etc. Green house gasses Observation Satellite / Cloud and Aerosol Imager: GOSAT[3]/CAI[4] is dedicated sensor for cloud and aerosol retrievals. Because that GOSAT/FTS (Fourier Transform Spectrometer[5]) data is affected by clouds and aerosols, GOSAT/CAI is carried on the same platform of GOSAT satellite. Therefore, cloud flag and its confidence level are evaluated from the GOSAT/CAI and provide as Level 2 of GOSAT products together with Level 1B product as a source of Level 2 product. As mentioned above, it is not so easy to detect cirrus clouds. Although cirrus detection wavelength channel (1.38 micrometer) is required for detection of cirrus clouds, GOSAT/CAI does not have such channel. Therefore, it is not possible to detect cirrus cloud essentially. On the other hand, Landsat-8/OLI has cirrus detection channel. It is expected that cirrus detection can be done with Landsat-8 OLI data. Thus, cirrus cloud screening can be done for GOSAT/FTS observations.

In order to check the capability of cirrus cloud detection, Light Detection and Ranging、Laser Imaging Detection and Ranging: LiDAR data which allows measurement of back scattering ratio and depolarization ratio is used [17]-[29]. The ground based LiDAR is equipped at one of the GOSAT validation sites which is situated at Saga University, Japan. Therefore, vertical profile of aerosol particles as well as cloud particles are detected which results in detection of aerosols and clouds including cirrus clouds. Meantime, sky view camera observes hemispherical cloud conditions. Although it is possible to detect thick clouds, it is not easy to detect cirrus clouds with sky view camera. Vertical cloud structure can be retrieved with Cloud Aerosol Lidar and Infrared Pathfinder Satellite Observations: Calipso[6] data. Therefore, detected cirrus clouds can be validated with Calipso data. In this paper, a specific representation of vertical cloud structure is proposed. That is to representation of the retrieved structure on the topographic map which is projected on the globe. Forward trajectory analysis is also made for retrievals of the original source areas of the cirrus in concern through consideration of atmospheric conditions.

In the next section, the proposed method for evaluation of cirrus detection accuracy of GOSAT/CAI and Landsat-8 is described followed by experiments (method and procedure as well as the results from the experiments). Then validation of the evaluation results with sky view camera data and Calipso data is described followed by the specific representation of vertical cloud structure on the earth. Finally, conclusion and some discussion are followed.

II. Method and Procedure

A. GOSAT/FTS and CAI

GOSAT satellite is operating since January 23 2009 as the joint project among Ministry of Environment, JAXA and National Institute Environmental Science: NIES. GOSAT carries FTS and CAI as mission instruments as shown in Fig.1.

Major mission of GOSAT is to measure total column of carbon dioxide and methane which can be done with FTS instrument. In order to avoid influence due to aerosols and clouds, TANSO/CAI is also carried on GOSAT.

[1] http://modis.gsfc.nasa.gov/
[2] http://landsat.usgs.gov/landsat8.php
[3] http://www.gosat.nies.go.jp/
[4] http://www.gosat.nies.go.jp/eng/gosat/page2.htm
[5] https://en.wikipedia.org/wiki/Fourier_transform_infrared_spectroscopy
[6] http://www.icare.univ-lille1.fr/drupal/calipso

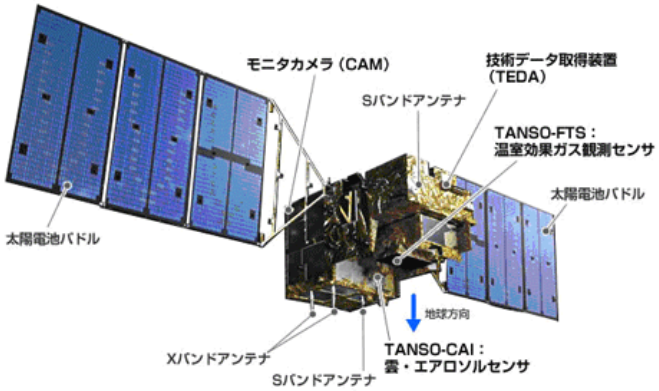


Fig. 1. Mission instruments onboard GOSAT satellite

B. GOSAT Validation Site

There are TCCON validation sites in the world. One of these is Saga University site in Japan. The location is shown in Fig.2. Fig.3 shows the LiDAR site (Laser light and the container in which LiDAR is equipped).

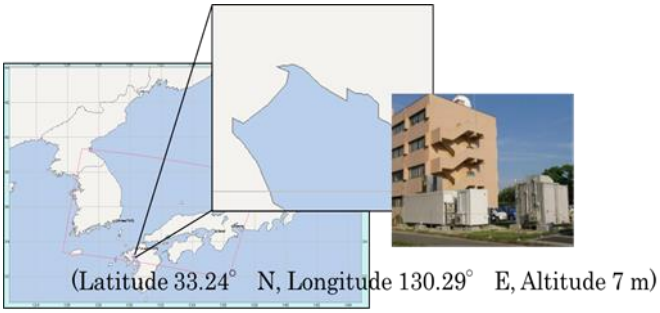


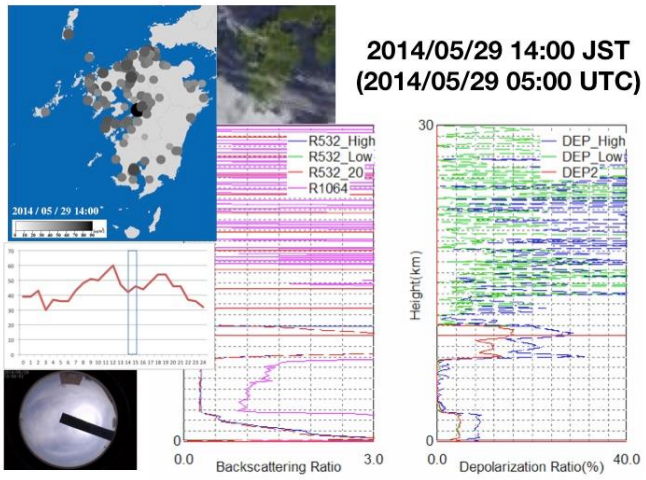
Fig. 2. Saga University TCCON site for GOSAT validation

Examples of the LiDAR data are shown in Fig.4 (a) together with PM2.5 data, CAI imagery data, the time series of PM2.5 data, and the sky view camera image. The right bottom graphs are the LiDAR data which is acquired at 14:00 Japan Standard time on May 29 2014. On the left, there is back scattering ratio data is situated while depolarization ratio is shown on the right. From these back scattering ratio and depolarization ratio, aerosol distribution and cloud vertical profile can be retrieved. Therefore, LiDAR data derived cloud vertical profile can be used for validation of CAI data derived clouds and Landsat-8 data derived clouds in particular, cirrus clouds. Meanwhile, Fig.4 (b) shows the LiDAR data which is acquired on April 26 2015. There is a peak of back scattered photon counts at around 10km of elevation (altitude above sea level). It is cirrus clouds. There is the ground based FTS for GOSAT validation which is situated just beside the LiDAR as shown in Fig.5. Other than these, there are sky view camera, sky radiometer which allows estimation of aerosol particle size distribution and refractive index retrievals for GOSAT validation. Examples of sky camera imagery data are shown in Fig.6 ((a) is for clear sky while (b) is for cloudy sky).

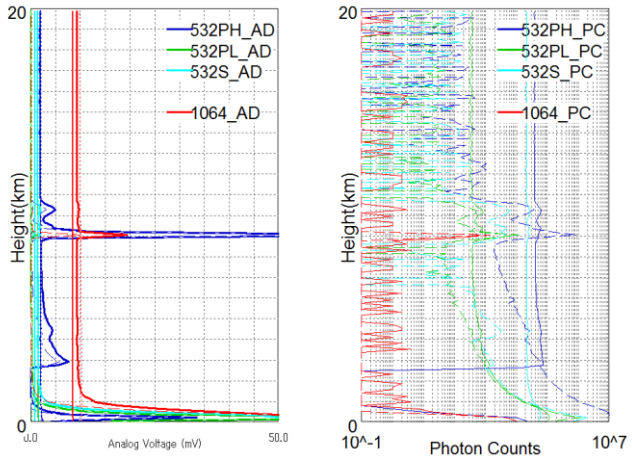
(a)Laser light

(b)Container

Fig. 3. LiDAR at Saga University GOSAT validation site



(a)Several data



(b)LiDAR data

Fig. 4. Examples of the LiDAR data together with the other measured data

Fig. 5. Ground based FTS for GOSAT validation

(a)Clear sky (b)Cloudy sky

Fig. 6. Examples of sky view camera data

C. Cloud Products Derived from CAI

There are two Level 2 cloud products, Cloud flag with 0 or 1 and Confidence level ranged from 0 to 1. Fig.7 shows lower and upper limits of clouds and clear sky. Using these definitions, four statistical tests are applied to the CAI data as shown in equation (1).

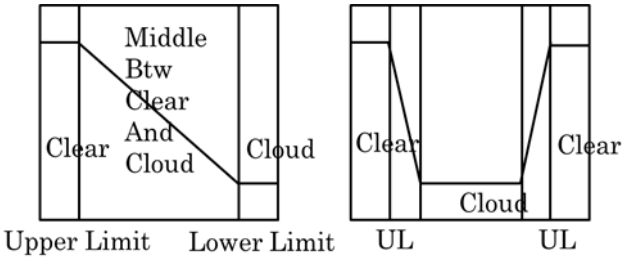


Fig. 7. Definition of lower and upper limits

$$Q = 1 - \sqrt[n]{(1-F_1)(1-F_2)...(1-F_n)} \quad (1)$$

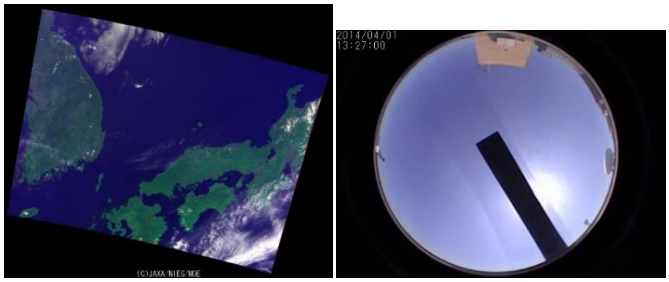
where F_i denotes statistical test results which are shown in Table 1.

TABLE I. FOUR TESTS FOR CALCULATION OF CONFIDENCE LEVEL IN CLOUD DETECTION

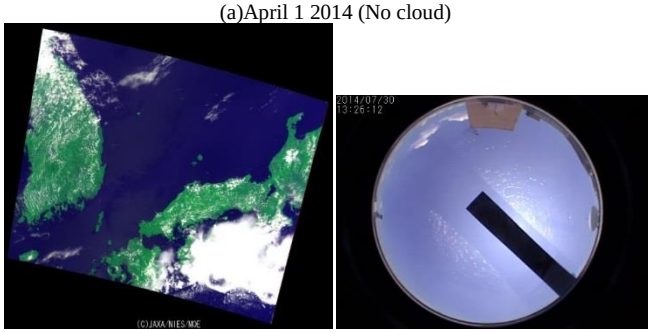
Tests		Lower Limit	Upper Limit
Band2R		＋0.195	+0.045
Band3R /Band2R	Smaller END LARGER END	0.9 1.1	0.66 1.7
NDVI	Smaller END LARGER END	-0.1 0.22	-0.22 0.46
Band3R /Band4R		1.06	0.86

D. GOSAT/CAI and Landsat-8 Imagery Data

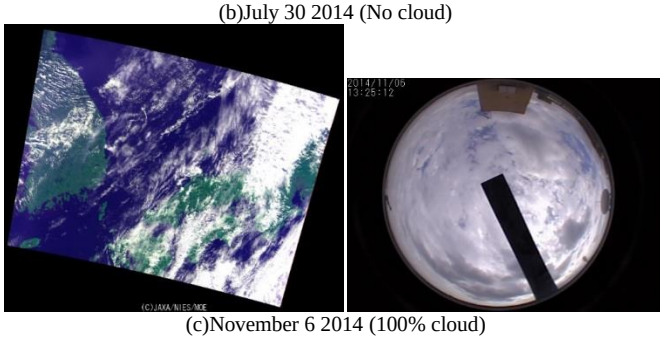
Fig.8 shows examples of the acquired color images of GOSAT/CAI imagery data together with sky view camera images which area acquired at the Saga University validation site at the same time as satellite over pass time.



