Consideration on Sensitivity for Multiple Correspondence Analysis

Masaaki Ida *

Abstract— **Correspondence analysis is a popular research technique applicable to data and text mining, because the technique can analyze simple and multiple cross tables containing some measure of correspondence between the rows and columns and also has the strength of graphical display in two or three dimensions. The results of correspondence analysis provide information similar to factor analysis. Those abilities will deepen the global understanding on the characteristics of accumulated numerical and text information and have possibility leads to new knowledge discovery. We have so far focused on the text information of Japanese higher education institution such as syllabuses. We conducted research on collecting and analyzing the information, and on text mining to analyze and visualize the information for grasping the global characteristics of curricula of universities. Moreover, we have considered the sensitivity of ordinary simple correspondence analysis in case of data perturbation. However, more considerations on multiple correspondence analysis are required. This article presents mathematical considerations on sensitivity of multiple correspondence analysis.**

Keywords: Multiple correspondence analysis, sensitivity analysis, visualization, text mining

1 Introduction

Correspondence analysis is a method to analyze corresponding relation between data in multiple categories expressed by cross table. Correspondence analysis is frequently utilized for visualization in text mining. Various comprehensive considerations on overall accumulated data can be taken by executing the correspondence analysis. This method is closely related to homogeneity analysis, dual scaling, and quantification methods.

In a practical situation, elements of the cross table, numbers of elements, include uncertain information. Therefore, it is necessary to examine perturbations of the result of correspondence analysis.

This paper presents an extension of our former studies for sensitivity of simple correspondence analysis to *multiple* correspondence analysis. Section two shows an example of comparative analysis for curricula of some higher education institutions with data perturbation. Section three presents mathematical consideration on sensitivity of *multiple* correspondence analysis,

*National Institution for Academic Degrees and University Evaluation, Gakuen-Nishimachi Kodaira, Tokyo, 187-8587 Japan. Email: ida@niad.ac.jp

2 Comparative Analysis of Curricula

2.1 Syllabus Sets and Cross Table

We developed the curriculum analysis system based on clustering and correspondence analysis for syllabus data, which is applied for some departments of Japanese universities [1, 2, 3, 4].

Procedure of the analysis system is simply described as follows: (1) Collection of curriculum information or syllabuses. (2) Categorizing (clustering) based on the technical terms in contents of syllabuses. (3) Visualization for the distribution of syllabus or technical term categories and curricula by using correspondence analysis.

As an example [4], we choose five departments or five universities whose department name include *media* in Japanese from Japanese universities. These department names including *media* are listed in Table 1.

Table 1: Departments of five universities: "media"

Univ.	Name of department
U1	Department of Media Technology
U2	Department of Media and Image Technology
U3	Department of Multimedia Studies
U4	Department of Media Arts, Science and Technology
U5	Department of Social and Media Studies

In the categorizing procedure (2), we use the library classification system, NDC: *Nippon Decimal Classification* [5]. This system is based on using each successive digit to divide into ten divided categories. In order to utilize NDC as term category system for our curriculum analysis, we we select approximately 9,600 terms or keywords as shown in Table 2.

For our *media* department case, totally 22,962 keywords are selected from their syllabuses. Cross table for 20 categories (clusters) and 5 universities is shown in Table 3. We can perform the considerations on this (20×5) table.

Detailed considerations on Table 3 are possible by *cor-*

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Table 2: 20 term-categories for curriculum analysis

	Category
C1	Knowledge and academics, Information science
C2	Psychology
C3	Social Sciences, Economics, Statistics, Sociology, Education
C4	Natural Sciences
C5	Mathematics
C6	Physics
C7	Chemistry
C8	Technology and Engineering
C9	Construction, Civil engineering
C10	Architecture
C11	Mechanical engineering, Nuclear engineering
C12	Electrical and Electronic engineering
C13	Maritime and Naval engineering
C14	Metal and Mining engineering
C15	Chemical technology
C16	Manufacturing
C17	Domestic arts and sciences
C18	Commerce, Transportation and Traffic, Communications
C19	Music and Dance, Theater, Motion Pictures, Recreation
C20	Language

Table 3: Category-University cross tabulation (20×5)

