Integration of Data Providing and Analyzing System and its Application to Higher Education Institutional Data

Masaaki Ida

National Institution for Academic Degrees and
Quality Enhancement of Higher Education
Tokyo, Japan 187–8587
Email: ida@niad.ac.jp

***Abstract*—There exist various kinds of data providing and analyzing service on Web sites. Japanese College and University Portraits is an information system consisting of databases with Web services for providing information concerning various activities undertaken by universities and junior colleges, covering national, prefectural, municipal, and private institutions. This paper describes the outline of this integrated system and related analysis systems. Especially, we focus on data providing service in several ways including research results conducted by the research department of National Institution for Academic Degrees and Quality Enhancement of Higher Education. A further advanced and integrated data analysis and data visualization system can be developed by using Web APIs with various multivariate analysis methods. Canonical correlation analysis is one of the basic and requisite data analysis and visualization skills for data analysts in this Big Data era. Therefore, because data is received through Web APIs, the development of an integrated data analysis system equipped with canonical correlation analysis is desirable. This article also presents a work-in-progress result of the canonical correlation analysis for higher education institutional data.**

Keywords–Higher education institutional data; data providing; Web API; visualization.

I. INTRODUCTION

Education-related databases are important for college selections or various quality assurance activities, such as reporting and data analysis in higher education institution. Therefore, data service of higher education institutional data is desired to be developed. However, Institutional data of universities, e.g., the number of various kinds of academic staffs, are difficult to analyze because they were not necessarily standardized and integrated in each university itself or even in national level education-related agencies. Some advanced higher education integrated data systems are progressively developing. The most famous and useful system is the Integrated Postsecondary Education Data System [1], which has been developed by National Center for Education Statistics (NCES) in the United States. The system collects and analyzes basic institution information about universities and colleges in the U.S. The system standardizes and accumulates this information nationwide. This system comprehensively holds general and basic institution data. Moreover, this system is equipped with data analysis tools to conduct university comparative analysis. There exist other web-based university database systems in the U.S. and other countries. These databases are well-organized and comprehensive systems with easy Web-based operation on their Web sites. However, in order to cooperate or integrate with other information systems, e.g., in-house database developed in individual institutions, or external database services, more improved systems are expected to be equipped with various Web service functions and standardized data sets. In this paper, in Section II, the integration of data providing and analyzing system in Japan is described. In Section III, Web API and data analysis is described. As an example of data analysis, canonical correlation analysis is introduced with a numerical example.

II. INTEGRATION OF DATA PROVIDING AND ANALYZING SYSTEM

A. Japanese college and university portraits

In Japan, Ministry of Education, Culture, Sports, Science and Technology collects basic information about higher education institutions in Japan. This law-based basic statistical data includes yearly information of higher education institutions, such as the number of faculties or staffs, the number of enrolled students by grade (undergraduate, graduate, foreign student), the number of graduates by subsequent course, the number of those who are employed after graduation by each industry and by occupation, faculties, facilities, and financial data. However, these are published as statistical data, so that detailed information of individual universities are not published.

Japanese College and University Portraits is an information system consisting of database with Web services for providing information concerning various activities undertaken by universities and junior colleges, covering national, prefectural, municipal, and private institutions [2]. System operation started in March 2015. The system is managed by National Institution for Academic Degrees and Quality Enhancement of Higher Education, Japan (NIAD-QE) associated with Promotion and Mutual Aid Corporation for Private Schools of Japan.

The purposes of the system are as follows

- Information Dissemination: The Portrait Website will be used not only by those who intend to participate in higher education as students, but also by stakeholders in various areas of society, such as government and industry. The database is also expected to be an information source contributing to improve international society's understanding of higher education institutions in Japan.
- Monitor and Analysis of Institution Activities: The system is expected to be used by higher education institutions to monitor and analyze the status of their own educational activities for internal quality assurance and enhancement.
- Workload Reduction: Collection and publication of fundamental and standardized data in the database system will assist higher education institutions when they

 ISBN: 978-1-61208-666-8

respond to various surveys and external evaluation. Workload reduction of institutions based on accurate data are expected to be accomplished by the system.