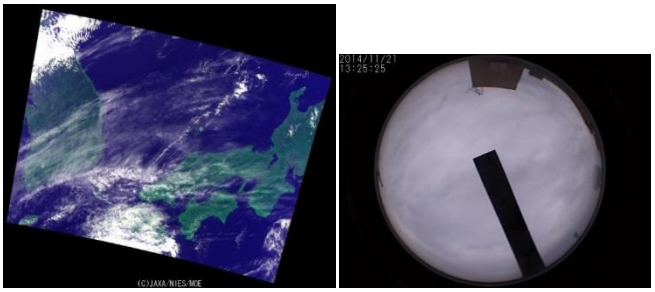
(a)April 1 2014 (No cloud)



(b)July 30 2014 (No cloud)



(c)November 6 2014 (100% cloud)



(d)November 21 2014 (100% cloud)

Fig. 8. Examples of GOSAT/CAI images and the sky view camera images at the Saga University validation site at the satellite over pass time

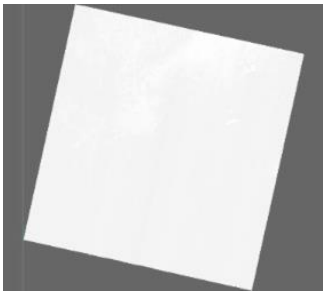
III. EXPERIMENTS

A. *Match-Up Data Between CAI and LiDAR as well as Landsat-8 OLI*

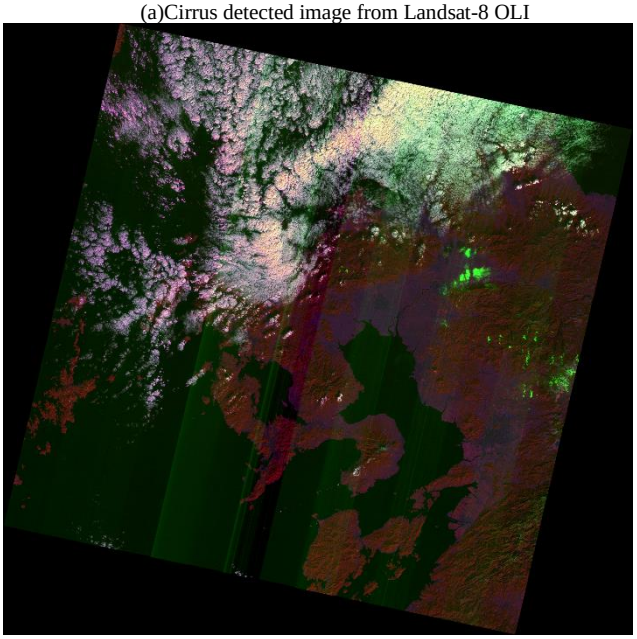
In order to evaluate cirrus detection capability of CAI and Landsa-8 OLI, match-up data have been searched for the term of the first half year of 2015. As the results from the search, the following two data are found,

January 20
April 26

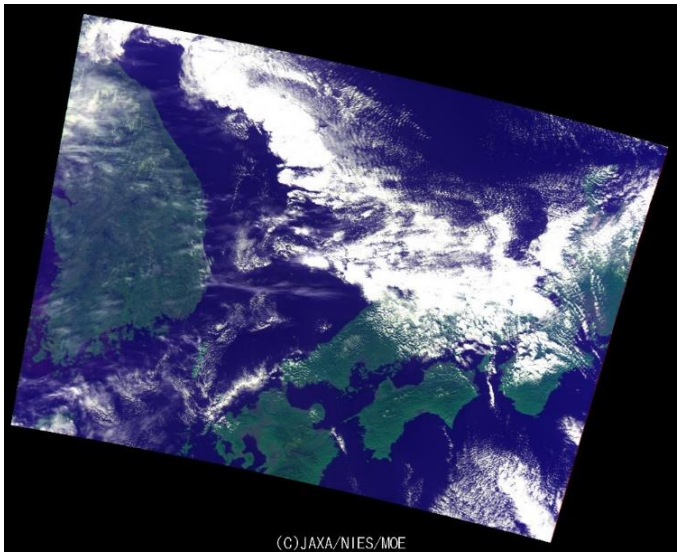
CAI and Landsad-8 OLI images of the match-up data on January 20 2015 are shown in Fig.9 while those for April 26 2015 are shown in Fig.10.



(a)Cirrus detected image from Landsat-8 OLI

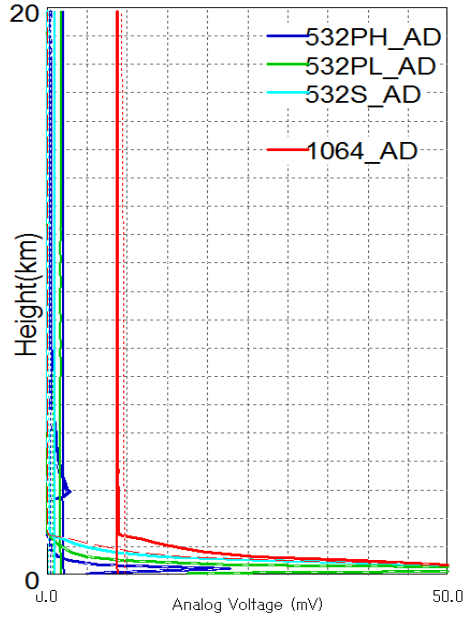


(b)Landsat-8 OLI image



(c)CAI image

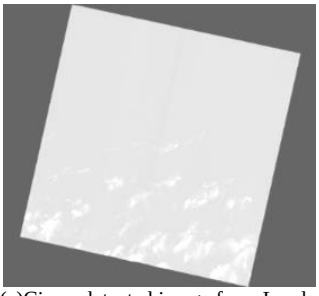
(d)Sky view image



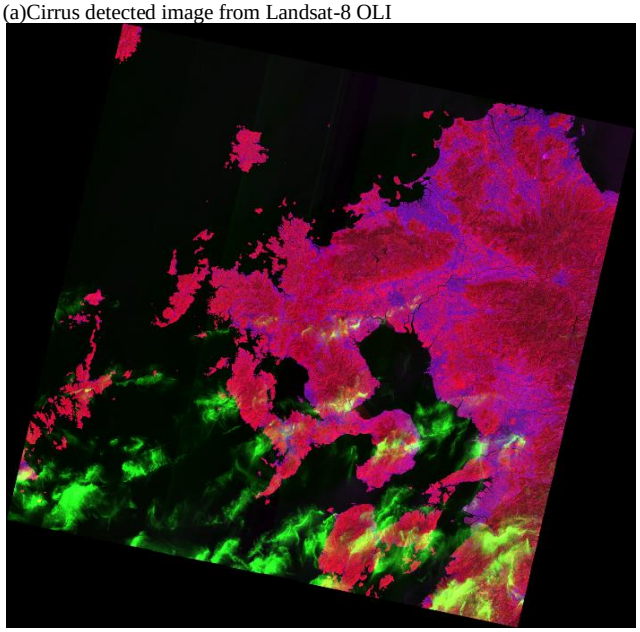
(e)LiDAR data

Fig. 9. Match-up data on January 20 2015

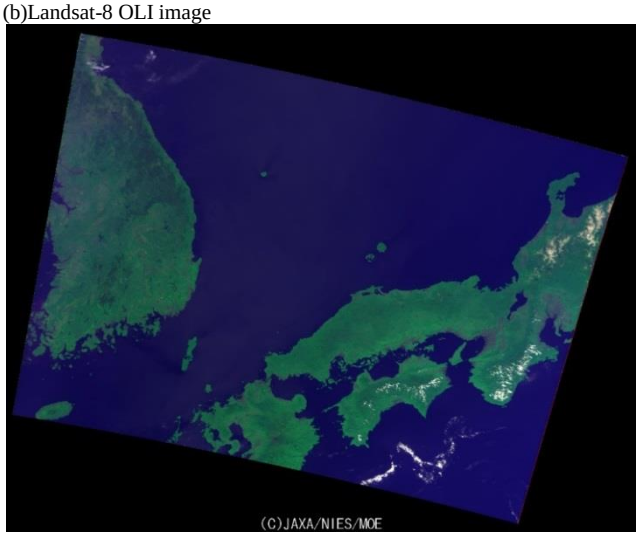
(a) shows cirrus detected image from Landsat-8 OLI imagery data. OLI band 9 is cirrus channel of which sensitive to 1.38 micrometer of wavelength. (b) shows color composite image of which red color is assigned to the near infrared channel, green color is assigned to band 9 of cirrus channel and blue color is assigned to mid-infrared wavelength channel, respectively. (c) shows natural color composite image of CAI.



(a)Cirrus detected image from Landsat-8 OLI

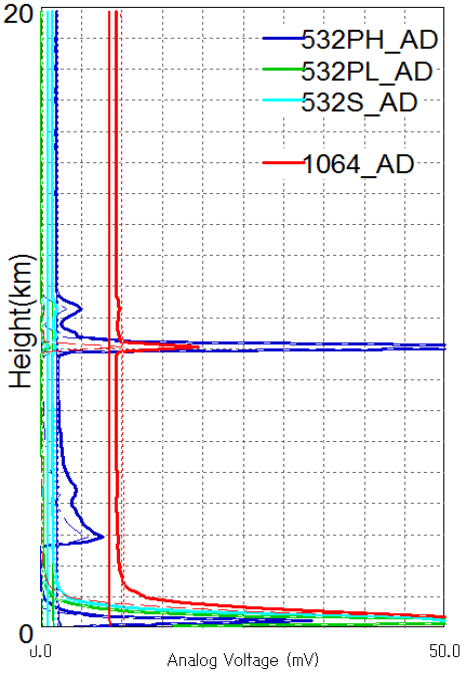


(b)Landsat-8 OLI image



(c)CAI image

(d)Sky view camera image



(e)LiDAR data

Fig. 10. Match-up data on April 26 2015

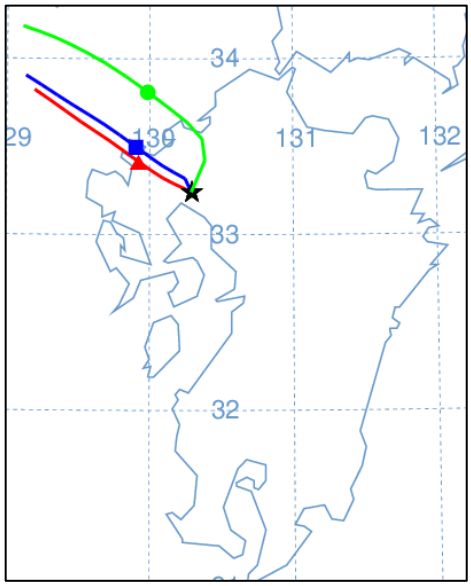
(d) shows sky view camera image while (e) shows LiDAR back scattered photon count data as the function of altitude. It is quite obvious that there is no cirrus cloud on January 20 2015 while there are cirrus clouds on April 26 2015. Green colored areas or pixels in (b) Landsat-8 OLI images indicate cirrus clouds.

Although Landsat-8 OLI image shows the cirrus clouds pixels in Fig.10 (b), Fig.10 (c) does not indicate any cirrus cloud at all. On the other hand, LiDAR data shows evidence of cirrus cloud existing as shown in Fig.10 (e). Therefore, it may say that Landsat-8 band 9 of cirrus channel does work to detect cirrus cloud while GOSAT/CAI does not work for detection of cirrus cloud due to missing cirrus channel.

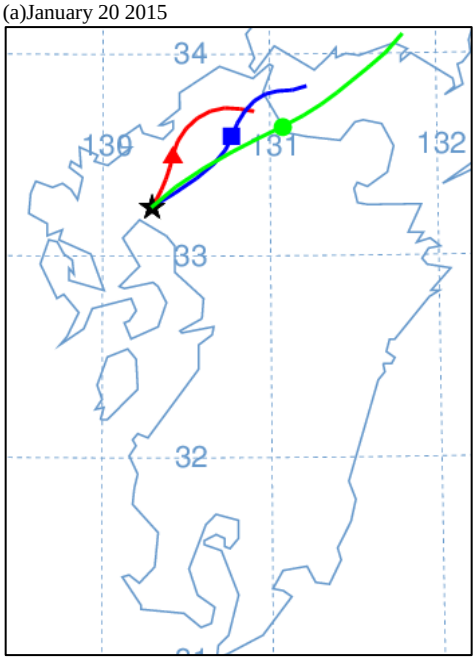
B. Adjectment of Acquisition Time Difference Between Landsat-8 and GOSAT

Local mean times of the orbits of Landsat-8 and GOSAT are different each other for 30 minutes. Therefore, some adjustment of the time difference between both is required. By using forward trajectory analysis software tool provided by NOAA, original positions of cirrus cloud (30 minutes before the acquisition time) are estimated. The results from the forward trajectory analysis are shown in Fig.11. For January 20 2015, there is North-West wind while there is North-East wind for April 26 2015.

Therefore, the cirrus cloud locations are shifted for the distance which is shown in Fig.11 within 30 minutes in those directions. Thus the cirrus cloud detection accuracy can be done through comparisons between LiDAR data and Landsat-8 OLI data which is acquired at 30 minutes apart from the LiDAR acquisition.