	U1	U2	U3	U4	U5	sum
C1	953	311	159	331	9	1763
C2	236	418	129	227	35	1045
C3	1744	2181	799	1104	260	6088
C4	95	129	37	60	7	328
C5	1472	703	518	694	105	3492
C6	280	1147	153	324	3	1907
C7	74	453	59	82	2	670
C8	407	564	43	258	6	1278
C9	190	250	69	209	24	742
C10	217	520	77	274	22	1110
C11	252	116	61	117	0	546
C12	347	708	58	391	33	1537
C13	5	12	5	4	0	26
C14	20	30	19	25	3	97
C15	58	59	36	41	0	194
C16	29	40	11	49	1	130
C17	20	21	14	29	10	94
C18	48	221	14	243	14	540
C19	194	88	116	97	52	547
C20	507	63	101	140	17	828
sum	7,148	8,034	2,478	4,699	603	22,962

respondence analysis [6]. In this case, term frequencies are standardized to the total value in each university. Eigenvalues for the analysis are {0.0994, 0.0650, 0.0207, 0.0065}. Therefore, it is sufficient to analyze in *two dimensions* because the accumulated contribution ratio is 86% for first and second values. Figure 1 shows two-dimension graphical allocation so that we can grasp the global feature of Table 3. Points in the figure correspond to the vectors of categories ($\boldsymbol{x}_1^1,\ \boldsymbol{x}_2^1$) and universities ($\boldsymbol{x}_1^2,\ \boldsymbol{x}_2^2$).

We can obtain global understanding on this curriculum comparison. Interpretations for Figure 1 are as follows: (1) Top center part is an area of computer science or language. University 1 is related to these categories. (2) Lower right is an area of social science or psychology. University 5 is related to this category. (3) Lower left is an area of natural science and physics. University 2 is related to these categories. (4) University 3 and University 4 are located in center area and they have various syllabuses. These departments of University 3 and 4 are regarded as with average curricula.

2.2 Perturbation in Cross Table

As seen in the previous sub-section, various comprehensive considerations on the overall accumulated data can be taken by executing the correspondence analysis. However, elements of the cross table, numbers of keywords in syllabuses inevitably contain uncertainty. Therefore, it is necessary to examine sensitivity of the result of correspondence analysis for data perturbation in cross table.

When a certain perturbation of cross table would make great change for the result of correspondence analysis, the interpretation on the data should be very careful. Sometimes we have to correct the interpretation if necessary. On the other hand, the result of correspondence analysis might not be influenced by the change of cross table. In such robust case, interpretation of the result of correspondence analysis has robust and essentially important features. Therefore, it is important to examine mathematically the variations by perturbation of cross table.

For our *media* department case, the calculation result of data perturbation for Table 3 is visualized as shown in Figure 2 [4].

Figure 2 shows the case of perturbation on *C6 of universities U1 (Department of Media Technology).* U1 is located in the Top upper part of each figure. In Figure 2, locations of categories and universities are corresponding to the locations in Figure 1. Arrows in the figures are the vectors of differential coefficients;

$$\left(\frac{d\boldsymbol{x}_1^1}{da_{pq}},\ \frac{d\boldsymbol{x}_2^1}{da_{pq}}\right),\quad \left(\frac{d\boldsymbol{x}_1^2}{da_{pq}},\ \frac{d\boldsymbol{x}_2^2}{da_{pq}}\right).$$

The (green) arrow in upper part of the figure means that if the number of keywords in syllabuses of U1 for Cluster 6 increase, then U1 moves to *left* direction. U1 goes more closer to the group of natural science as *physics* or *chemistry.*

Therefore, we are able to understand sensitivities, and expect change for the result of correspondence analysis clearly for variations in text analysis.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

