TOPIC ANALYSIS OF SCHOLAR PAPER USING EARTH OBSERVATION VOCABULARY

[a] Masafumi ONO, [b] Masahiko NAGAI, [c] Ryosuke SHIBASAKI

[a] Earth Observation Data Integration and Fusion Research Initiative, the University of Tokyo,
Cw503 Institute of Industrial Science 4-6-1 Komaba, Meguro-ku,Tokyo 153-8505, Japan;
Tel: +81-3-5452-6417;
Email: maono@iis.u-tokyo.ac.jp

[b] Geoinformatics Center (GIC), Asian Institute of Technology (AIT) ,
km 42, Paholyothin Highway, P.O. Box 4, Klong Luang, Pathumthani 12120, THAILAND;
Tel: + 66-2-524-5599;
E-mail: nagaim@ait.ac.th

[c] Center for Spatial Information Science, the University of Tokyo,
Cw503 Institute of Industrial Science 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan;
Tel: + 81-3-5452-6417;
E-mail: shiba@csis.u-tokyo.ac.jp

KEY WORDS: Earth observation, Vocabulary, Paper analysis, Naïve Bayes, Text mining

ABSTRACT

This paper presents a pilot analysis for Asian Conference on Remote Sensing (ACRS) proceedings by using Earth Observation Vocabulary (EOV) which consists of 146 observation items and its definitions. Although the papers in ACRS proceedings are classified into official session topics, the topics change every year and also have wide varieties including academic theme or project-based categories. So only with the official classification, it is difficult to analyze over years. Therefore, in this paper we attempt to characterize the ACRS proceedings from the viewpoint of observation items in EOV. For the analysis, we use Naïve Bayes which is a popular text classification model. In this paper, firstly we evaluate the accuracy of the Naïve Bayes in order to find the suitable conditions. Subsequently, we apply the Naïve Bayes model for proceedings and calculate the similarity to EOV. The result shows the priority of observation items and the trend in ACRS.

INTRODUCTION

The volume of earth observation data is increasing with the current improvement of global observation technologies. In parallel with this, released papers on earth observation research are also increasing. However, the detail "what kinds of earth observation research are well-published?" is not obvious. This is a motivation that we operate a pilot analysis with earth observation vocabulary (EOV).

Every year many researches using observation data in Asia are introduced through Asian Conference on Remote Sensing (ACRS). As the output, many papers are published in ACRS proceedings. The ACRS papers are officially classified into the representative session topics that reflect the contents of each ACRS paper. However, the session topics have several varieties including not only academic theme or observation items but also project-based categories such as "JAXA SAFE" in 2011 special sessions. Of course, we respect this official classification. But, because the session topics changes every year, it is difficult to analyze over years. Therefore, we consider that it is meaningful to extract characteristic from the viewpoint of observation items.

EOV is developed by University of Tokyo and Group on Earth Observation (GEO) members. It consists of 146 observation parameters and its definition. The EOV is originated from a list of prioritized earth observation items suggested in the report of "GEO task US-09-01a" (GEO UIC 2010). Currently, the EOV is used for accessing GEO resources or classifying observation dataset in Data Integration and Analysis System (DIAS; ONO 2010). Because observation data basically can be classified into observation parameters, we think EOV is compatibility with remote sensing researches.

In analysis, we use Naïve Bayes as classification method and similarity adaptation to EOV. Naïve Bayes is a very popular probabilistic learning model in the field of machine learning. The characteristic of Naïve Bayes is easy to apply and fast to perform. It is generally said that Support Vector Machine has better accuracy than Naïve Bayes in

binary classification. However, Naïve Bayse is more suitable for this research because we will attempt to calcurate multi class simiralities to EOV which has many classes over 100.

In this paper, firstly we explain the detail of Naïve Bayes methods. And next, we evaluate the accuracy of the Naïve Bayes in order to find the suitable conditions. Subsequently, we analyze for ACRS proceedings through EOV. Then finally, we will conclude with remarks.