Data items stored in the system are in multiple levels, Institution level and faculty level, for example, general information of higher education institution, objectives of education and research, characteristics, education system organization, campus, university evaluation, student support, policies of education, academic program, admission, faculty, enrollment, scholarship, completion, post graduate pathways, employment, research activities, international activity, student life, financial information and so on.

The database system consists of three databases, three circles in Figure 1, with basic organization data located in the common part of circles, which is regarded as university data warehouse. These database are classified in the following three categories:

- University common publication data: published common education data over national, prefectural, private institution,
- University basic data: corresponding to school basic survey,
- National university evaluation data: used for national university evaluation.

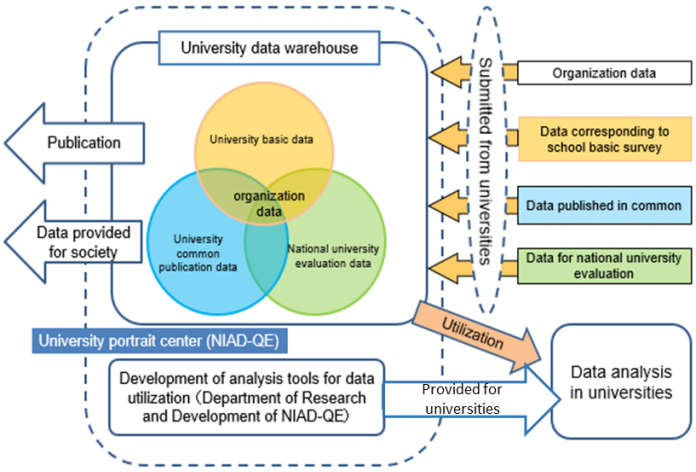

Figure 1. Japanese College and University Portraits.

Each data is registered by data-sheet submission from higher education institutions. The stored data are used for publication and provided for data utilization for the sake of society and universities.

B. Data providing service

Data providing and utilization services are explained in the following ways. These functions are partially equipped in the working system. Data analysis tools for data utilization are now being developed by research department of NIAD-QE.

1) Data providing via Website: The system has search (retrieval) functions, (1) simple search by university name, faculty name and location, (2) detailed search, e.g., by entrance examination, student financial aid and so on, and (3) keyword search. Adding to searching and utilizing the data which are ordinary in table format in Web pages, we can download data tables, and utilize them in spreadsheet software in user side. Then it is possible to extract necessary data for analysis, and to conduct data analysis by using personal analysis tools or personal Business Intelligence (BI) tools on user's local environment, which are popular tools in these days.

2) Data providing via BI tool: Highly-detailed and flexible data analysis can be attained by Structured Query Language (SQL). However, expert ability is required for such advanced treatment of database. In case that we intend to try advanced data analysis without expert ability, full-scale BI tools are candidates of effective analysis with great potential, which is equipped in the Portrait system. BI tool makes it possible for system registered users to utilize the database more conveniently with some useful BI functions, such as filter, formula, chart and drill-down functions. We can generate various kinds of easily understandable data tables and charts, and also generate data analysis report file in PDF format or spread sheet format. This BI tool of the system was used to generate data analysis reports in National University Corporation Evaluation in Japan.

3) Data providing via Web-based analysis system: Data analysis and data visualization tool are being developed by research department of NIAD-QE. Figure 2 shows an example of comparative analysis of universities. We refer European university comparison and visualization systems, U-Map [3] and U-Multirank [4]. They are new higher education transparency tools for multi-dimensional mapping and ranking [5]. In this figure, data table includes selected eight indicators in columns for selected 13 universities in rows for corresponding fiscal year. Values of indicators are transformed into relative classes or groups (e.g., four level: quantile point), which are expressed by the number of star marks. Chart in lower side shows feature of three universities selected from universities in this data table. Fan-shaped parts of the charts, surrounding center circle, correspond to the amount of indicators. Relative analysis by class or group is helpful for understanding whole aspect of higher education institutions with multiple features.

4) Data system integration: Data providing via Web API: Web API is a Web Application Programming Interface for performing computer processing via Internet. This mechanism makes it possible for registered users to access external database through the internet. The advantages to use Web API are to obtain data when necessary, to obtain only necessary part of data by query (search), to obtain standardized and latest data that might be updated recently, and to have possibility to provide more useful and valuable information combining with other multiple external data sources provided by other Web APIs such as official government statistics or location information Web service. Moreover, this type of Web services has an effect for developing application modules with independency, which leads to improvement of maintenance and redesign of database application system.