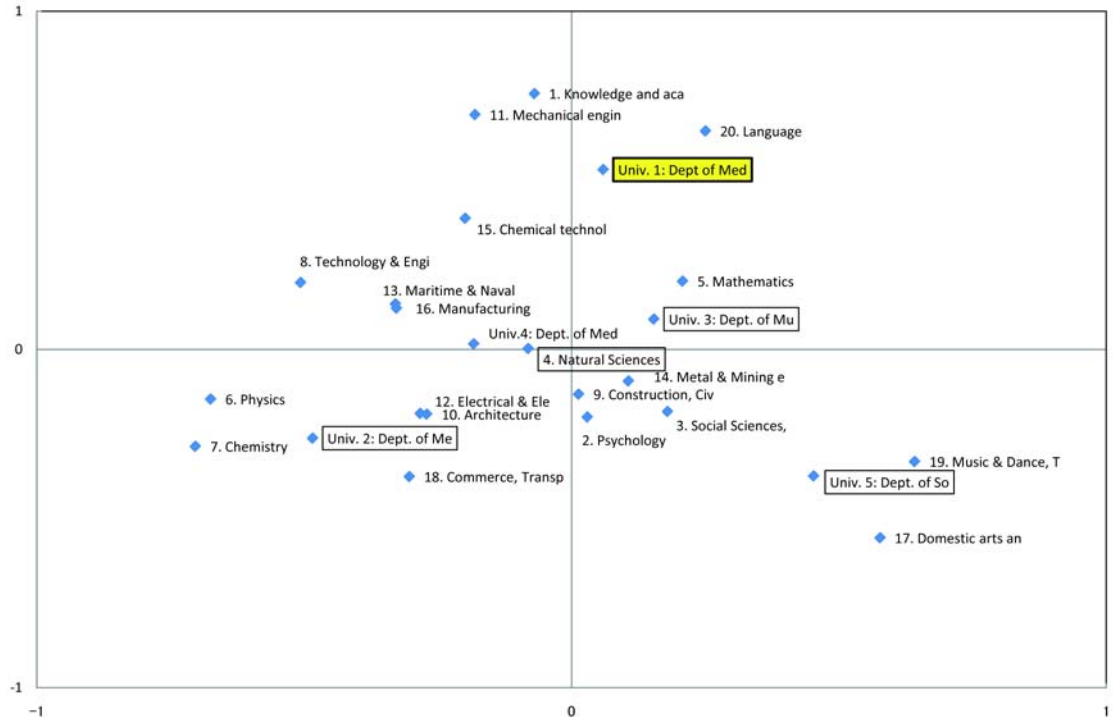

Figure 1: Result of correspondence analysis: *scores* in two dimension for 5 universities and 20 categories.

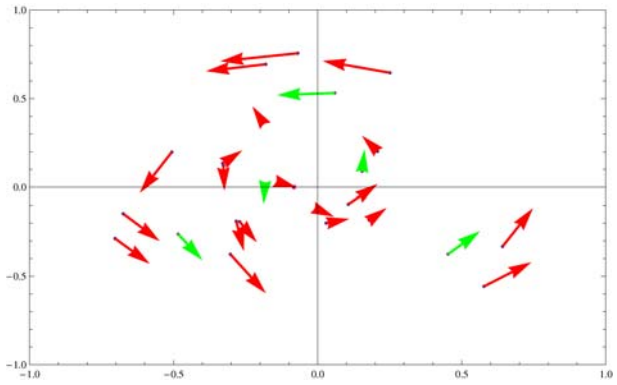

Figure 2: Sensitivity for C6 of universities Univ.1 (department of Media Technology).

3 Multiple Correspondence Analysis

3.1 Formulation of Multiple Correspondence Analysis

Multiple correspondence analysis is an extension of simple correspondence analysis which is related to various forms of multivariate analysis methods as *third method of quantification*, *dual scaling*, and *homogeneity analysis* and others, that are essentially based on singular value decomposition (e.g., [6]). In this section, mathematical considerations on multiple correspondence analysis are discussed.

As an example, Table 4 illustrates a part of r categorical data for a certain survey. Each row of this table corresponds to a binary coding of factors, also called dummy variables.

Table 4: Example of data table

	G^1			G^2					$\cdots$	G^r	
1	0	0	1	1	0	0	0	0	$\cdots$	1	0
2	1	0	0	0	0	1	0	0	$\cdots$	1	0
3	0	1	0	1	0	0	0	0	$\cdots$	0	1
4	0	1	0	0	0	0	1	0	$\cdots$	1	0
5	0	0	1	0	1	0	0	0	$\cdots$	0	1
6	0	1	0	0	0	0	1	0	$\cdots$	0	1
$\vdots$		$\vdots$				$\vdots$			$\vdots$	$\vdots$	

This table is expressed by an indicator matrix G as

$$G = (G^1, G^2, \cdots, G^r). \quad (1)$$

where G^k is a binary matrix that exactly one element in each row is equal to one.

A cross table or contingency table shown in Table 5 is the joint distribution or correspondence between the vari-

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

ables with categories $(A^s_1, A^s_2, \cdots, A^s_m)$ and the variables with categories $(A^t_1, A^t_2, \cdots, A^t_m)$. These categories are all inclusive. Each element (a^{st}_{ij}) shows the number of specific combination of responses to the categories A^s_i and A^t_j.