(a)January 20 2015



(b)April 26 2015

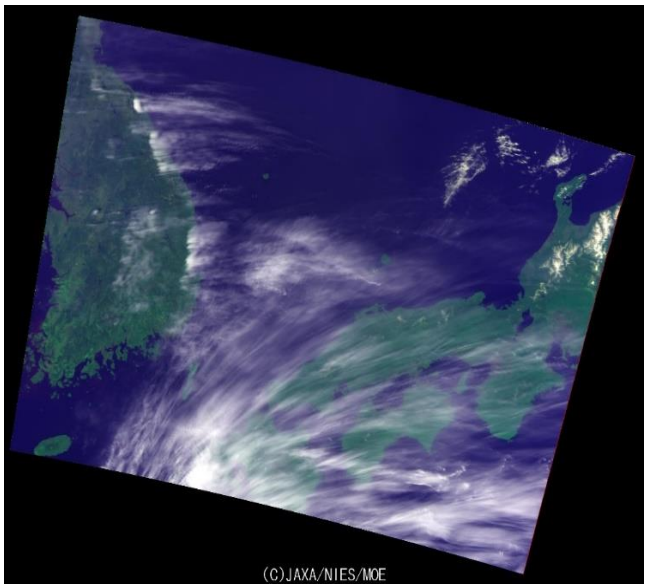
Fig. 11. Results from forward trajectory analysis

C. Summary of the Experimental Results

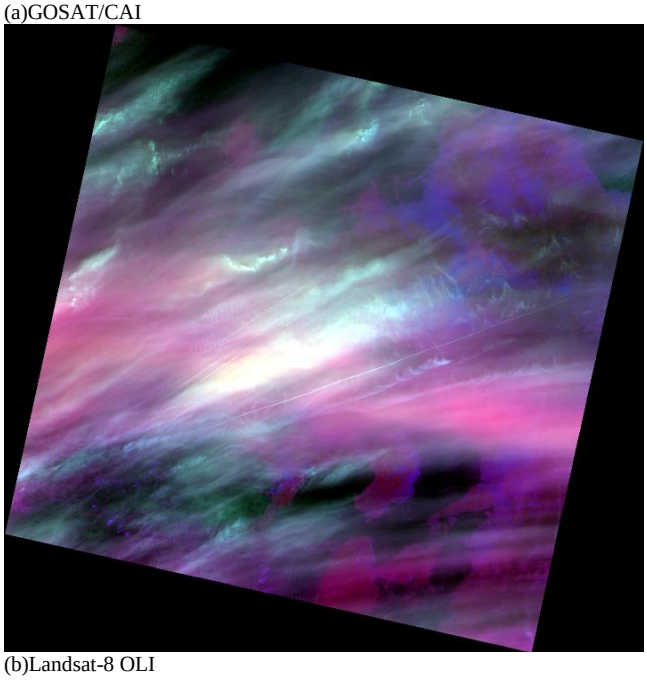
LiDAR data are acquired for 173 days within 518 days from April 1 2014 to August 31 2015 (Revisit cycle of the GOSAT satellite is 3 days). Within 173 days, LiDAR data are acquired 48 days (Acquisition ratio of LiDAR data to the total available days is just 33.01%). Cirrus clouds are observed for 11 days out of 48 days. Meanwhile, cirrus clouds are detected with CAI for just 8 days out of 11 days. On the other hand, cloud free situations are found with CAI for 18 days out of 37 days which is confirmed with LiDAR data. Due to the fact that the revisit cycle of Landsat-8 satellite is 16 days, just two match-up data between LiDAR and Landsat-8 OLI are collected for check cirrus detection accuracy. Two of match-up data show good coincidence between Landsat-8 OLI data utilized cirrus detection (no cirrus cloud and cirrus cloud existing situations).

D. Another Comparison Between Landsat-8 OLI data and GOSAT/CAI Imagery Data

Another match-up data between Landsat-8 OLI and GOSAT/CAI imagery data is found for April 7 2014 (Unfortunately LiDAR data is not acquired on that day). Fig.12 (a) shows GOSAT/CAI imagery data, (b) shows Landat-8 OLI imagery data on that day. Meanwhile, Fig.12 (c) shows the sky view camera image while (d) shows the results from the forward trajectory analysis for adjustment of the data acquisition time difference of 30 minutes between GOSAT/CAI and Landsat-8 OLI data.

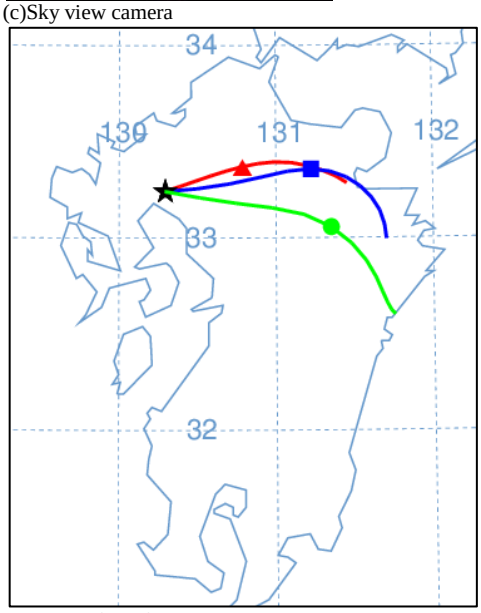


(a)GOSAT/CAI



(b)Landsat-8 OLI

(c)Sky view camera



(d)Results from forward trajectory analysis

Fig. 12. Another match-up data between GOSAT/CAI and Landsat-8 OLI

Although GOSAT/CAI cloud product indicates there are cirrus clouds at the intensive study area of Saga University GOSAT Validation site, Landsat-8 OLI indicates there is no cirrus cloud at all. As shown in Fig.12 (c), there are thick clouds in the sky above the test site at the GOSAT satellite over pass time. These, however, are not cirrus clouds at all. Forward trajectory analysis result shows that there is West wind at that time.

E. Confirmation of Cirrus Cloud Detection Capability with Calipso Data

Cirrus clouds can be confirmed with Calipso data. As shown in Fig.13, vertical profile of the existing clouds are investigated with Calipso data.

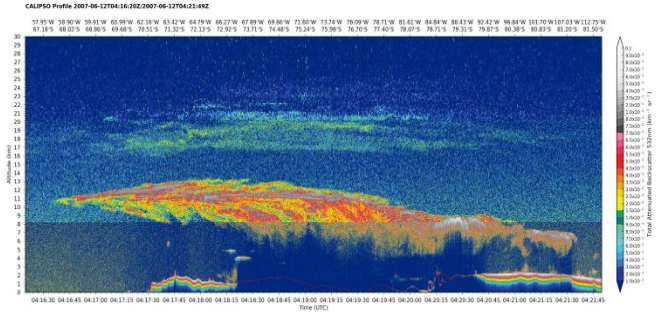


Fig. 13. Calipso data derived vertical profile of cloud structure

The horizontal axis of the Fig.13 is sub-satellite track while vertical axis shows back scattered phone count from the cloud particles.

IV. CONCLUSION

Cirrus cloud detection accuracy of GOSAT/CAI and Landsat-8 is evaluated with a ground based Laser Radar: Lidar data and sky view camera data. Also, the evaluation results are confirmed with Calipso data together with a topographic representation of vertical profile of cloud structure. Furthermore, origin of cirrus clouds is estimated with forward trajectory analysis. The results show that GOSAT/CAI de4rived cirrus clouds is not accurately enough due to missing of cirrus cloud detection spectral channel while Landsat-8 derived cirrus cloud.

ACKNOWLEDGEMENTS

Authors would like to thank Dr. Shuji Kawakami of JAXA, Prof. Dr. Hirofumi Ohyama of Nagoya University, Dr. Isamu Morino and Dr. Osamu Uchino of NIES and their research staff for their efforts to conduct the experiments and valuable discussions.

REFERENCES

[1] Remote Sensing Society of Japan Edt. Kohei Arai, et al., Remote Sensing -An Introductory Textbook-, Maruzen Planet Publishing Co. Ltd., Chapter 9 Radiometric Correction and Cloud Detection, 155-172, p.301, ISBN978-4-86345-185-8, 2013.

[2] Kohei Arai, A Merged Dataset for Obtaining Cloud Free IR Data and a Cloud Cover Estimation within a Pixel for SST Retrieval, Asian-Pacific Remote Sensing Journal, Vol.4, No.2, pp.121-127, Jan.1992

[3] Kohei Arai, Yasunori Terayama, Yoko Ueda, Masao Moriyama, Cloud Coverage Estimation Within a Pixel by Means of Category decomposition, Journal of the Journal of Japan Society of Photogrammetry and Remote Sensing,, Vol.31, No.5, pp.4-10, Oct.1992.

[4] Kohei Arai, Tasuya Kawaguchi, Adjacency Effect Taking Into Account Layered Clouds Based on Monte Carlo Method, Journal of Remote Sensing Society of Japan, Vol.21, No.2, pp.179-185, (2001).

[5] Kohei Arai, Tatsuya Kawaguchi, Adjucency Effect of Layered Clouds taking Into Account Phase Function of Cloud Particles and Multi-Layered Plane Parallel Atmosphere Based on Monte Carlo Method, Journal of Japan Society of Photogrammetry and Remote Sensing, Vol.40, No.6, 2001.

[6] Kohei Arai, Adjacency effect of layered clouds estimated with Monte-Carlo simulation, Advances in Space Research, Vol.29, No.19, 1807-1812, 2002.

[7] K.Arai, Merged dataset with MOS-1/VTIR and NOAA/AVHRR enhancing cloud detectability, Proc. of the 28th COSPAR Congress, MA4-3.1, 1-8, 1990.

[8] K.Arai, Estimation of partial cloud coverage within a pixel, Proc. of the Pre-ISY International Symposium, 99-106, 1991.

[9] K.Arai, Sea surface temperature estimation taking into account partial cloud within a pixel, Proc. of the ISY conference on Earth and Space Information Systems, 10/13, 1992.

[10] K.Arai, Y.Ueda and Y.Terayama, Comparative study on estimation of partial cloud coverage within a pixel -Proposed adaptive least square method with constraints- Proc. of the European ISY Conference, 305/310, 1992.

[11] K.Arai, SST estimation of the pixels partially contaminated with cloud, Proc. of the Asian-Pacific ISY (International Space Year) Conference, 1992.

[12] Kohei Arai, Adjacency effect due to a box type of 3D clouds estimated with Monte Carlo simulation considering the phase function of cloud particles, Abstracts of the 33rd COSPAR Scientific Assembly, A1.2-0061, Warsaw, Poland, July 16-23, (2000).

[13] S.Sobue, K.Arai and N.Futamura, Development of a method of cloud detection in Japanese Earth observation satellites, Proceedings of the ISTS (International Space Science and Technology Symposium), N-6, 2002-n-21,2002.

[14] Shin-ichi Sobua and Kohei Arai, Development of method for cloud detection in ASTER image data, Proc. of the 24th International Symposium on Space Technology and Science (ISTS), n-01, (2004)

[15] T. Sakai, O. Uchino, I. Morino, T. Nagai, S. Kawakami, H. Ohyama, A. Uchiyama, A. Yamazaki, K.Arai, H. Okumura, Y. Takubo, T. Kawasaki, T. Akaho, T. Shibata, T. Nagahama, Y. Yoshida, N. Kikuchi, B. Liley, V. Sharlock, J. Robinson, T. Yokota, Impact of aerosol and cirrus clouds on the GOSAT observed CO2 and CH4 inferred from ground based lidar, skyradiometer and FTS data at prioritized observation sites,(2013), Proceedings of the 9th International Workshop on Greenhouse Gas measurements from Space, IWGGMS-9, 2013

[16] H.Okumura, Kohei Arai, Improvement of PM2.5 density distribution visualization system using ground-based sensor network and Mie Lidar, Proceedings of the Conference on Remote Sensing of Clouds and the Atmosphere, SPIE Remote Sensing, ERS 15-RS 104-50, 2015

[17] Osamu Uchino, Tetsu Sakai, Tomohiro Nagai, Masahisa Nakasato, Isamu Morino, Tatsuya Yokota, Tsuneo Matsunaga, Nobuo Sugimoto, Kohei Arai, Hiroshi Okumura, Development of transportable Lidar for validation of GOSAT satellite data products, Journal of Remote Sensing Society of Japan, 31, 4, 435-443, 2011

[18] O.Uchino, T.Sakai, T.Nagai, I.Morino, K.Arai, H.Okumura, S.Takubo, T.Kawasaki, Y.mano, T.Matsunaga, T.Yokota, On recent stratspheric aerosols observed by Lidar over Japan, Journal of Atmospheric Chemistry and Physics, 12, 11975-11984, 2012(doi:10.5194/acp-12, 11975-2012).

[19] Testu Sakai, Osamu Uchino, Isamu Morino, Tomohiro Nagai, Taiga Akaho, Kawasaki Tsuyoshi, Tetsu Sakai, Hiroshi Okumura, Kohei Arai, Akihiro Uchiyama, Akehiro Yamazaki, Tsuneo Matsunaga, Tatsuya Yokota, Vertical profile of volcanic prumes form Sakurajima volcano detected by Lidar and skyradiometer situated Saga and its optical property, Journal of Remote Sensing Society of Japan, 34, 3, 197-204, 2014

[20] 314. H. Okumura, S.Takubo, T.Kawsaki, I.N.Abdulah, T.Sakai, T.Maki, K.Arai, Web based data acquisition and management system for GOSAT validation Lidar data analysis, Proceedings of the SPIE Vol.8537, Conference 8537: Image and Signal Processing for remote Sensing , Paper #8537-43, system, 2012.