Research department of NIAD-QE is developing various kinds of Web APIs and their applications which are suitable for data analysis and data dissemination. Web APIs for university basic survey (for national and prefectural) and university financial data have being developed. Output form can be selected in JSON or XML formats. University basic survey sheets consist of detailed university information cards in university level or department level, e.g., institutional structure, faculty member and staff (sheet 7: number of students, number of

 ISBN: 978-1-61208-666-8

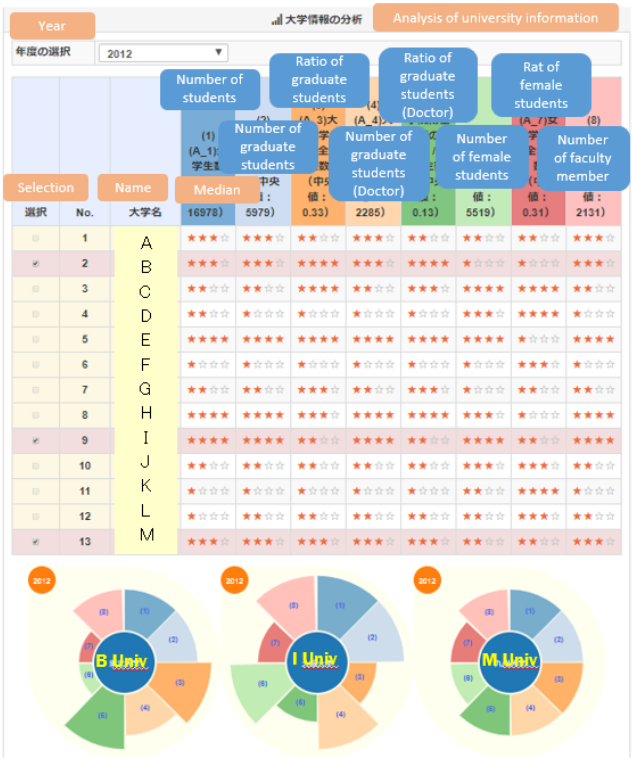

Figure 2. Data Analysis and ata Visualization Tool.

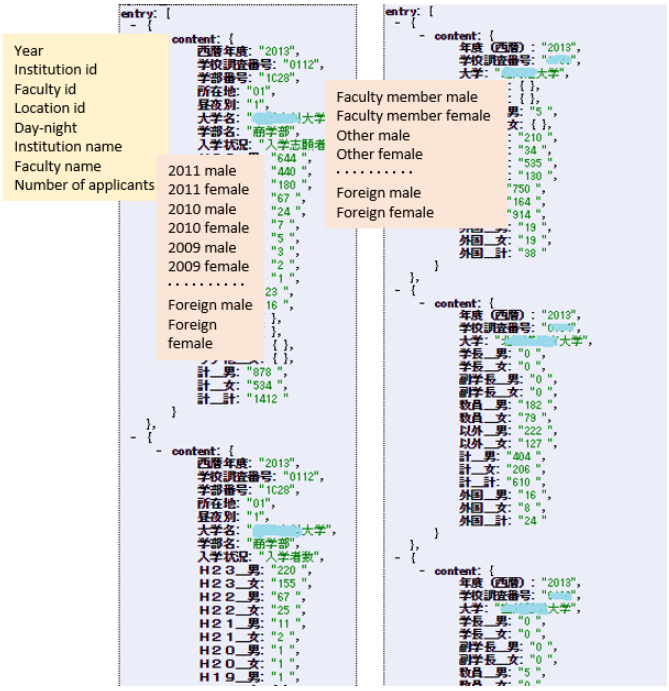

Figure 3. Web API of university basic information (JSON format).

academic staffs), student (sheet 8: number of students of each department), graduate student (sheet 9), foreign student (sheet 11), facility (sheet 20), graduation and employment (sheet 30) and so on. This type of detailed university Web API development is early attempt in higher education field.