Table 5: Cross table for $A^s \times A^t$

	A^t_1	A^t_2	$\cdots$	$A^t_{n^t}$
A^s_1	a^{st}_{11}	a^{st}_{12}	$\cdots$	$a^{st}_{1n^t}$
A^s_2	a^{st}_{21}	a^{st}_{22}	$\cdots$	$a^{st}_{2n^t}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$A^s_{m^s}$	$a^{st}_{m^s1}$	$a^{st}_{m^s2}$	$\cdots$	$a^{st}_{m^sn^t}$

This cross table for two dimensions is denoted by $m^s \times n^t$ matrix A^{st} (sometimes we abbreviate s and t),

$$A^{st} = \begin{pmatrix} a^{st}_{11} & a^{st}_{12} & \dots & a^{st}_{1n^t} \\ a^{st}_{21} & a^{st}_{22} & \dots & a^{st}_{2n^t} \\ \vdots & \vdots & \vdots & \vdots \\ a^{st}_{m^s1} & a^{st}_{m^s2} & \dots & a^{st}_{m^sn^t} \end{pmatrix}. \tag{2}$$

The matrix A^{st} has the following relation with the matrices G^s and G^t.

$$A^{st} = (G^s)^{\mathrm{T}} G^t \tag{3}$$

where "T" denotes transform of the matrix.

We denote $m^s \times m^s$ diagonal matrix by D^s (and $n^t \times n^t$ diagonal matrix by $\tilde{D}^t$) that has diagonal elements equal to the sum of each column (and row) of the matrix A^{st} as follows:

$$D^s = \mathrm{diag}(d^s_{11}, \cdots, d^s_{m^sm^s}),\ d^s_{ii} = \sum_{j=1}^{n^t} a^{st}_{ij} \tag{4}$$

$$\tilde{D}^t = \mathrm{diag}(d^t_{11}, \cdots, d^t_{n^tn^t}),\ d^t_{jj} = \sum_{i=1}^{m^s} a^{st}_{ij}. \tag{5}$$

It is natural to assume that $d^s_{ii} > 0$ and $d^t_{jj} > 0$. We have the relation as

$$D^s = (G^s)^{\mathrm{T}} G^s \tag{6}$$

$$\tilde{D}^t = G^t (G^t)^{\mathrm{T}}. \tag{7}$$

The vectors given to multiple variables are denoted by $\boldsymbol{x}^s = (x^s_1, \dots, x^s_{m^s})^{\mathrm{T}}$, $\boldsymbol{x}^t = (x^t_1, \dots, x^t_{n^t})^{\mathrm{T}}$ and so on. These vectors are called *scores* in correspondence analysis. The scores of multiple correspondence analysis are the solutions for the following eigenvalue problem:

$$\begin{pmatrix} D^1 & A^{12} & \dots & A^{1r} \\ A^{21} & D^2 & \dots & A^{2r} \\ \vdots & \vdots & & \vdots \\ A^{r1} & A^{r2} & \dots & D^r \end{pmatrix} \begin{pmatrix} \boldsymbol{x}^1 \\ \boldsymbol{x}^2 \\ \vdots \\ \boldsymbol{x}^r \end{pmatrix}$$

$$= r\,\lambda \begin{pmatrix} D^1 & & & 0 \\ & D^2 & & \\ & & \ddots & \\ 0 & & & D^r \end{pmatrix} \begin{pmatrix} \boldsymbol{x}^1 \\ \boldsymbol{x}^2 \\ \vdots \\ \boldsymbol{x}^r \end{pmatrix}. \tag{8}$$

Namely,

$$G^{\mathrm{T}} G \boldsymbol{x} = r\lambda D \boldsymbol{x} \tag{9}$$

$$D = \mathrm{diag}(G^{\mathrm{T}} G) \tag{10}$$

where $\boldsymbol{x} = ({\boldsymbol{x}^1}^{\mathrm{T}}, \cdots, {\boldsymbol{x}^r}^{\mathrm{T}})^{\mathrm{T}}$.