[21] 315. T.Sakai, H. Okumura, T.Kawsaki, I.N.Abdulah, O.Uchino, I.Morino, T.Yokota, T.Nagai, T.Sakai, T.Maki, K.Arai, Observation of aerosol parameters at Saga using GOSAT product validation Lidar, Proceedings of the SPIE Vol.8526, Conference 8526: Lidar Remote Sensing for Environmental Monitoring XIII, SPIE Asia-Pacific Remote Sensing, Paper #8295A-50,IP1, 2012.

[22] 332. Hiroshi Okumura, Shoichiro Takubo, Takeru Kawasaki, Indra Nugraha Abdulah, Osamu Uchino, Isamu Morino, Tatsuya Yokota, Tomohiro Nagai, Tetu Sakai, Takashi Maki, Kohei Arai, Improvement of web-based data acquisition and management system for GOSAT validation Lidar data analysis(2013), SPIE Electronic Imaging Conference, 2013.

[23] 335. Hiroshi Okumura, Kohei Arai, Observation of aerosol properties at Saga using GOSAT product validation LiDAR, Proceedings of the Conference on Image and Signal Processing fo Remote Sensing, SPIE #ERS13-RS107-38, 2013

[24] 337. T. Sakai, O. Uchino, I. Morino, T. Nagai, S. Kawakami, H. Ohyama, A. Uchiyama, A. Yamazaki, K.Arai, H. Okumura, Y. Takubo, T. Kawasaki, T. Akaho, T. Shibata, T. Nagahama, Y. Yoshida, N. Kikuchi, B. Liley, V. Sharlock, J. Robinson, T. Yokota, Impact of aerosol and cirrus clouds on the GOSAT observed CO2 and CH4 inferred from ground based lidar, skyradiometer and FTS data at prioritized observation sites,(2013), Proceedings of the 9th International Workshop on Greenhouse Gas measurements from Space, IWGGMS-9, 2013

[25] 338. Shuji Kawakami, Hirofumi Ohyama, Kei Shiomi, T.Fukamachi, Kohei Arai, C.Taura, H.Okumura, Observations of carbon dioxide and methane column amounts measured by high resolution of FTIR at Saga in 2011-2012, Proceedings of the International Symposium on Remote Sensing, ISRS-TCCOC 2013.(2013)

[26] 339. Osamu Uchino, T.Sakai, T.nagai, I.Morino, H.Ohyama, S.Kawakami, K.Shiomi, T Kawasaki, T.Akaho, H.Okumura, Kohei Arai, T.matsunaga, T.Yokota, Comparison of lower tropospheric ozone column observed by DIAL and GOSAT TANSO-FTS TIR, Proceedings of the AGU Fall Meeting 2013.(2013)

[27] 340. I Morino, T Sakai, T.Nagai, A.Uchiyama, A.Yamazaki, S Kawakami, H.Ohyama, Kohei Arai, H.Okumura, T.Shibata, T.Nagahama, N.Kikuchi, Y.Yoshida, Ben Liley, Vannessa Sherlock, John Robinson, O. Uchino, T.Yokota, Impact of aerosols and cirrus on the GOSAT onboard CO2 and CH4 inferred from ground based Lidar, skyradiometer and FTS data at prioritized observation sites, Proceedings of the AGU Fall Meeting 2013.(2013)

[28] 343. H.Okumura, K.Arai, Development of PM2.5 density distribution visualization system using gournd-level sensor network and Mie-Lidar, Proceedings of the SPIE European Remote Sensing Coference, ERS 14-RS107-97, 2014.

[29] 346. H.Okumura, Kohei Arai, Improvement of PM2.5 density distribution visualization system using ground-based sensor network and Mie Lidar, Proceedings of the Conference on Remote Sensing of Clouds and the Atmosphere, SPIE Remote Sensing, ERS 15-RS 104-50, 2015

AUTHORS PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission "A" of ICSU/COSPAR since 2008. He wrote 33 books and published 510 journal papers. He is now Editor-in-Chief of IJACSA and IJISA.

An Empirical Comparison of Tree-Based Learning Algorithms: An Egyptian Rice Diseases Classification Case Study

Mohammed E. El-Telbany
Computers and Systems Department
Electronics Research Institute Cairo,
Egypt

Mahmoud Warda
Computers Department
National Research Center
Cairo, Egypt

***Abstract*—Applications of learning algorithms in knowledge discovery are promising and relevant area of research. The classification algorithms of data mining have been successfully applied in the recent years to predict Egyptian rice diseases. Various classification algorithms can be applied on such data to devise methods that can predict the occurrence of diseases. However, the accuracy of such techniques differ according to the learning and classification rule used. Identifying the best classification algorithm among all available is a challenging task. In this study, a comprehensive comparative analysis of a tree-based different classification algorithms and their performance has been evaluated by using Egyptian rice diseases data set. The experimental results demonstrate that the performance of each classifier and the results indicate that the decision tree gave the best results.**

***Keywords*—Data Mining, Classification, Decision Trees, Bayesian Network, Random Forest, Rice Diseases.**

I. Introduction

Processing the huge data and retrieving meaningful information from it is a difficult task. Data mining is a wonderful tool for handling this task. The major components of the architecture for a typical data mining system are shown in Fig 1. The term Data Mining, also known as Knowledge Discovery in Databases (KDD) refers to the non trivial extraction of implicit, previously unknown and potentially useful information from data in databases [1]. They are several different data mining techniques such as *clustering*, *association*, *anomaly detection* and *classification* [2]. The classification process has been identified as an important problem in the emerging field of data mining as they try to find meaningful ways to interpret data sets. The goal of classification is to correctly predict the value of a designated discrete class variable, given a vector of predictors or attributes by produces a mapping from the input space to the space of target attributes [3]. There are various classification techniques each technique has its pros and cons. Recently, Fernandez-Delgado *et al.* [4] evaluate 179 classifiers arising from 17 families (e.g. statistics, symbolic artificial intelligence and data mining, connectionist approaches, and others are ensembles). The classifiers show strong variations in their results among data sets, the average accuracy might be of limited significance if a reduced collection of data sets is used [4]. For example, the largest merit of neural networks (NN) methods is that they are general: they can deal

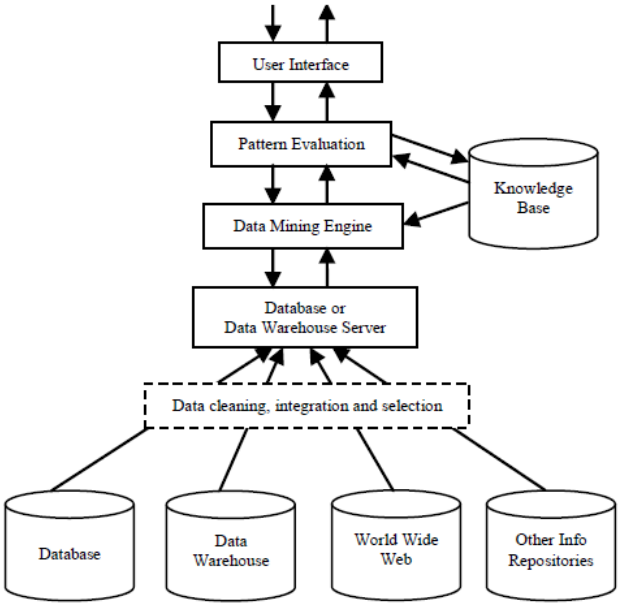


Fig. 1: Architecture of a Typical Data Mining System [1]

with problems with high dimensions and even with complex distributions of objects in the n-dimensional parameter space. However, the relative importance of potential input variables, long training process, and interpretative difficulties have often been criticized. Although the *support vector machine* (SVM) has a high performance in classification problems [5], the rules obtained by SVM algorithm are hard to understand directly and costly in computation. Due to the above-mentioned drawbacks of NN and SVM, the purpose of this paper, is to explore the performance of classification using various decision tree approaches which have the following advantages as follows [6]:

1) Decision trees are easy to interpret and understand;
2) Decision trees can be converted to a set of $if-then$ rules; and
3) Decision trees don't need priori assumptions about the nature of data, it is a *distribution-free*.

Since decision trees have the described advantages, they have proven to be effective tools in classification of Egyptian rice

disease problems [7]. Specially, the transfer of experts from consultants and scientists to agriculturists, extends workers and farmers represent a bottleneck for the development of agriculture on the national. This information can be used as part of the farmers decision-making process to help to improve crop production. The aim of this paper is to evaluate the tree-based classifiers to select the classifier which more probably achieves the best performance for the Egyptian rice diseases which cause losses that estimated by 15% from the yield, malformation of the leaves or dwarfing of the plants. Discovering and controlling of diseases are the main aims and have a large effect for increasing density of Fadden and increasing gain for farmer then increasing the national income. Actually, the original contribution of this research paper is to measure and compare the performances of tree-based classification algorithms for Egyptian rice diseases. In particular, we have focused on the Bayesian network, random forest algorithms, comparing its performances with a decision tree using a variety of performance metrics. In this paper, four classification algorithms are investigated and presented for their performance. Section II, presents the related previous work. The proposed used classification algorithms are explained in section III. In section IV, our problem is formally described. Section V, describes data set used in this paper. In section VI an experimental results described for investigated types of classification algorithms including their performance measures. Finally, the conclusions are explained in section VII.

II. Related Work

The objectives of applying data mining techniques in agriculture is to increase of productivity and food quality at reduced losses by accurate diagnosis and timely solution of the field problem. Using data mining classification algorithms, it become possible to discover the classification rules for diseases in rice crop [7], [8]. The image processing and pattern recognition techniques are used in developing an automated system for classifying diseases of infected rice plants [9]. They extracted features from the infected regions of the rice plant images by using a system that classifies different types of rice disease using self-organizing map (SOM) neural network. Feature selection stage was done using rough set theory to reduce the complexity of classifier and to minimize the loss of information where a rule base classifier has been generated to classify the different disease and provide superior result compare to traditional classifiers [9]. Also, SVM is used to disease identification in the rice crop from extracted features based on shape and texture, where a three disease leaf blight, sheath blight and rice blast are classified [10]. In another work, the brown spot in rice crop is identified using K-Means method for segmentation and NN for classification of disease [11]. The NN is used to identify the three rice diseases namely (i) Bacterial leaf blight, (ii) Brown spot, and (iii) Rice blast [12]. The fuzzy entropy and probabilistic neural network are used to identify and classifying the rice plant diseases. Developed a mobile application based on android operating system and features of the diseases were extracted using fuzzy entropy [13].

III. Classification Algorithms

A total of four classification algorithms have been used in this comparative study. The classifiers in Weka have been categorized into different groups such as Bayes and Tree based classifiers, etc. A good mix of algorithms have been chosen from these groups that include decision tree, Naive Bayes net, random trees and random forest. The following sections briefly explain about each of these algorithms.

A. Decision Tree

The first algorithm used for comparison is a decision tree, which generates a classification-decision tree for the given data-set by recursive partitioning of data [14]. The algorithm considers all the possible tests that can split the data set and selects a test that gives the best information gain. For each discrete attribute, one test with outcomes as many as the number of distinct values of the attribute is considered. For each continuous attribute, binary tests involving every distinct values of the attribute are considered. In order to gather the entropy gain of all these binary tests efficiently, the training data set belonging to the node in consideration is sorted for the values of the continuous attribute and the entropy gains of the binary cut based on each distinct values are calculated in one scan of the sorted data. This process is repeated for each continuous attributes [15], [16]. In particular entropy, for an attribute is defined as in equation 1.

$$H(X) = -\sum_{j}^{m} p_j log_2(p_j) \tag{1}$$

Where p_j is defined as $P(X = V_j)$, the probability that X takes value V_j, and m is the number of different values that X admits. Due to their recursive partitioning strategy, decision trees tend to construct a complex structure of many internal nodes. This will often lead to over fitting. Therefore, the decision tree algorithm exhibits meta-parameters that allow the user to influence when to stop tree growing or how to prune a fully-grown tree.

B. Random Decision Tree

The second chosen algorithm for the comparison is the *random decision tree* presented by Fan *et al.* in [17]. The Random decision tree is an ensemble learning algorithm that generates many individual learners. It employs a bagging idea to produce a random set of data for constructing a decision tree. In the standard tree each node is split using the best split among all variables. The choice is bind on the type of the attribute, in particular if the feature can assume values in a finite set of options it cannot be chosen again in the sub tree rooted on it. However, if the feature is a continuous one, then a random threshold is chosen to split the decision and it can be chosen again several times in the same sub tree accordingly with the ancestor's decision. To enhance the accuracy of the method, since the random choice may leads to different results, multiple trees are trained in order to approximate the true mean. Considering k as the number of features of the dataset and N as the number of trees, then the confidence probability to have is:

$$1 - (1 - \frac{1}{k})^N \tag{2}$$

Considering the k features of the dataset, and the i classifying attributes, the most diversity among trees is with depth of

$$\frac{k}{2} \tag{3}$$

since the maximum value of the combination is

$$\binom{i}{k} \tag{4}$$

Once the structure is ready the training may take place, in particular each tuple of the dataset train all the trees generated in order to read only one time the data. Each node counts how many numbers of examples go through it. At the end of the training the leaves contain the probability distribution of each class, in particular for the tree *i*, considering *n[y]* the number of instances of class y at the node reached by x, is:

$$P_i(y|x) = \frac{n[y]}{\sum_y n[y]} \tag{5}$$

The classification phase retrieves the probability distribution from each tree and average on the number of trees generated in the model:

$$P(y|x) = \frac{1}{N}\sum_{i=1}^{N} P_i(y|x) \tag{6}$$

C. Bayesian Network

Bayesian Networks encode conditional interdependence relationships through the position and direction of edges in a directed acyclic graph. The relationship between a node and its parent is quantified during network training. This classifier learns from training data the *conditional probability* of each attribute X_i given the class label C. Classification is then done by applying Bayes rule to compute the probability of C given the particular instances of $X_1 \ldots X_n$ and then predicting the class with the highest posterior probability. The goal of classification is to correctly predict the value of a designated discrete class variable given a vector of predictors or attributes [2]. In particular, the naive Bayes classifier is a Bayesian network where the class has no parents and each attribute has the class as its sole parent. Bayesian and neural network seem to be identical in their inner working. Their difference exist in the construction. Nodes in a neural network don't usually have clearly defined relationship and hidden node are more "discovered" than determined, whereas the relationships between nodes in Bayesian network are due to their conditional dependencies [18], [2].