The followings are examples retrieved by survey year and institution code of Japanese universities. Figure 3 (left) shows an example of Web API output concerning applicant and enrollment figures by undergraduate school in faculty level. The elements in Japanese language mean university name, faculty name, and the number of undergraduate students in every fiscal year. Figure 3 (right) shows an example of output of the Web API concerning faculty members from survey in university level.

API key is issued for registered user to access and use API functions for security. Retrieved data is provided through cryptographic protocol that provides communication security over computer network.

III. Web API and Data Analysis

A. Analysis and visualization process

Generally, there exist various kinds of data analysis services on Web sites. Further complex and advanced data analysis tool and data visualization applications can be developed by data integration mechanism using Web API functions. By utilizing Web APIs, we can develop flexible integrated Web applications with data tables and charts generation, and data analysis system. With flexibilities of API mechanism more useful and user-friendly data visualization system can be developed.

The analysis and visualization process is as follows:

1) Database query by university name or department name with various indicators is submitted to university information Web API site with API key for registered user, which is also developed by research department of NIAD-QE.
2) Data in JSON or XML format are received by Web programming on server side or client side.
3) Analysis and visualization of various indicators with effective graphic libraries are conducted, and comparison of multiple indicators with sorting functions on data tables or charts is made.
4) Moreover, analysis system can be programmed to combine with other databases using API functions, e.g., various official statistical data API or map API on outer Web service sites. These Web service combination, or mash up programing, can be easily applied using Web API functions.

B. Canonical correlation analysis

Canonical correlation analysis (CCA) is a core analysis method in multivariate analysis field. CCA is a generalized method of corresponding analysis that is useful for questionnaire analysis [6]–[8]. Two multiple variable data matrices, X and Y are expressed with $n \times p$ and $n \times q$ real data matrices, X_R and Y_R. We define the following matrices for data average and deviation:

$$Q_n = I_n - (1/n)\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$$

$$X = Q_n X_R, \quad Y = Q_n Y_R$$

where $\mathbf{1}_n$ means $(1, 1, \cdots, 1)^{\mathrm{T}}$, and Q_n means averaging.

Calculating correlation matrices, R_{XX}, R_{YY}, R_{XY}, for X, Y, then, the result of canonical correlation analysis is the singular value and corresponding singular vectors, μ with $\boldsymbol{a}$ and $\boldsymbol{b}$ satisfy the following matrix equation:

$$\begin{pmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY} \end{pmatrix} \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{pmatrix} = (1+\mu) \begin{pmatrix} R_{XX} & O \\ O & R_{YY} \end{pmatrix} \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{pmatrix}$$

We call the singular value μ as the first canonical correlation coefficient μ^1 for maximum value, and μ^2 for second one, and μ^3 for third one and so on. And its corresponding vectors, $\boldsymbol{a}^i$ and $\boldsymbol{b}^i$ for μ^i, are called score vectors. Similarly, we calculate the vectors $\boldsymbol{f}^i$ and $\boldsymbol{g}^i$ as follows;

$$\boldsymbol{f}^i = X\boldsymbol{a}^i \quad (1)$$

$$\boldsymbol{g}^i = Y\boldsymbol{b}^i \quad (2)$$

In this paper, we consider $\mu^i, \boldsymbol{a}^i, \boldsymbol{b}^i, \boldsymbol{f}^i, \boldsymbol{g}^i$ for understanding the arrangement of each element and tendency of whole data set.

C. Numerical example

As an example of canonical correlation analysis with Web API, we show the result of analysis for university financial data (work in progress); In this case, the number of items for X expressing incomes is four (management expenses grant, tuition, research grant, donation), and the number of items for Y expressing expenses is three (general management expenses, research expenses, education expenses). Figure 4 shows the result of canonical correlation analysis in two dimensions; $(\boldsymbol{a}^1, \boldsymbol{a}^2)$ and $(\boldsymbol{b}^1, \boldsymbol{b}^2)$ for μ^1, μ^2. For $\boldsymbol{f}^i, \boldsymbol{g}^i$, Figure 5 show the arrangement of each university incomes and expenses in two dimensions.