We denote $\boldsymbol{u}^s$ and H^{st} as follows:

$$\boldsymbol{u}^s = (D^s)^{\frac{1}{2}} \boldsymbol{x}^s \tag{11}$$

$$(D^s)^{\frac{1}{2}}_{ii} = \sqrt{d^s_{ii}} \tag{12}$$

$$H^{st} = (D^s)^{-\frac{1}{2}} A^{st} (\tilde{D}^t)^{-\frac{1}{2}}. \tag{13}$$

Then,

$$K \begin{pmatrix} \boldsymbol{u}^1 \\ \boldsymbol{u}^2 \\ \vdots \\ \boldsymbol{u}^r \end{pmatrix} = r\lambda \begin{pmatrix} \boldsymbol{u}^1 \\ \boldsymbol{u}^2 \\ \vdots \\ \boldsymbol{u}^r \end{pmatrix} \tag{14}$$

where the matrix I means identity matrix,

$$K = \begin{pmatrix} I^1 & H^{12} & \dots & H^{1r} \\ H^{21} & I^2 & \dots & H^{2r} \\ \vdots & \vdots & & \vdots \\ H^{r1} & H^{r2} & \dots & I^r \end{pmatrix}. \tag{15}$$

Eigenvalue is denoted by λ_i, and its corresponding eigenvector by $\boldsymbol{u}_i$ which corresponds to $\boldsymbol{x}_i$:

$$\boldsymbol{u}_i = ({\boldsymbol{u}_i^1}^{\mathrm{T}}, {\boldsymbol{u}_i^2}^{\mathrm{T}}, \cdots, {\boldsymbol{u}_i^r}^{\mathrm{T}})^{\mathrm{T}}. \tag{16}$$

Hence

$$K\boldsymbol{u}_i = r\lambda_i \boldsymbol{u}_i. \tag{17}$$

We have $\lambda_i \geqq 0$ in the equation. Moreover, the maximal eigenvalue $\lambda_0 = 1$ and $\boldsymbol{x}_0 = \mathbf{1}$ ($\boldsymbol{u}_0 = D^{\frac{1}{2}}\mathbf{1}$). Therefore, we omit this trivial case with $\lambda_0 = 1$. In the following discussion, we consider the case that the eigenvalues, $0 < \lambda_i < 1$, and $\lambda_i \geqq \lambda_j$ for $(i < j)$ and as follows:

$$0 < \lambda_i < 1, \tag{18}$$

$$\boldsymbol{u}_i \cdot \boldsymbol{u}_i = 1, \text{ and } \boldsymbol{u}_i \cdot \boldsymbol{u}_j = 0\ (i \neq j). \tag{19}$$

Concerning the value of χ^2 we have the following relation:

$$\chi^2 = \sum_{s,t,s\neq t} \sigma^{st}(\mathrm{tr}((H^{st})^{\mathrm{T}} H^{st}) - 1) + \sum_s \tilde{\sigma}^s (m^s - 1) \tag{20}$$

where $\sigma^{st} = \sum_{i,j} a^{st}_{ij}$ and $\tilde{\sigma}^s = \sum_i d^s_{ii}$.

In the case of Section 2, the value of χ^2 is 116.9.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

3.2 Sensitivity of Multiple Correspondence Analysis

In this subsection we consider the case that there exists data perturbation in the element of matrix A^{st}. The differential coefficients of the eigenvectors for a_{pq}^{st} are shown as follows:

$$(K - \lambda_i I)\boldsymbol{u}_i = \mathbf{0} \tag{21}$$

then following equations are obtained according to the characteristics of eigenvalue and eigenvector:

$$\left(\frac{dK}{da_{pq}^{st}} - \frac{d\lambda_i}{da_{pq}^{st}}\right)\boldsymbol{u}_i + (K - \lambda_i I)\frac{d^{st}\boldsymbol{u}_i}{da_{pq}^{st}} = \mathbf{0}. \tag{22}$$

We obtain the following equations with vector standardization:

$$\frac{d\lambda_i}{da_{pq}^{st}} = \boldsymbol{u}_i^{\mathrm{T}}\left(\frac{dK}{da_{pq}^{st}}\right)\boldsymbol{u}_i \tag{23}$$

$$\frac{d\boldsymbol{u}_i}{da_{pq}^{st}} = \sum_{j\neq i}\frac{\boldsymbol{u}_j^{\mathrm{T}}\left(\frac{dK}{da_{pq}^{st}}\right)\boldsymbol{u}_i}{\lambda_i - \lambda_j}\,\boldsymbol{u}_j \tag{24}$$

where

$$\frac{dK}{da_{pq}^{st}} = \begin{pmatrix} 0 & \cdots & \cdots & 0 \\ \vdots & 0 & (\frac{dH^{st}}{da_{pq}^{st}}) & \vdots \\ \vdots & (\frac{dH^{st}}{da_{pq}^{st}})^{\mathrm{T}} & 0 & \vdots \\ 0 & \cdots & \cdots & 0 \end{pmatrix}. \tag{25}$$