D. Random Forest

The random forest classifier, described by Ho [19], [1], works by creating a bunch of decision trees randomly. Each single tree is created in a randomly selected subspace of the feature space. Trees in different subspaces complement each other's classifications. Actually, random forest is an ensemble classifier which consists of many decision tree and gives class as outputs i.e., the mode of the class's output by individual trees. Prediction is made by aggregating (majority vote for classification or averaging for regression) the predictions of the ensemble. Random Forests gives many classification trees without pruning [20]. The success of an ensemble strategy depends on two factors, the strength (accuracy) of individual base models and the diversity among them.

TABLE I: Possible value for each attribute from the Egyptian rice database

Attribute	Possible Values
Variety	gizal71,gizal77, gizal78 sakhal0l, sakhal02, sakhal03 , sakhal04
Age	Real values
Part	leaves, leaves spot, nodes, panicles, grains, plant, flag leaves, leaf sheath, stem
Appearance	Spots, elongate, spindle, empty, circular, oval, fungal, spore balls, twisted, wrinkled, dray, short, few branches, barren, small, deformed, seam, few stems, stunted, stones, rot, empty seeding
Colour	gray, olive, brown, brownish, whitish, yellow, green, orange, greenish black, white, pale, blackish, blac k
Temperature	Real values
Disease	Blight, brown spot, false smut, white tipe, stem rot

IV. Problem Definition

The main aim of this work is to produce a comparison among different inductive learning the *optimal model* for a target function $t = F(x)$, given a training set of size n, $(x_1; t_1), ..., (x_n; t_n)$, an inductive learner produces a model $y = f(x)$ to approximate the true function $F(x)$. Usually, there exists x such that $y \neq t$. In order to compare performance, a loss function is introduced $L(t, y)$. Given the loss function $L(t, y)$, that measures the discrepancy between our function's class and reality, where t is the true class and y is the predicted class, an optimal model is one that minimizes the average loss $L(t, y)$ for all examples. The optimal decision y^* for x is the class that minimizes the expected loss $E_t(L(t, y^*)$ for a given example x when x is sampled repeatedly.

V. Data Set Description

Rice is the worlds most common staple food for more than half of mankind. Because of its important, rice is considered a strategic resource in Egypt has been assigned as a high priority topic in its Agricultural Strategic Plans. Successful Egyptian rice production requires for growing a summer season (May to August) of 120 to 150 days according to the type of varieties as Gizal77 needs 125 day and Sakhal04 needs 135 day. Climate for the Egyptian rice is that daily temperature maximum = $30 - 35^o$, and minimum = $18 - 22^o$; humidity = 55%-65%; wind speed = $1 - 2m$. Egypt increase productivity through a well-organized rice research program, which was established in the early eighties. In the last decade, intensive efforts have been devoted to improve rice production. Consequently, the national average yields of rice increased by 65% i. e., from $(2.4t/fed.)$ during the lowest period $1984 - 1986$ to $(3.95t/fed.)$ in 2002 [21]. Many affecting diseases infect the Egyptian rice crop; some diseases are considered more important than others. In this study, we focus into the most important diseases, which are five; blight, brown spot, false smut, white tip nematode and stem rot sequence. Each case in the data set is described by seven attributes. We have a total of 206 samples and the attribute and possible values are listed in Table I.

VI. Experimental Evaluation

To gauge and investigate the performance on the selected classification methods or tree-based learning algorithms, many experiments are implemented within the WEKA framework

[22]. The Weka is an open source data mining workbench software which is used for simulation of practical measurements. It contains tools for data preprocessing, classification, regression, clustering, association rule and visualization. It does not only supports data mining algorithms, but also data preparation and meta-learners like bagging and boosting [22]. In order to test the efficiency of tree-based classification algorithms, training and test sets are used. Usually disjoint, subsets, the training set to build a classification tree(s) and the test set so as to check and validate the trained model. Also, cross-validation process applied where same sized disjoint sets are created so as to train the model fold wise. n-fold cross-validation, (usually $n = 10$) is used to divide the data into equally sized k subsets/ folds. In such case the model is trained using $(k-1)$ folds and the k^{th} fold is used as a test set. The whole process is repeated n times in an attempt to use all the folds for testing thus allowing the whole of the data to be used for both training and testing. In our data, ten cross-validation bootstraps, each with 138 (66%) training cases and $68(34\%)$ testing cases, were used for the performance evaluation. The simulation results are partitioned into two parts for easier analysis and evaluation. On the first part, correctly and incorrectly classified instances will be partitioned in percentage value and subsequently Kappa statistics, mean absolute error and root mean squared error will be in numeric value only. We also show the relative absolute error and root relative squared error in percentage for references and evaluation. The results of the simulation are shown in Tables II and III. Table II mainly summarizes the result based on accuracy and time taken for each simulation. Meanwhile, Table III shows the result based on error during the simulation.

TABLE II: Evaluation results of different classification algorithms

Alg.	Correctly %	Incorrectly %	time (sec.)	Kappa statistics
Decision Trees	97.57	2.42	0.01	0.97
Random Trees	94.66	5.33	0.07	0.92
Bayes Net	93.68	6.31	0.06	0.93
Random Forest	95.63	4.36	0.07	0.94

TABLE III: The errors of different classification algorithm

Alg.	Mean Abs. Error	Root Mean Squ. Error	Relative Abs. Error(%)	Root Relative Squ. Error(%)
Decision Tree	0.04	0.12	12.8	30.7
Random Trees	0.06	0.133	19.61	33.7
Bayes Net	0.129	0.199	41.31	50.4
Random Forest	0.036	0.124	11.44	31.4

Figure 2 shows the evaluation of different classification algorithms which are summarized in Table III. From the confusion matrix to analyse the performance criterion for the classifiers in disease detection accuracy, precision, recall and Mathews correlation coefficient (MCC) have been computed for the dataset as shown in Table IV. MCC is a special case of the linear correlation coefficient, and therefore also scales between +1 (perfect correlation) and -1 (anti correlation), with 0 indicating randomness. Accuracy, precision (specificity), recall (sensitivity) and MCC are calculated using the equations (7), (8), (9) and (10) respectively, where T_p is the number of true positives, T_n is the number of true negatives, F_p is the number of false positives and F_n is the number of false negatives.

$$accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{7}$$

$$specificity = \frac{T_p}{T_p + F_p} \tag{8}$$

$$sensitivity = \frac{T_p}{T_p + F_n} \tag{9}$$

$$MCC = \frac{T_p * T_n - F_p * F_n}{\sqrt{(T_p + F_n)(T_p + F_p)(T_n + F_n)(T_n + F_p)}} \tag{10}$$

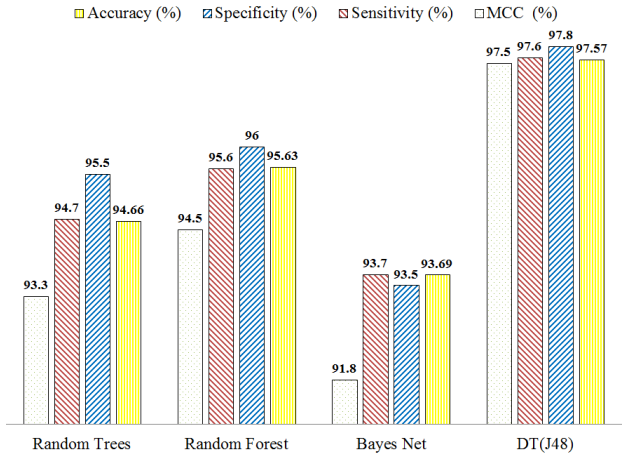


Fig. 2: The Roote Mean Square (RMSE) of each algorithm

TABLE IV: Accuracy, Specificity, Sensitivity and MCC of different classification algorithm

Alg.	Accuracy (%)	Specificity (%)	Sensitivity (%)	MCC
Decision Tree	97.57	97.8	97.6	0.95
Random Tree	93.69	93.5	93.7	0.92
Bayes Net	95.63	96.0	95.6	0.95
Random Forest	94.66	95.5	94.7	0.068

VII. Conclusions and Future Work

Data mining in agriculture is a very interesting research topic and can be used in many applications such as yields prediction, disease detection, optimizing the pesticide usage and so on. There are many algorithms that have been presented for classification in diagnosing the Egyptian rice diseases data set. However, we have choose four algorithms the J48 decision tree, Bayes net, random trees and random forest that belongs to the Tree-based category which are easy to interpret and understand. we conduct many experiments to evaluate the four classifiers for Egyptian rice diseases. The above analysis shows that for the J48 decision tree achieves highest sensitivity, specificity and accuracy and lowest RMS error, than Bayes net, random trees and random forest. J4.8 gave the best results due to the pruning process which simplify the tree and remove

unrelevant branches. Moreover, the random forest superior over random trees due to boosting process [23], [24].

Lastly, it should be mentioned that the predictive accuracy is the probability that a model correctly classifies an independent observation not used for model construction. A tree that involves irrelevant variables is not only more cumbersome to interpret but also potentially misleading. selecting an informative features and removing irrelevant/redundant features drastically reduces the running time of a learning algorithm and yields a more general classifier [25], [26]. So, in future works we intend to apply relevant methods for *feature selection* in classification to improve our results as a preprocessing stage before the classification process.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their revision and help to enhancement the paper writing. Also, we are indebted to Central Laboratory for Agricultural Expert Systems staff for providing us with their experiences and data set. And above all, God for His continuous guidance.

REFERENCES

[1] J. Han, M. Kamber and J.Pei, Data mining: concepts and techniques, Elsevier Inc., 3^{rd} edition, 2012.

[2] Y. Nong, Data mining: theories, algorithms, and examples, CRC Press, 2014.

[3] M. Zaki and W. Meira, Data mining and analysis: foundations and algorithms, Cambridge University Press, 2014.

[4] M. Fernandez-Delgado, E. Cernadas, S. Barro and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," in Machine Learning Research, 15, pp. 3133-3181, 2014.

[5] C. Bishop, Pattern recognition and machine learning, Springer New York, 2006.

[6] Y. Zhao and Y. Zhang, *Comparison of decision tree methods for finding active objects*, arXiv:0708.4274v1, 2007.

[7] M. El-Telbany, M. Warda and M. El-Borahy, "Mining the classification rules for Egyptian rice diseases," in International Arab Journal of Information Technology (IAJIT), Jordan, Vol. 3, No. 4, 2006.

[8] A. Nithya, V. Sundaram, "Classification rules for Indian Rice diseases," in International Journal of Computer Science (IJCSI), Vol. 8, Issue 1, 2011.

[9] S. Phadikar, J. Sil and A. Das, "Rice diseases classification using feature selection and rule generation techniques," in Comput. Electron. Agric., 90, pp. 7685, 2013.

[10] Q. Yao, Z. Guan, Y. Zhou, J. Tang, Y. Hu and B. Yang, "Application of support vector machine for detecting rice diseases using shape and color texture features," in International Conference on Engineering Computation, pp.79-83, 2009.

[11] D. Al-Bashish, M. Braik, S. Bani-Ahmad, "A Framework for Detection and Classification of Plant Leaf and Stem Diseases," in International Conference on Signal and Image Processing, pp. 113-118, 2010.

[12] J. Orillo, J. Cruz, L. Agapito, P. Satimbre and I.Valenzuela, "Identification of diseases in rice plant (oryza sativa) using back propagation Artificial Neural Network." in 7^{th} International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), 2014.

[13] K. Majid, Y. Herdiyeni and A.Rauf, "I-PEDIA: Mobile application for paddy disease identification using fuzzy entropy and probabilistic neural network," in International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2013.

[14] T. Hastie, R. Tibshirani and J. Friedman, The elements of statistical learning: data mining, inference and prediction, the Mathematical Intelligence, 27(2): pp. 83-85, 2005.

[15] J. Quinlan, "Induction of decision trees," in Machine Learning, 1(1), pp. 81-106, 1986.

[16] T. Mitchell, Machine Learning, McGraw Hill, 1997.

[17] W. Fan, H. Wang, P. Yu, and S. Ma, "Is random model better? On its accuracy and efficiency," in 3^{rd} IEEE International Conference on Data Mining, 2003.