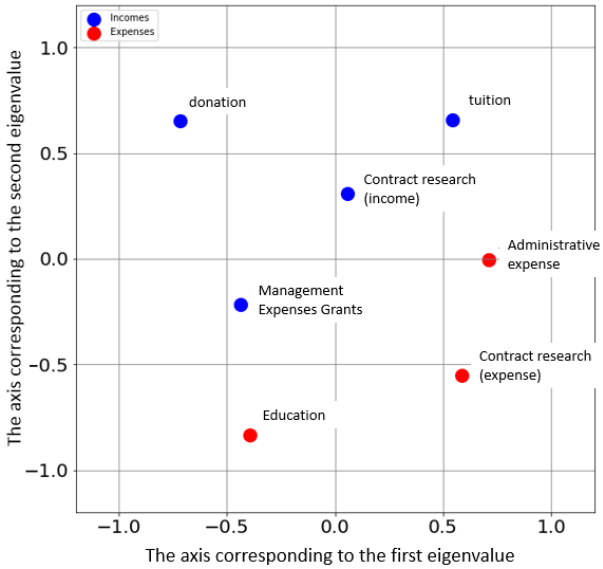

Figure 4. Example of canonical correlation analysis (1).

Figure 4 shows visually summarized information in two dimensions, which are high accumulation contribution of eigenvalues. We can grasp the global feature of financial situations of universities by this arrangement. Figure 5 shows the proximity between universities (university ID) and the tendency of whole data set.

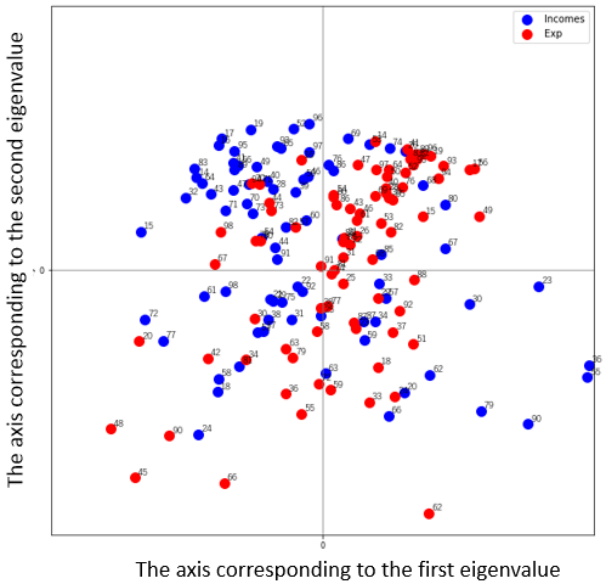

Figure 5. Example of canonical correlation analysis (2).

In this way, we can take a global view of the clustering. various comprehensive considerations on overall of accumulated data can be taken by executing the analysis. Various comprehensive considerations on overall accumulated data can be taken by executing the analysis. Those abilities will deepen the global understanding on the relations of accumulated multiple information, and which have promising possibility leads to new knowledge discovery.

IV. Conclusion

This paper describes the outline of the integrated system, Japanese College and University Portraits and related data analysis systems. Especially, we focus on several data providing services and utilization of Web API function. This type of university Web API development is early attempt in higher education field. In order to handle more general university data, coordination of differences between the data definition is needed for useful comparison. We hope that our development and attempt will play an important role as an infrastructure for data utilization and data analysis in higher education quality assurance.

References

[1] "Integrated Postsecondary Education Data System, IPEDS", URL; nces.ed.gov/ipeds [accessed: 2018-07-20].

[2] "Japanese College and University Portraits", URL; top.univ-info.niad.ac.jp [accessed: 2018-07-20]

[3] "U-Map", URL; http://www.u-map.eu [accessed: 2018-07-20].

[4] "U-Multirank", URL; http://www.umultirank.org [accessed: 2018-07-20].

[5] D. F. Westerheijden, "Multi-dimensional Mapping and Ranking New Higher Education Transparency Tools", NIAD-UE University Quality Assurance Forum 2014 Keynote speech, URL; www.niad.ac.jp/n_kenkyukai/no13_2014forum_keynote.pdf [accessed: 2018-07-20].

[6] B. Thompson, *Canonical Correlation Analysis: Uses and Interpretation*, Sage Publications, 1985.

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

[8] M. Greenacre, *Correspondence Analysis in Practice, Second Edition*, Chapman and Hall/CRC, 2007.

 ISBN: 978-1-61208-666-8