METHODS

Naïve Bayes

Naïve Bayse treats all documents as frequency information of words. In this model, words are assumed to be independent each other. For leaning process, Naïve Bayes learner creates the occurrence table of word binded to every class $c_j \in C$ from tokenized documents. This is used as the training set T for classification. In the classification process, classifier takes the word occurrence from T and counts up for each word $w_i \in W$ in a target document d. In the principle of Multinominal Naïve Bayes (MNB), which is a basic model of Naïve Bayes, finally we can get the estimating formula is as

$$b_{MNB}(d) = \ argmax_c \left[\log p(c_j) + \sum_i f_i \log \frac{N_{ci} + \alpha}{N_c + \alpha|V|} \right].$$

$p(c_j)$ is a prior class probability to estimate the documents belonging to class c_j for all documents. f_i indicates the frequency of word w_i in target document d. N_{ci} is the number of word w_i assigned to class c_j in T. N_c is the total number of all word occurrences assigned to class c_j in T. |V| is the total number of vocabulaly in training set T, except overlapped words. α is a smoothing parameter. In this research, we set $\alpha = 1$.

If it finds the most suitable class for the target document d, the single class with the maximum score will be selected. However, this research uses Naïve Bayes not only for class identification but also for similarity adaptation (Aiguzhinov 2010). So, we caluculate the similarity to each class by using

$$ms_{cj}(d) = \ \log p(c_j) + \sum_i f_i \log \frac{N_{ci} + \alpha}{N_c + \alpha|V|}.$$

Acorrding to abobe formulas, we can understand that MNB deals with paramers only related to a single class c_j when comparing. In contrast, Compliment Naïve Bayes (CNB), which is another model of Naïve Bayes, handles many parameters gotten via all class except c_j (Jason 2003). The rule of CNB is

$$b_{CNB}(d) = \ argmax_c \left[\log p(c_j) - \sum_i f_i \log \frac{N_{\bar{c}i} + \alpha}{N_{\bar{c}} + \alpha|V|} \right].$$

$N_{\bar{c}i}$ is the number of word w_i in all classes except class c_j in the training set T. $N_{\bar{c}}$ is the total number of all words in classes except c_j in T. It is generally expected that CNB works more stable for bias in T than MNB, because CNB handles parameters from more classes. In classification, CNB classifier assigned the target document to the class with the maximum score. As a simirality evaluation of CNB, this reserach also uses

$$cs_{cj}(d) = \ \log p(c_j) - \sum_i f_i \log \frac{N_{\bar{c}i} + \alpha}{N_{\bar{c}} + \alpha|V|}.$$

EXPERIMENT FOR ACCURACY ON NAÏVE BAYES

Outline

Generally, there are many noises in word frequency information initially leanred from documents. In some cases, the training set through data cleaning has a possibility to improve Naïve Bayes classifier. Accordingly, in this section we compare the accuracies of MNB and CNB per several boundary conditions and subsequently we will find suitable parameters for this research.

In order to evaluate the accuracy on Naïve Bayes, we use ACRS2010 and ACRS 2011 oral papers classified into official topic in the program. Table 1 shows the numbers of topic and oral paper per each program. Because overlaps are removed, for example session topic "TS2-1 Data Processing-1" and "TS2-5 Data Processing-2" in ACRS 2011 are couted as a common topic. As an except case, the topic "TS7-1 New Development of GIS" in ACRS 2011 is integrated to "TS6-1 GIS and Environment", because we think that both topic are similar and that it is difficult to classify whether "new development" or not on text clasification methods. Further, although the real number of all oral papers in ACRS2010 is 261, we removed several blank papers and set 244 as the usable number.

Table 1: Number of unique topics and usable papers in ACRS 2010 and 2011

	ACRS 2010 program	*ACRS 2011 program*
Session Topic	31	31
Oral paper	227	244

The topics of ACRS2010 and ACRS2011 are totally different. But it can be considered that both has several common topics. For example, "TS09 : Urban Change / Monitoring" in ACRS2010 program and "TS1-4 Urban Change Monitoring" in ACRS2011 are same. So in this experiment, we take papers in the common topics as a test set and then we evaluate the accuracy for classification and similarity through Naïve Bayes moel. Table 3 shows the expected document maping from ACRS2011 topic to ACRS2010.