[18] N. Friedman, D. Geiger and M. Goldszmidt, "Bayesian network classifiers," in Machine Learning, 29, pp. 131-163, 1997.

[19] K. Ho , "Random decision forests," in IEEE Proceedings of the 3^{d} International Conference on Document Analysis and Recognition, pp. 278-282, 2005.

[20] L. Breiman, Random Forests, Machine learning, Springer, pp. 5-32, 2001.

[21] Sakha Research Center, "The results of rice program for rice research and development," Laboratory for Agricultural Expert Systems, Ministry of Agriculture, Egypt, 2002.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. Witten,", The WEKA data mining software an update," in ACM SIGKDD Explorations Newsletter, 2009.

[23] L. Rokach and O. Maimon, Data mining with decision trees: theory and applications, World Scientific Publishing, 2^{nd} ed. 2015.

[24] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, Boosting, and Randomization," in Machine Learning, pp. 1-22, 1999.

[25] C. Aggarwal, Data mining: the textbook, Springer, 2015.

[26] S. Garca, J. Luengo and F. Herrera, Data preprocessing in data mining, Springer, 2014.

Bidirectional Extraction of Phrases for Expanding Queries in Academic Paper Retrieval

Yuzana Win
Graduate School of Engineering
Nagasaki University
Nagasaki, Japan

Tomonari Masada
Graduate School of Engineering
Nagasaki University
Nagasaki, Japan

***Abstract*—This paper proposes a new method for query expansion based on bidirectional extraction of phrases as word *n*-grams from research paper titles. The proposed method aims to extract information relevant to users' needs and interests and thus to provide a useful system for technical paper retrieval. The outcome of proposed method are the trigrams as phrases that can be used for query expansion. *First*, word trigrams are extracted from research paper titles. *Second*, a co-occurrence graph of the extracted trigrams is constructed. To construct the co-occurrence graph, the direction of edges is considered in two ways: *forward* and *reverse*. In the forward and reverse co-occurrence graphs, the trigrams point to other trigrams appearing after and before them in a paper title, respectively. *Third*, Jaccard similarity is computed between trigrams as the weight of the graph edge. *Fourth*, the weighted version of PageRank is applied. Consequently, the following two types of phrases can be obtained as the trigrams associated with the higher PageRank scores. The trigrams of the one type, which are obtained from the forward co-occurrence graph, can form a more specific query when users add a technical word or words before them. Those of the other type, obtained from the reverse co-occurrence graph, can form a more specific query when users add a technical word or words after them. The extraction of phrases is evaluated as additional features in the paper title classification task using SVM. The experimental results show that the classification accuracy is improved than the accuracy achieved when the standard TF-IDF text features are only used. Moreover, the trigrams extracted by the proposed method can be utilized to expand query words in research paper retrieval.**

***Keywords*—word n-grams; Jaccard similarity; PageRank; TF-IDF; query expansion; information retrieval; feature extraction**

I. Introduction

In these days, it is an important but complex task to get valuable information by searching the Web. With the rapid increase of information, users often perceive the difficulty of accessing the rich information resource effectively and of obtaining the information associated with their needs accurately. When users want to find the information relevant to their needs, they are required to find appropriate query words or phrases. However, the search results may not be relevant due to the inability of the queries to represent the needs accurately. Especially in academic paper retrieval, in many cases, users also want to find the papers focusing on specific and precise research topics, not general and vague topics. It can be considerably difficult for users to formulate a query for retrieving the papers discussing clear and specific topics. If the query contains only a single word, the search result consists of papers discussing a wide range of topics. That is, while the recall is high, the precision is low. If the query contains too many words, users may get only a limited number of academic papers as a search result. That is, while the precision is high, the recall is low. To overcome the above problems, the solution of this paper is to provide users with help in extracting from a large text set phrases that can be used to expand a less specific query. By expanding queries with the extracted phrases, users may get a search result containing a sufficient number of papers talking about specific research topics.

This paper proposes a new method for extracting important phrases as word *n*-grams from research paper titles. The extracted phrases are expected to be fruitful in query expansion for academic information retrieval. The proposed method is special in the following sense. The method extracts two types of phrases, each of which realizes a different query expansion, i.e., the expansion to the left and the expansion to the right. For example, the proposed method gives "a framework for" and "in sensor networks" as its outcome. The phrase "a framework for" can expand queries like "clustering", "classification", etc., to the left and give more specific queries like "a framework for clustering", "a framework for classification", etc. The phrase "in sensor networks" can expand queries like "clustering", "classification", etc., to the right and give more specific queries like "clustering in sensor networks", "classification in sensor networks", etc.

A brief explanation of the proposed method is given as follows. *First*, the proposed method extracts word trigrams as phrases that can be used for query expansion from a large number of research paper titles. There are two reasons why we focus on trigrams. The one reason is that, while word *n*-grams will be useful for text analysis, longer *n*-grams may cause data sparseness problem. Because the *n*-grams longer than three may be too long to obtain a sufficient large number of technical papers as a search result. The other is that unigrams and bigrams are too short to make a single word query express a specific and precise topic. *Second*, the proposed method builds a co-occurrence graph of the extracted trigrams. To construct the co-occurrence graph, the extracted word trigrams are used as nodes and the co-occurrence relations of trigrams appearing in the same paper titles as edges. Here, both the forward and reverse directions of edges are considered. In the forward co-occurrence graph, the trigram points to other trigrams appearing after it in a paper title. In the reverse co-occurrence graph, the trigram points to other trigrams appearing before

it in a paper title. *Third*, the proposed method evaluates the Jaccard similarity for all co-occurring pair of trigrams and utilizes the similarity as the edge weight. And *fourth*, the proposed method applies a weighted version of PageRank on the forward and reverse co-occurrence graphs. As a result, we can get the top-ranked trigrams with reference to PageRank scores. Many of the top-ranked trigrams given from these two co-occurrence graphs can be regarded as important phrases. Details will be explained later.

Our first paper [18] describes a method for exploring technical phrase frames by extracting word *n*-grams. However, this paper introduces a new approach that applies weighted PageRank algorithm on the forward and reverse co-occurrence graphs of trigrams. The distinction between these two types of co-occurrence graphs does not appear in [18]. As a result, the two types of top-ranked trigrams are obtained. The performance of the extracted trigrams are evaluated as additional features in paper title classification using SVM. This evaluation is also not included in [18].

The remainder of this paper is divided into four sections. Section 2 describes the related work. Section 3 explains the proposed method. Section 4 contains the results of the evaluation experiment. The final section concludes the paper with discussion on future work.

II. Related Work

The extraction of important word sequences, e.g. keyphrases and key sentences, is relevant to our problem. There are two types of extraction, i.e., supervised [2], [6], [7], [9] and unsupervised methods [1], [3], [4], [8], [10], [11]. Natural language processing techniques [12], [13], [14] have also been used for keyphrase extraction.

Mihalcea [15] proposed an unsupervised method for automatic sentence extraction using graph-based ranking algorithms. The author used a text graph to represent the interconnection of words or other text entities with meaningful relations, ranked the entire sentences in weighted graphs manner, sorted in reversed order of their scores and selected the top ranked sentences for summary. The author evaluated the method in text summarization task. The experimental results show that graph-based ranking algorithms (HITS and PageRank) are useful for sentence extraction when applied to graphs extracted from texts.

Litvak et al. [17] analyzed two graph-based approaches, i.e., unsupervised and supervised ones, which enhance to extract keywords to be used in summarizing documents. The researchers built a graph to represent the co-occurrence in a window of a fixed number of words. They used HITS algorithm to get the top-ranked keyword and identified the keywords in order to generate the summarization. As a result, they argued that if a large number of summarized documents were available then supervised classification was the most accurate to identify the keywords in a document graph. Unless the number of summarized documents are large, unsupervised classification is better to extract the keywords in a graph.

Wan et al. [16] proposed CollabRank, a collaborative approach to single-document keyphrase extraction from multiple documents. They implemented the CollabRank to obtain document clusters by using the clustering algorithm. They used the graph-based ranking algorithm to extract the keyphrases within each document cluster. They built a graph based on all candidate words in the documents of the given cluster and evaluated the candidate phrases in the document based on the scores of the words contained in the phrases. Finally, they chose a few phrases with highest scores as the keyphrases of the document.

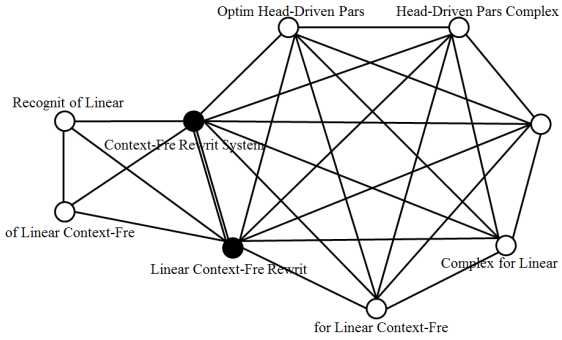


Fig. 1. A small portion of the co-occurrence graph

Contribution. This paper proposes a method that applies weighted PageRank algorithm on the forward and reverse co-occurrence graphs of trigrams. Consequently, the method can extract two different types of trigrams that can be used for query expansion: 1) Many of the trigrams obtained from the forward co-occurrence graph can form a more specific query when users add a word *before* them (e.g. "**clustering** for web search"); 2) Many of the trigrams obtained from the reverse co-occurrence graph can form a more specific query when users add a word *after* them (e.g. "automatic extraction of **clustering**"). This kind of bidirectional nature of extraction was not achieved by any of the PageRank-type methods described above.

III. The Proposed Method

In this section, the four steps of the proposed method are explained.

A. Word Trigrams

First, the proposed method extracts trigrams from a large set of research paper titles after applying stemming. For example, the proposed method extracts from the paper title "Recognition of Linear Context-Free Rewriting Systems" the following trigrams: "Recognit of Linear","of Linear Context-Fre", "Linear Context-Fre Rewrit", and "Context-Fre Rewrit System". Word trigrams are extracted by using the natural language toolkit for python (NLTK).

B. Co-occurrence Graph

The next step of the proposed method is to construct a co-occurrence graph of the extracted trigrams. In order to build the co-occurrence graph, the extracted word trigrams are used as nodes. When two trigrams appear in the same title, they are connected by an edge. Fig. 1 shows a small portion of the co-occurrence graph. This portion is obtained from the

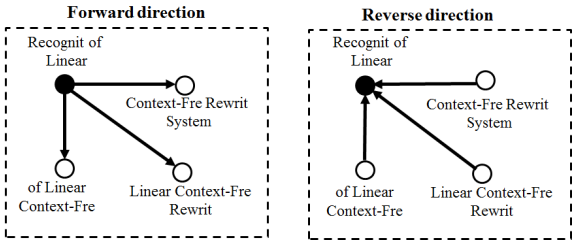


Fig. 2. Co-occurring pairs of trigrams according to forward and reverse directions

following two paper titles: "Recognition of Linear Context-Free Rewriting Systems" and "Optimal Head-Driven Parsing Complexity for Linear Context-Free Rewriting Systems".

Further, the direction of edges is specified according to the order of trigrams. The direction of edges is determined in two ways: *forward* and *reverse* directions, as shown in Fig. 2. On the left panel of Fig. 2, the trigram "Recognit of Linear" points to the trigrams "of Linear Context-Free", "Linear Context-Fre Rewrit", and "Context-Fre Rewrit System", because the latter three trigrams appear *after* the trigram "Recognit of Linear" in the paper title "Recognition of Linear Context-Free Rewriting Systems". This direction is called *forward* direction. In contrast, on the right panel of Fig. 2, the same trigram "Recongnit of Linear" is pointed by the other three trigrams. In this case, each trigram points to the trigrams appearing *before* it. This direction is called *reverse* direction. According to the forward and reverse directions of edges, the two co-occurrence graphs, i.e., forward co-occurrence graph and reverse co-occurrence graph, can be constructed.

C. Jaccard Similarity

In the third step, the Jaccard similarity is evaluated for all co-occurring pairs of trigrams and the similarity is utilized as the edge weight. Let (t_1,t_2) denote a pair of trigrams whose similarity is to be calculated. Let $S(t_i)$ denote the set of paper titles that contain the trigram t_i. The Jaccard similarity is computed between two trigrams t_1 and t_2 as follows:

$$sim(t_1, t_2) = \frac{|S(t_1) \cap S(t_2)|}{|S(t_1) \cup S(t_2)|} \quad (1)$$

After assigning the Jaccard similarity to each edge, a weighted version of PageRank algorithm is applied. The survey paper [5] analyzed many binary similarity measures. There are two reasons why we compute the Jaccard similarity. The first one is that it is simple to compute. The second one is that the Jaccard similarity is measured with the exclusion of *negative matches* [5]. In our approach, negative matches are related to the research paper titles where both of the trigrams under consideration do not appear and are not that important.