We can calculate the value of $\frac{dH^{st}}{da_{pq}^{st}}$ as follows:

$$\frac{dH^{st}}{da_{pq}^{st}} = L_{pq}^{st} + M_p^{st}H^{st} + H^{st}N_q^{st} \tag{26}$$

where L_{pq}, M_p, N_q are $m\times n$, $m\times m$, $n\times n$ sparse matrices, only one element of that has a value (other elements are zero) as follows:

$$L_{pq}^{st} = p\begin{pmatrix} 0 & \overset{q}{\vdots} & 0 \\ \cdots & \frac{1}{\sqrt{d_{pp}^{st}d_{qq}^{st}}} & \\ 0 & & 0 \end{pmatrix} \tag{27}$$

$$M_p^{st} = p\begin{pmatrix} 0 & \overset{p}{\vdots} & 0 \\ \cdots & -\frac{1}{2d_{pp}^{st}} & \\ 0 & & 0 \end{pmatrix} \tag{28}$$

$$N_q^{st} = q\begin{pmatrix} 0 & \overset{q}{\vdots} & 0 \\ \cdots & -\frac{1}{2d_{qq}^{st}} & \\ 0 & & 0 \end{pmatrix}. \tag{29}$$

We have

$$\frac{d\boldsymbol{x}_i^s}{da_{pq}^{st}} = \frac{d(D^s)^{-\frac{1}{2}}}{da_{pq}^{st}}\boldsymbol{u}_i + (D^s)^{-\frac{1}{2}}\frac{d\boldsymbol{u}_i^s}{da_{pq}^{st}}. \tag{30}$$

Calculating the values of $\frac{dH}{da_{pq}}$, we can easily obtain each numerical value of

$$\left(\frac{d\boldsymbol{x}_1}{da_{pq}^{st}}, \frac{d\boldsymbol{x}_2}{da_{pq}^{st}}, \cdots\right)$$

with the matrices

$$\frac{d(D^s)^{-\frac{1}{2}}}{da_{pq}^{st}}.$$

Using these derivatives, we can visualize the perturbations of the scores in correspondence analysis.

Moreover, according similar discussion we can calculate the values of

$$\frac{d\chi^2}{da_{pq}^{st}}.$$

For example, the case of $r = 2$ that corresponds to the problem in Section tow, we calculate the sensitivity of result of correspondence analysis. The new result is coincidence to the result of Section two that is shown in Figure 2.

4 Conclusion

In this paper, we consider the sensitivity for *multiple correspondence analysis* and curriculum comparison. Especially, derivatives of the scores for the multiple correspondence analysis was deduced. We will refine mathematical discussion on perturbation problem, for example on χ^2, and also improve the web-based visualization system for larger sensitivity analysis problem that easily visualizes the results for many variations simultaneously.

References

[1] Nozawa, T., Ida, M., Yoshikane, F., Miyazaki, K. and Kita, H., "Construction of Curriculum Analyzing System based on Document Clustering of Syllabus Data", *Journal of Information Processing Society of Japan*, Vol. 46, No.1, pp.289–300, 2005.

[2] Ida, M., Nozawa, T., Yoshikane, F., Miyazaki, K. and Kita, H., "Development of Syllabus Database and its Application to Comparative Analysis of Curricula among Majors in Undergraduate Education", *Research on Academic Degrees and University Evaluation*, Japanese, No.2, pp.85-97, 2005.

[3] Ida, M., Nozawa, T., Yoshikane, F., Miyazaki, K. and Kita, H., "Syllabus Database System and its Application to Comparative Analysis of Curricula",

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

6th International Symposium on Advanced Intelligent Systems, 2005.

[4] Ida, M, "Textual Information and Correspondence Analysis in Curriculum Analysis", *Proceedings of FUZZIEEE2009*, 2009.

[5] NDC, Nippon Decimal Classification, http://en.wikipedia.org/wiki/Nippon_Decimal_Classification

[6] Benzecri, J.-P., *Correspondence Analysis Handbook*, Marcel Dekker, 1992.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)