Table 2: Expected mapping from ACRS 2010 topic to ACRS 2011

ACRS 2011 topic	*Counts*		*ACRS 2010 topic*
TS1-4: Urban Change Monitering	6	→	TS07, TS09: Urban Change / Monitering
TS1-6: Oceanography	5	→	TS17, TS29 : Atmosphere / Oceanography
TS2-1, TS2-5, TS2-6: Data Processing	21	→	TS16, TS33 : Data Processing
TS2-3, TS2-4, TS4-5: Land Cover and Land Use	15	→	TS20, TS31, TS39: Land Use / Land Cover
TS2-7, TS4-9, TS4-10: Algorithm	16	→	TS28, TS38: Algorithm and Modeling
TS4-4, TS4-5 : Forest Resources	10	→	TS02, TS08 : Forest Resources

Example of Training set

Figure 1 shows an example of the training set created from all oral papers in ACRS 2010. Each word token has the information of the occurrence and the part-of-speech (POS). In the example, it is assigned to the class such as "Land Use / Land Cover" which is one of settion topic in ACRS 2010. "28347" is the total number of words assigned in the clasas.

```
Land Use / Land Cover: 28347
{'area': (353, 'NN'), 'land': (337, 'NN'), 'forest': (220, 'NN'), 'cover': (215, 'RB'), 'classification': (210, 'NN'),
 'data': (207, 'NNS'), 'image': (197, 'NN'), 'study': (145, 'NN'),…
 …
…'decision-making': (1, 'JJ'), 'seasonal': (1, 'JJ'), 'technological': (1, 'JJ'), 'cultural': (1, 'JJ'), 'questionable': (1, 'JJ')}
```

Figure 1: an example of the training set created from all oral papers in ACRS 2010

Metric

When evaluating the accuracy, we introduce two metrics "Precision for Class identification (PC)" and "Similarity Coverage (SC)". The metric PC evaluates whether the test documents set $d_k \in D$ could be totally classified into the best single classes. PC is estimated with the following formula $pc(D)$. If d_k is classified into the correct class through $b_{MNB}(d)$ or $b_{CNB}(d)$, it takes $g(d_k)$ =1. If not classified, it takes $g(d_k)$ =0.

$$pc(D) = \frac{1}{n}\sum_{k=1}^{n} g(d_k)$$

The metric SC evaluates totally how similar each document is to the correct class. SC is estimated with the bottom formula $sc(D)$. In it, Z_c means the number of all classes in the training set. $rank(d_k)$ is the rank of a document d_k to the correct class in similarity ranking .

$$sc(D) = \frac{1}{n}\sum_{k=1}^{n} \frac{Z_c - \{rank(d_k) - 1\}}{Z_c}$$

Evaluation for thresholds

At first, we comare the accuraccy of MNB and CNB per different threshold for occurrence of word. This is because it's expected that some vocaburalies and noises with less contribution for characterizing documents might be removed through the limitation. In this experiment, the vocaburalies with the occurrence greater than or equal to the determined threshold are valid on calculation. The others with less occurrence are ignored.

Figure 2 shows the scores of *pc(D)* and *sc*(*D*) with thresholds from 1 to 5. The left of the figure is the result of *pc*(*D*) and the right is the result of *sc*(*D*). The red bar is the score of MNB, while the blue bar is the score of CNB. Larger score means better accuracy in both figure. When we tried prior experiments with a threshold over 5, vocabularies in some papers almost disappeared so we decided the maximum threshold is 5.