D. Weighted PageRank Algorithm

The last step of the proposed method applies weighted PageRank algorithm on both forward and reverse co-occurrence graphs of the extracted trigrams. Let P(t_i) denote the PageRank scores of the trigram t_i. Let w_{ji} denote the weight assigned to the edge connecting the two co-occurring pairs of nodes, t_i and t_j. w_{ji} is set to the corresponding Jaccard similarity. Then the PageRank score of the trigrams is calculated t_i by applying the Eq. (2) as below:

$$P(t_i) = \frac{1-d}{N} + d \times \sum_{t_j \in M(t_i)} \frac{w_{ji}}{\sum_{t_k \in M(t_j)} w_{jk}} P(t_j) \quad (2)$$

where $M(t_i)$ denotes the set of nodes which point to t_i and N is the total number of extracted trigrams. The parameter d is the damping factor that is usually set to 0.85. $\sum_{t_k \in M(t_j)} w_{jk}$ is the sum of the weights assigned to each neighbor t_k in $M(t_j)$. Intuitively, if a node is pointed by many high-scored neighbors, the node may get a high score. However, the proposed method combines the Jaccard similarity and weighted PageRank algorithm. Therefore, if a node is pointed by many high-scored neighbors with large Jaccard similarities, then the node may obtain a high score.

TABLE I. DATA SETS

Fields	*Venue*
NLP	ACL, EACL, COLING, CICLing, NAACL, IJCNLP, EMNLP, NLDB, TSD
DM	SIGMOD, VLDB, PODS, SIGIR, WWW, KDD, ICDE, ISWC, CIDR, ICDM, ICDT, EDBT, SDM, CIKM, ER, ICIS, SSTD, WebDB, SSDBM, CAiSE, ECIS, PAKDD
ALG	STOC, FOCS, ICALP, STACS, ISAAC, MFCS, FSTTCS, FCT, COCOON, CSR, WoLLIC
PRG	POPL, PLDI, ECOOP, OOPSLA, ISMM, ICLP, ICFP, CGO, ESOP, FOSSACS, CP, CC, LOPSTR, FLOPS, HOPL, AOSD

IV. EXPERIMENTAL RESULTS

A. Evaluation in Text Classification

The trigrams extracted by the proposed method were evaluated as additional features in the paper title classification task. We used SVM (Support Vector Machine) for classification and checked whether the trigrams extracted by the proposed method improved the classification accuracy.

The proposed method was tested in the binary classification of the paper titles obtained from DBLP (Digital Bibliography & Library Project) [1]. Each DBLP record included a list of authors, title, conference name or journal name, year, page numbers, etc. Academic conferences were chosen in the four fields: Natural Language Processing (NLP), Data Management (DM), Algorithms and Theory (ALG), and Programming Languages (PRG). We only selected top conferences and used the research paper titles presented in the conferences shown in Table I. As a result, the total number of paper titles contained in NLP, DM, ALG, PRG data sets are 10,666, 27,573, 16,468, and 9,434, respectively. In the preprocessing, stop words were removed and Porter Stemmer [2] was used to stem words to their root forms.

Classification was conducted on the four data sets, i,e., NLP paper titles, DM paper titles, ALG paper titles and PRG paper titles. From these four data sets, six different pair of data sets were obtained as ALG_PRG, DM_ALG, DM_PRG, NLP_ALG, NLP_DM and NLP_PRG. For each pair, the data was randomly split into 90% of the paper titles for training and 10% for testing, and the classification was performed

[1] http://www.dblp.com/
[2] http://www.tartarus.org/martin/PorterStemmer/

with SVM. The accuracies were averaged over the ten results obtained from the 10-fold cross-validation.

TF-IDF term weighting was used to compose a feature vector for each paper title based on the formula: $\text{tf\_idf}(t,d) = \text{tf}(t,d) \times \log(N/\text{df}(t))$, where $\text{tf}(t,d)$ is the frequency of term t in document d, and $\text{df}(t)$ is the document frequency of t, i.e., the number of documents where *t* appears. N is the total number of documents in the corpus. In the experiment, the TF in TF-IDF was modified by using the trigrams obtained by the proposed method to improve the classification accuracy. First, we find the trigram having the largest PageRank score in each paper title. For each paper title d, the set of the three words of the trigram is denoted having the largest PageRank score by $W(d)$. Then, $\text{weight}(t,d)$ is used, defined by Eqs. (3) and (4) in place of $\text{tf}(t,d)$:

$$\text{weight}(t,d) \equiv \alpha \times \text{tf}(t,d) + 1 \text{ for } t \in W(d) \quad (3)$$

$$\text{weight}(t,d) \equiv \alpha \times \text{tf}(t,d) \text{ for } t \notin W(d) \quad (4)$$

In Eqs. (3) and (4), α is the term reweighting parameter and is chosen as an integer. All word counts are increased by a factor of α and then the word counts are increased by one only for the words of the trigram having the largest PageRank score (cf. Eq. (3)). For example, we assume that the trigram "probabilist inform flow" has the largest PageRank score among the trigrams appearing in the paper title "Decidability of Parameterized Probabilistic Information Flow" after stemming. Then only the counts of the three words "probabilist", "inform", and "flow" are increased by one after we increase the counts of all words by a factor of α. If the trigrams extracted by the proposed method are important in the sense that they are closely related to a particular research topic and thus help discriminating the research topic from other topics, the reweighting described above may improve the classification accuracy.

SVM was trained with linear kernel by setting $C = 1.0$ and the classification accuracy was obtained in terms of Area Under the ROC curve (AUC). The term reweighting parameter α was varied from 3 to 27, and the mean and standard deviation of AUC in the 10-fold cross validation were recorded.

Tables II and III summarize the p-values obtained by comparing the standard TF-IDF (i.e., TF-IDF without modification of TF) and the TF-IDF based on the TF modified by Eqs. (3) and (4) in terms of AUC. The p-values are obtained in a paired two-sided t-test. If the classification accuracy of the proposed method is not as high as the frequency-based method, the p-value is assigned with a minus symbol. When the p-value is less than 0.05, we can say that the improvement is statistically significant and thus give the p-value in bold in Tables II and III. The results of Table II are given by using the trigrams obtained from the forward co-occurrence graph. On the other hand, the results of Table III are given by using trigrams obtained from the reverse co-occurrence graph. Tables II and III show the term reweighting factor α yielding the best p-values on each data set. Only for the two pairs, i.e., DM_PRG and NLP_PRG, in Table III, we could not get a statistically significant improvement. For all remaining cases in Tables II and III, we could get a significantly better accuracy than the standard TF-IDF. Based on these results, it can be said that the classification accuracy is improved by modifying the TF in TF-IDF by the trigrams the proposed method gives. So we claim that the proposed method can extract the features that are useful in discriminating different research topics as the trigrams having large PageRank scores.

TABLE II. p-VALUES FOR PAIRED T-TEST ON ROC CURVE AUC (FORWARD)

Data sets	*Standard TF-IDF*	*Modified TF-IDF*	*p-value*
ALG_PRG	0.942075	0.942493 (α = 3)	**0.009**
DM_ALG	0.978106	0.978225 (α = 8)	**0.021**
DM_PRG	0.971507	0.971669 (α = 10)	**0.029**
NLP_ALG	0.989345	0.989452 (α = 8)	**0.003**
NLP_DM	0.954356	0.954432 (α = 27)	**0.048**
NLP_PRG	0.985577	0.985633 (α = 19)	**0.047**

TABLE III. p-VALUES FOR PAIRED T-TEST ON ROC CURVE AUC (REVERSE)

Data sets	*Standard TF-IDF*	*Modified TF-IDF*	*p-value*
ALG_PRG	0.942075	0.943616 (α = 9)	**0.013**
DM_ALG	0.978106	0.978163 (α = 8)	**0.020**
DM_PRG	0.971507	0.971566 (α = 23)	0.052
NLP_ALG	0.989345	0.989422 (α = 12)	**0.016**
NLP_DM	0.954990	0.954883 (α = 20)	−0.053
NLP_PRG	0.985899	0.985959 (α = 22)	**0.041**

B. Comparing with Frequency-based Trigram Extraction

To discuss the special nature of the trigrams extracted by the proposed method, we compared the proposed method with a simple method for the extraction of trigrams, i.e., the frequency-based extraction. In the frequency-based method, the same data sets were used and the same preprocessing were applied as in the proposed method. Then, the number of occurrences, i.e., frequency, were counted for every trigram, and the higher-ranked trigrams based on their frequencies were obtained. The difference between two methods are clarified by displaying examples.

Tables IV and V summarize the trigrams obtained by the frequency-based method and by the proposed method for ALG and DM data sets, respectively. For example, "the complex of" and 381 in the top cell of the left column of Table IV mean that the frequency is 381 for the trigram "the complex of". Moreover, "and relat problem" and 6.05 $\times 10^{-4}$ in the top cell of the middle column of Table IV mean that the PageRank is 6.05 $\times 10^{-4}$ for the trigram "and relat problem" in the forward co-occurrence graph. Furthermore, "on the complex" and 19.59 $\times 10^{-4}$ in the top cell of the right column of Table IV mean that the PageRank is 19.59 $\times 10^{-4}$ for the trigram "on the complex" in the reverse co-occurrence graph.

We can observe that many trigrams obtained from the forward co-occurrence graph can expand queries to the right. For example, the trigram "in web search" can expand the queries like "ranking" and "queries" to give more specific queries like "ranking in web search" and "queries in web search". On the other hand, many trigrams obtained from the reverse co-occurrence graph can expand queries to the left. For example, the trigram "efficient algorithm for" can expand the queries like "computing" and "mining" to give more specific queries like "efficient algorithm for computing" and "efficient algorithm for mining". This is a remarkable feature of the proposed method. In contrast, the frequency-based method

TABLE IV. TOP-10 (STEMMED) TRIGRAMS OF ALG

Frequency		*PageRank*			
		Forward ($\times 10^{-4}$)		*Reverse* ($\times 10^{-4}$)	
the complex of	381	and relat problem	6.05	on the complex	19.59
lower bound for	259	in polynomi time	5.52	the complex of	19.23
approxim algorithm for	225	in linear time	5.46	lower bound on	8.89
algorithm for the	209	and it applic	4.10	approxim algorithm for	8.69
on the complex	162	term rewrit system	3.75	lower bound for	8.12
the power of	103	in the plane	3.72	a note on	7.20
with applic to	100	constraint satisfact problem	3.24	bound on the	7.09
bound on the	91	in planar graph	3.05	effici algorithm for	6.84
lower bound on	81	of complex class	2.37	the power of	6.13
effici algorithm for	73	and their applic	2.34	on the power	5.76

TABLE V. TOP-10 (STEMMED) TRIGRAMS OF DM

Frequency		*PageRank*			
		Forward ($\times 10^{-4}$)		*Reverse* ($\times 10^{-4}$)	
a case studi	229	on the web	4.71	the impact of	8.06
the role of	219	for inform retriev	2.91	the effect of	6.56
the impact of	216	in inform retriev	2.68	the role of	4.41
a framework for	201	a case studi	2.62	a framework for	3.39
the effect of	184	in web search	2.50	a comparison of	3.04
for inform retriev	140	in social network	2.33	the influenc of	2.83
the case of	140	an empir studi	2.03	a studi of	2.71
in inform system	137	inform retriev system	2.00	the use of	2.06
on the web	134	an exploratori studi	1.94	effici process of	1.85
of inform system	123	in social media	1.90	an analysi of	1.84

TABLE VI. p-VALUES FOR PAIRED T-TEST ON ROC CURVE AUC (FORWARD)

Data sets	*Frequency-based*	*Proposed*	*p-value*
ALG_PRG	0.94904 (α = 5)	0.94928 (α = 3)	0.058
DM_ALG	0.98194 (α = 4)	0.98202 (α = 2)	0.546
DM_PRG	0.97613 (α = 2)	0.97622 (α = 2)	0.333
NLP_ALG	0.99140 (α = 5)	0.99148 (α = 4)	**0.031**
NLP_DM	0.96251 (α = 2)	0.96276 (α = 2)	0.079
NLP_PRG	0.98699 (α = 5)	0.98706 (α = 4)	0.336

TABLE VII. p-VALUES FOR PAIRED T-TEST ON ROC CURVE AUC (REVERSE)

Data sets	*Frequency-based*	*Proposed*	*p-value*
ALG_PRG	0.94245 (α = 4)	0.94256 (α = 3)	0.689
DM_ALG	0.97828 (α = 5)	0.97822 (α = 5)	−0.384
DM_PRG	0.97235 (α = 3)	0.97214 (α = 3)	−0.195
NLP_ALG	0.98942 (α = 5)	0.98941 (α = 5)	−0.774
NLP_DM	0.95450 (α = 6)	0.95422 (α = 6)	**−0.024**
NLP_PRG	0.98563 (α = 3)	0.98548 (α = 5)	−0.054

cannot give these two types of trigrams separately, because all trigrams are mixed in the same ranking, as shown in the left columns of Tables IV and V.

Further, we can observe that the frequency-based ranking tends to provide trigrams having a general meaning like “lower bounds for”, “the power of”, “with applications to”, “bounds on the”, “lower bounds on”, “a case study”, “the case of”, etc., where the original form is recovered from the root form of each word. In contrast, the proposed method tends to provide trigrams having a specific meaning, e.g. like “in polynomial time ”, “in linear time”, “in planar graphs”, “term rewriting systems”, “constraint satisfaction problems”, “of complexity classes”, “information retrieval system”, “an empirical study”, etc., with respect to the forward co-occurrence graph. Also with respect to the reverse co-occurrence graph, many trigrams given by the proposed method have at least as specific a meaning as the trigrams given by the frequency-based method. Therefore, it can be said that, at least with respect to the forward co-occurrence graph, the top-ranked trigrams obtained by the proposed method have a more specific meaning than those obtained by the frequency-based method.