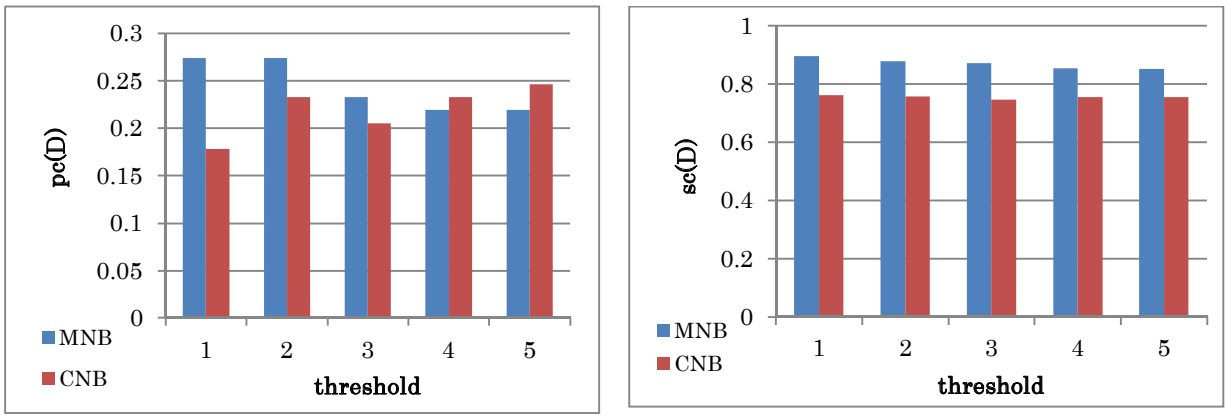

Figure 2: *pc*(*D*) and *sc*(*D*) per different thresholds

In *pc*(*D*) evaluation, the scores on MNB are decreasing with larger thresholds. In contrast, the scores on CNB are almost increasing except for the thresholds 2 . Threfore, MNB is better than CNB in smaller thresholds and then this tendency is reversed at the threshold 4. As the total scope, the minimum is 0.1780 at threshold 1 on CNB. The maximum is 0.2739 at threshold 1 on MNB.

In *sc*(*D*) estimation, the score on MNB is always better than CNB. However, the scores on MNB are decreasing, while scores on CNB are almost stable. Totall, the minimum is 0.7453 at threshold 3 on CNB and the maximum is 0.8955 at threshold 1 on MNB.

Evaluation for parts of speech

There are many numeric values in achademic papers to show evidences on the research. Because numeric value is neutral information, they do not basically contribute to improving topic classificationm. Also, conjunctions such as 'and', '&' and 'or' do not characterize each document. Thefore, if a specific part-of-speech (POS) is removed or suitably selected, it might be expected for improving accuracy. Thus, we comared the accuracy per different POS. These results are shown in Figure 3.

In Figure 3, the horizontal axis shows the results with training sets having different POS. The left side of the axis indicates the result with more POS. The end of the right side indicates the result by a single POS only. More specifically, "Plain" means that all POSs are contained on calculation. "CD, CC, MD removed" means that cardinal number (CD), coordinating conjunction (CC) and modal(MD) expected to have no critical effects on classification are removed. "N, V, J, R" means that noun, verb, adjective and adverb which are basic parts to make sentence are calculated. "N, V, J" means that noun, verb, and adjective are contained. "N and V" means that noun and verb which is the minimum parts for sentence are pickuped. "N only" means that only noun to characterize documents most are only used for calculation.

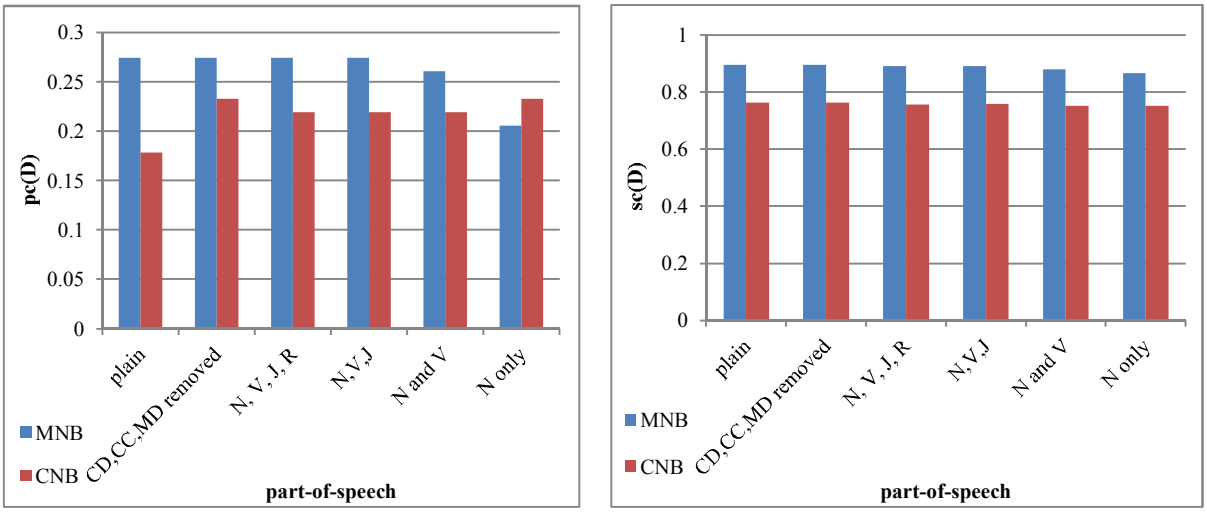

Figure 3: *pc*(*D*) and *sc*(*D*) per different POS

In *pc*(*D*) evaluation, the scores on MNB are equal between "plain" and "N, V, J" and the score at the "N only" is remarkably low. On the other hands, the scores on CNB are equal between "N, V, J, R" and "N, V, J". The scores at the "CD, CC, MD removed" and "N only" are highest in CNB. The minimum is 0.1780 at "plain" on CNB. The maximum is 0.2739 between "plain" and "N, V, J" on MNB.

In *sc*(*D*) evaluation, the score on MNB is always better than CNB. The scores on MNB and CNB are almost equal but they tend to be slightly decreasing. The minimum is 0.7513 at "N and V" on CNB. The maximum is 0.8955 at the "plain" on MNB.

Totally speaking, it seemed that these tendencies of increasing and decreasing on POS are similar with the result of threshods.

Evaluation on different resources

In above two experiments, we used the same resouce of ACRS 2010 papers as a training set. At this third experiment, we attempt to compare the accuracies between ACRS 2010 and Wikipedia-en which is another different resouce. The training set of Wikipedia-en is created from the aggregation of pages which are related to the topic of ACRS 2010. The parameters in both training sets are same as threshols = 1 and POS = "plain". These results are shown in Figure 4.

In pc(D) evaluation, the class identification on MNB with Wikipedia-en are all failed. The scores on CNB are about one-third of ACRS 2010. As the total scope, the minimum is 0.0 with the training set of Wikipedia-en on MNB. The maximum is 0.2739 with ACRS 2010 on MNB.
In sc(D) evaluation, although the score of ACRS2010 on MNB is better than CNB, the scores of Wikipedia-en on MNB is worse than CNB. As the total scope, the minimum is 0.5424 at Wikipedia-en on MNB and the maximum is 0.8955 at the ACRS2010 on MNB.

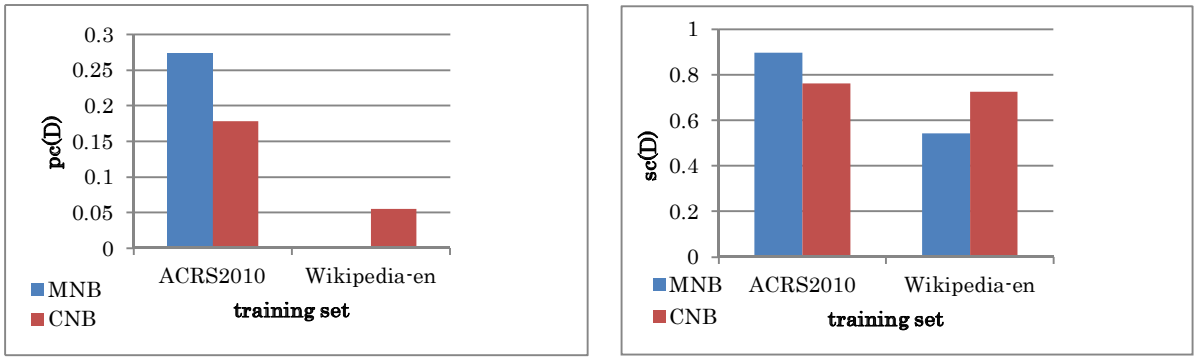

Figure 4: *pc*(*D*) and *sc*(*D*) per different resources

Discussion

Throuth these experiments, we found the MNB maks the better accuracy with rich vocaburalies in the low thresholds and the full coverd POS,while CNB maks the better accuracy with well-cleaned or trimmed vocaburaies in class identification. Also, it's expected that the best result is gotten by MNB and with the training set without data cleaning specially. But, the experiment comparing with Wikipedia-en indicates that CNB might have a better accuracy if the quality of the original resouce for learning is not sofisticated. In above many cases, although the score of MNB is better than CNB, CNB is more stable because CNB basically handles information from all classes in calculation.

In addition, we found that it is unfortunately difficult to apply Naïve Bayes as a class identification method for practical uses, because the maximum *pc*(*D*) scores is under 0.30. We think that some ACRS topics such as special sessions are certainly not easy to be classified by word frequency information.

On the other hands, we understand that Naïve Bayes is useful for the total similarity adaptation as a statistical analysis, because the scores of *sc*(*D*) are good through all experiments. For example, the score 0.8955 of *sc*(*D*) means that the the average similarity rank for the correct topic of ACRS paper marks at least within rank 4 in 31 topics . This score can be satisfied for practical uses. Thefore, we will use Naïve Bayes as similarity based model for the analysis with EOV in the next chapter.