However, it is possible that the proposed method may degrade the quality of the extracted trigrams by providing them in two separate rankings. Therefore, we compared the proposed method with the frequency-based method also in text classification task described in Section IV-A. We also used SVM (Support Vector Machine) for classification and checked if the trigrams extracted by the proposed method were as useful as the trigrams extracted by the frequency-based method.

To obtain the best classification accuracy in terms of Area Under the ROC curve (AUC), SVM was trained with two different kernels, namely linear kernel by setting $C = 1.0$ and rbf (Radial Basis Function) kernel by setting $C = 2.0$ and $gamma = 2.0$. We selected the term reweighting parameter α yielding the best case from each kernel and recorded the mean and standard deviation of AUC in the 10-fold cross validation.

Tables VI and VII summarize the p-values obtained by comparing the frequency-based method and the proposed method based on the TF modified by Eqs. (3) and (4) in

terms of AUC. The p-values are obtained in a paired two-sided t-test. The p-value is assigned with a minus symbol if the classification accuracy of the proposed method is not as high as the frequency-based method. When the p-value is less than 0.05, it can be said that the improvement is statistically significant. The results of Table VI are given by using the trigrams obtained from the forward co-occurrence graph, where SVM is trained by using the rbf kernel. On the other hand, the results of Table VII are given by using the trigrams obtained from the reverse co-occurrence graph, where SVM is trained by using the linear kernel. For all but one case in Tables VI and VII, we could get as good an accuracy as the frequency-based method. We could not get a comparable accuracy only for the NLP_DM data set pair in Table VII. Consequently, the result showed that the proposed method at least could extract as effective trigrams as the frequency-based method. It can be said that the bidirectional nature of the proposed method is an extra gain, which cannot be achieved by the frequency-based method.

C. A Possible Application: Query Expansion

Based on the experimental results, it can be said that the trigrams extracted by the proposed method represent technical research topics well. We here discuss how such trigrams can be used in query expansion for information retrieval.

For example, as presented in Fig. 3, the query word "clustering" can be expanded to the right by the trigrams "in sensor networks", "for web search", "for text categorization", etc., which are obtained by the proposed method from the forward co-occurrence graph. These trigrams can be used for the *right* expansion in this manner, because their first word (i.e., "for","in","of", etc.) is a function word that mainly follows a noun. As we discussed in Section IV-B, the trigrams obtained by the proposed method tend to represent a specific meaning, especially with respect to the forward co-occurrence graph. Therefore, we may expect that the search results obtained by the queries expanded in this manner will relate to specific research topics. Fig. 4 gives another example. The query word "clustering" is expanded to the left by the trigrams "a framework for", "automatic extraction of", "efficient algorithm for", etc., which are obtained from the reverse co-occurrence graph. These trigrams can be used for the *left* expansion, because their last word (i.e., "for", "of", etc.) is a function word that is mainly followed by a noun.

It should be noted that a similar expansion cannot be straightforwardly achieved by the trigrams obtained by the frequency-based method, because the trigrams that can be used for the right expansion and those that can be used for the left expansion are mixed in the same ranking as shown in the left columns of Tables IV and V. However, the proposed method provides two types of trigrams in two different rankings, as shown in the middle and right columns of Tables IV and V.

We here verify how the search results obtained by the expanded queries can focus on more specific research topics. Fig. 5 shows the three types of search results obtained from Google Scholar. Fig. 5(a) gives the search results for the query "clustering". Fig. 5(b) gives the search results obtained by the proposed method from the forward co-occurrence graph. Fig. 5(c) gives the search results obtained by the proposed

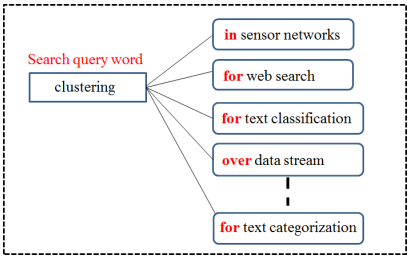


Fig. 3. Example of possible right expansions of the query word 'clustering' by using the trigrams obtained from the forward co-occurrence graph for DM data set

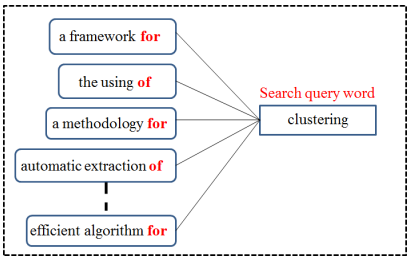


Fig. 4. Example of possible left expansions of the query word 'clustering' by using the trigrams obtained from the reverse co-occurrence graph for DM data set

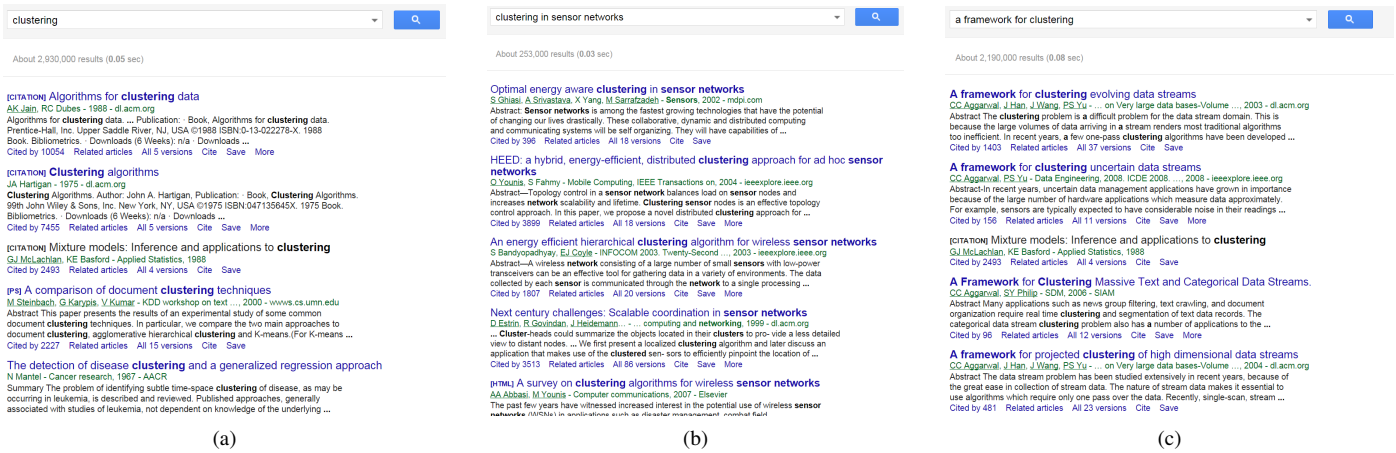
method from the reverse co-occurrence graph. As presented in Fig. 5(a), we can get the search result having a general meaning when we only input a single query word "clustering". For example, the topics like "Algorithms for clustering data" and "A comparison of document clustering techniques" tend to provide a general meaning consisting of the words like "algorithms" and "techniques". These words tend to represent a wide range of topics. Consequently, the single query word has not exploited users' needs and interests, but users can't get relevant topics when each user has a specific need.

In contrast, when a single query word "clustering" is expanded to the right by the phrase "in sensor networks", we can get the search results focusing on more specific topics as shown in Fig. 5(b). Most of the words or phrases appearing in the search results, e.g., "hybrid", "ad hoc", "hierarchical", and "wireless", have a specific meaning. On the other hand, when a single query word "clustering" is expanded to the left by the phrase "a framework for", we can also get the search results focusing on more specific topics as shown in Fig. 5(c). Some of the words or phrases occurring in the search results, e.g., "data streams", "high dimensional", and "Text and Categorical", represent a specific meaning.

Therefore, it is found that the query expansion, i.e., the expansion to the left and the expansion to the right, can give the search results relating to specific topics. Further, we can get two different types of search results due to the bidirectional nature of the proposed method. These results are more specific when we expand query words than when we only use a single query word. We can observe that the proposed method works as a new query expansion scheme more oriented toward actual user needs and interests for informational retrieval.

V. Conclusion

In this paper, we proposed a new method for query expansion based on bidirectional extraction of phrases. The proposed



(a) (b) (c)

Fig. 5. Example of the search results for the query (a) 'clustering' (b) 'clustering in sensor networks' and (c) 'a framework for clustering'

method extracted important phrases as trigrams based on a procedure consisting of four processing steps. The trigrams extracted by the proposed method were evaluated as additional features in the paper title classification task using SVM. The experimental results showed that the accuracy was improved. We also compared the trigrams given by the proposed method with those given by the frequency-based method. According to the experimental results, the proposed method could provide as good trigrams as the frequency-based method. However, the proposed method has an extra gain, i.e., the bidirectional nature of trigrams extraction, which cannot be achieved by the frequency-based method. Further, we discussed how we could use such trigrams for query expansion. A search system using this type of query expansion can give search results relating to specific topics.

We have a future plan to perform a quantitative evaluation of the search results obtained by the query expansion based on the proposed method in information retrieval task.

ACKNOWLEDGMENT

This work has been supported by the Grant-in-Aid for enhancement of engineering higher education of the Japan International Cooperation Agency (JICA) from 2014 to 2017. We are grateful for their support.

REFERENCES

[1] K.S. Hasan and V. Ng, "Conundrums in unsupervised keyphrase extraction: making sense of the start-of-the-art," in *Proc. of the 23rd International Conference on COLING 2010*, Beijing, pp. 365–373, August 2010.

[2] C. Caragea, F. Bulgarov, A. Godea, S.D. Gollapalli, "Citation-enhanced keyphrase extraction from research papers: a supervised approach," in *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1435-1446, Doha, Qatar, October 2014.

[3] R. Mihalcea, P. Tarau and E. Figa, "PageRank on semantic networks with application to word sense disambiguation," in *Proc. of the 20th International Conference on Computational Linguistics*, Stroudsburg, PA, USA, no. 1126, August 2004.

[4] S.D. Gollapalli and C. Caragea, "Extracting keyphrases from research papers using citation networks," in *Proc. of the 28th AAAI Conference on Artificial Intelligence*, pp. 1629–1635, June 2014.

[5] S. Choi, S. Cha, and C.C. Tappert, "A survery of binary similarity and distance measures," *J.Syst. Cybern, Inf.*, vol. 8, no. 1, pp. 43–48, 2010.

[6] D. X. Wang, X. Gao, and P. Andreae, "Automatic keyword extraction from single-sentence natural language queries," in *Proc. of the 12th Pacific Rim International Conference on Artificial Intelligence*, Kuching, Malaysia, pp. 637–648, September 2012.

[7] K. Zhang, H. Xu, J. Tang, and J. Li, "Keyword extraction using support vector machine," in *Proc. of the 7th International Conference on WAIM*, Hong Kong, China, pp. 85–96, June 2006.

[8] T. Nomoto and Y. Matsumoto, "A new approach to unsupervised text summarization," in *Proc. of SIGIR2001*, pp. 26–34, 2001.

[9] M. R. Amini and P. Gallinari, "The use of unlabeled data to improve supervised learning for text summarization," in *Proc. of SIGIR2002*, 105-112, 2002.

[10] S. N. Kim and M. Kan, "Re-examining automatic keyphrase extraction approaches in scientific articles," in *Proc. of 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009*, pp. 9-16, Suntec, Singapore, August 2009.

[11] Z. Liu, P. Li, Y. Zheng and M. Sun, "Clustering to find exemplar terms for keyphrase extraction," in *Proc. of 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 257-266, Singapore, August 2009.

[12] T. Tomokiyo and M. Hurst, "A language model approach to keyphrase extraction," in *Proc. of ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, vol. 18, pp. 33–40, 2003.

[13] K. Barker and N. Cornacchia, "Using noun phrase heads to extract document keyphrases," in *Proc. of 13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence*, pp. 40–52, May 2000.

[14] S. N. Kim, T. Baldwin and M. Kan, "Evaluating n-gram based evaluation metrics for automatic keyphrase extraction," in *Proc. of 23rd international conference on COLING 2010*, pp. 572-580, Beijing, August 2010.

[15] R. Mihalcea, "Graph-based ranking algorithms for sentence extraction, applied to text Summarization," in *Proc. of ACL 2004 on Interactive Poster and Demonstration Sessions*, Article no. 20, Stroudsburg, PA, USA, July 2004.

[16] X. Wan and J. Xiao, "CollabRank: towards a collaborative approach to single-document keyphrase extraction," in *Proc. of 22nd International Conference on COLING 2008*, pp. 969–976, Manchester, August 2008.

[17] M. Litvak and M. Last, "Graph-based keyword extraction for single-document summarization," in *Proc. of 08 MMIES on COLING 2008*, pp. 17-24, Manchester, August 2008.

[18] Y. Win and T. Masada, "Exploring technical phrase frames from research paper titles," in *Proc. of 29th IEEE International Conference on WAINA-2015*, pp. 558–563, Gwangju, South Korea, 2015.