ANALYSIS WITH EOV

Earth Observation Vocabulary

EOV is developed by University of Tokyo and Group on Earth Observation (GEO) members. It consists of 146 observation parameters and its definition. This originated from the earth observation priorities suggested in the report of User Interface Committee(UIC) in GEO.

UIC is a task team to coordinate user requirements and to provide earth observation data and information for global users. In the task activity in 2010, the task team harvested over 1700 documents including the existing publicly-available documents, such as international reports, workshop summaries, conference proceedings, and national- and regional-level reports (GEO USI 2010). As the output meta-analyzed from the documents, the prioritized 146 observation items are list-upped.

However, when the list of observation items was suggested in the report, each definition was not obvious. Then, the semantic group in GEO and our University of Tokyo members worked to assign each observation item to the suitable definition which are basically referenced from the authoritative glossaries and then decided through several discussions with experts. This product is named as EOV. Subsequently, we digitized the observation items and converted them into SKOS (W3C 2009) based format for system-available. Figure 5 shows the part of EOV in SKOS format.

Figure 5: Earth Observation Vocabulary in SKOS format

Metric for analysis

In this analysis, we will attempt to reveal statistic tendency for all papers published in ACRS. The tendency is estimated through the summation of $ms_{cj}(d_k)$ given by the MNB based formula. The training set used for calculating $ms_{cj}(d_k)$ is leaned from EOV. The total similarities of all ACRS papers to the observation items are calculated and ranked by

$$sim_{cj}(D) = \sum_{k=1}^{n} ms_{cj}(d_k).$$

Result

Ranking:

Table 3 shows the result of EOV based ranking of ACRS 2010 all papers. The representative top 10 and bottom 10 in ranking are shown in table 3. The scores are normalized from $sim_{cj}(D)$.

For example, we can see “Land Cover”, which is really used for 14 titles of ACRS 2010 oral papers, marks at the rank 2. As another example, the rank of “Stratospheric Ozone” is low because there are few researches related to

ozone in ACRS 2010. Accordingly, we expect that this ranking actually reflects the characteristics of ACRS 2010 papers from the viewpoint of observation items.

Table 3: EOV based Ranking in ACRS 2010

rank	Observation item	score (x10^{-2})
1	Land Surface Temperature	0.9528
2	Land Cover	0.9125
3	Elevation	0.9049
4	Land Use	0.8891
5	Leaf Area Index (LAI)	0.8845
6	Field Cover (Continuous)	0.8832
7	Ice Depth	0.8633
8	River Flow Observations(=runoff)	0.8533
9	Sea Surface Temperature	0.8511
10	Sea Level(=sea surface height)	0.8451
…	…	…

rank	Observation item	score (x10^{-2})
…	…	…
137	Ambient Particulate Matter…(*1)	0.4278
138	Ocean Salinity	0.0030
139	Cloud Water/Ice Amounts	0.3895
140	Column Ozone Concentration	0.2997
141	Ambient Ozone Concentration	0.2817
142	Methane Concentration	0.2729
143	Stratospheric Ozone	0.2033
144	Carbon Dioxide Concentration	0.1278
145	Carbon Dioxide Partial Pressure	0.0372
146	Water Quality & Composition...(*2)	0.0000

(*1) Ambient Particulate Matter Composition (coarse)
(*2) Water Quality & Composition, pH and salinity, Dissolved Oxygen Content

Next, we calculate the EOV based ranking with ACRS 2011 papers. The result is shown in Table 4. Regarding the top and bottom 10 ranking, it seems similar to the result in 2010. In other word, it might be said that major topics and minor topics do not dramatically change in ACRS every year. As the changed points, "NDVI", "Vegetation Cover" and "Glacier/Ice Cap Elevation" rank in top 10.

Table 4: EOV based Ranking in ACRS 2011

rank	Observation item	score (x10^{-2})
1	Land Surface Temperature	0.9144
2	Elevation	0.8855
3	Leaf Area Index (LAI)	0.8740
4	Land Cover	0.8629
5	Ice Depth	0.8580
6	Land Use	0.8579
7	NDVI (*3)	0.8454
8	Sea Level(=sea surface height)	0.8438
9	Vegetation Cover	0.8403
10	Glacier/Ice Cap Elevation	0.8353
…	…	…

rank	Observation item	score (x10^{-2})
…	…	…
137	Currents	0.4810
138	Ocean Salinity	0.4484
139	Cloud Water/Ice Amounts	0.3692
140	Column Ozone Concentration	0.3243
141	Ambient Ozone Concentration	0.3081
142	Methane Concentration	0.2930
143	Stratospheric Ozone	0.2241
144	Carbon Dioxide Concentration	0.1392
145	Carbon Dioxide Partial Pressure	0.0285
146	Water Quality & Composition...(*4)	0.0000

(*3) Normalized Difference Vegetation Index
(*4) Water Quality & Composition, pH and salinity, Dissolved Oxygen Content

Trend change:

We calculated the difference between 2010 and 2011 by using above two rankings. The result is shown in Table 5. In 2011, there are huge disasters happened such as Tsunami in Japan. Also, the session "Earthquake and Tsunami" is proposed in ACRS 2011. Despite the situation, we could not find the scores of water related observation items increases. You wonder why? We guess the following for the question. Originally, the total rate of water related researches does not change because for example the titles including the keyword "Water" are counted as 13 in ACRS 2010 oral papers and as 12 in ACRS 2011. But, in 2011, the water ralated experts might concentrate on and shift to more applicable researches for Tsunami or the other disasters. This effected that the score of "Ocean Topography" or "Water Infiltration" became relatively low.

rank	Observation item	score (x10^{-2})
1	NDVI (*5)	0.0326
2	Suspended particulates/ … (*6)	0.0291
3	Ambient Ozone Concentration	0.0263
4	Stand Density/Height/Volume	0.0251
5	Column Ozone Concentration	0.0246
6	Surface Deformation	0.0244
7	Gross Primary Productivity	0.0233
8	Wave Direction	0.0228
9	Column Nitrogen Dioxide … (*7)	0.0215
10	Biodiversity	0.0209
…	…	…

rank	Observation item	score (x10^{-2})
…	…	…
137	Evapotranspiration	-0.0363
138	Land Surface Temperature	-0.0384
139	Sea Ice Surface (Skin) Temperature	-0.0388
140	Water Infiltration/… (*8)	-0.0410
141	Water run-off	-0.0461
142	Cloud Cover (cloud index)	-0.0477
143	Field Cover (Continuous)	-0.0490
144	Land Cover	-0.0496
145	River Flow Observations(=runoff)	-0.0563
146	Ocean Topography	-0.0626

(*5) Normalized Difference Vegetation Index
(*6) Suspended particulates/turbidity/water attenuation coefficient
(*7) Column Nitrogen Dioxide Concentration
(*8) Water Infiltration/Percolation-Land Surface

CONCLUSION

In this paper we operated the pilot analysis for ACRS proceedings with EOV. Through several experiments, we revealed the characteristics of ACRS papers and the trend from the viewpoint of observation items.

The fact, EOV is currently used for applications in GEOSS or DIAS. So, then the trial which links ACRS resources to the others is operated, this kind of EOV-based analysis would be required. Therefore, next we hope that we try to integrate academic papers and earth observation data.

In addition, although ACRS oral papers are officially classified into session topics, poster papers are not classified. So, if it is needed to classify the ACRS poster papers, this method might be useful.

ACKNOWLEDGEMENT

This study is supported by DIAS (Data Integration and Analysis System) project and the Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Japan.

REFERENCES:

Ono, M., Nagai, M., Shibsaki, R., 2010. Development of the multi-referential reverse dictionary for interoperating earth observation data. The 31nd Asian Conference on Remote Sensing,

GEO UIC, 2010. Task US-09-01a Final Report to the UIC. http://sbageotask.larc.nasa.gov/Cross-SBA_Report_GEO_US0901a.pdf.

Jason D. M. Rennie, Lawrence Shih, Jaime Teevan and David R. Karger., 2003. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003

Artur Aiguzhinov, Carlos Soares, Ana Paula Serra, 2010. A Similarity-Based Adaptation of Naive Bayes for Label Ranking: Application to the Metalearning Problem of Algorithm Recommendation, 16-26. In Discovery Science.

W3C, 2009. SKOS Simple Knowledge Organization System Reference. http://www.w3.org/TR/skos-reference/