Developing a Web-based GNH Survey Application and a Web Service for Identifying Missing Data Mechanism in Incomplete Survey Data

Sonam TSHERING[†], Takeo OKAZAKI[††] and Satoshi ENDO[†††]

Department of Information Engineering University of the Ryukyus, Okinawa 903-0213 Japan

Summary
This paper proposes a web-based GNH survey application, capable of identifying missing data mechanism which is crucial in effectively handling missing data in survey questionnaire. First, we've stated the rationale behind GNH survey briefly highlighting the concept and the birth of GNH. Second, we've described the present approach of conducting GNH survey with special emphasis on the problems associated with it. Third, we've shown our proposed GNH survey application, demonstrating how the problems of current approach can be rectified and improved upon. In addition, the issues of missing data in a database and as to how to handle it via a web service are also discussed and implemented. Finally, we've presented how a web service has been developed and integrated into GNH survey application.

Key words:
GDP, GNH, survey, application, missing data mechanism, web services.

1. Introduction

Gross Domestic Product popularly known as GDP is the market value of all officially recognized final goods and services produced within a country in a given period. Although it was purely designed to measure the market value of production that flows through the economy, it became the main tool for measuring the welfare of a country. While it has led to improved standard of living, it has also caused social exclusion, poverty, misery, environmental pollution and degradation, etc.

Pained by these grievous consequences, the Government of Bhutan adopted Gross National Happiness widely known as GNH as its development philosophy. This philosophy is premised on the belief that citizens' happiness is more important than the economic development.

While GDP is a sum of Consumption (C), Investment (I), Government Spending (G) and Net Exports (X-M), GNH is a sum of Economic development, Psychological, Emotional, and Spiritual well-being.

Symbolically;

$$GDP = C + I + G + (X - M);\ \text{while}$$

GNH =Living standards + Health system + Education system + Good governance + Community vitality + Culture + Ecology + Time use + Psychological well-being

Manuscript received August 5, 2012
Manuscript revised August 20, 2012

Statistics on these components of GNH are vital as in the case of GDP to operationalize GNH philosophy or to incorporate it into development plans and programs. For this purpose, the Bhutanese government conducts a survey and collects data pertaining to the 9 domains of GNH. But the current paper-and-pencil method of data collection has a couple of problems, so we proposed and developed a web-based GNH survey application along with a web service for identifying missing data mechanism which is crucial in handling missing data.

1. Current Approach of Conducting GNH Survey: Paper-and-Pencil Method

The present method of carrying out GNH survey is that an enumerator goes to a respondent with a questionnaire, interviews the respondent and records the answers on a questionnaire. Next, officials at office collect all questionnaires and then manual recoding, cleansing, and recording take place before analysis. But a number of problems arise from this method of data collection, recoding, cleansing and recording: It is tedious, time consuming, and expensive since everything—from data collection to data entry to data validation to skip pattern has to be done manually.

Taking these problems into account we've proposed and developed a web-based survey application based on the requirement analysis of the current GNH survey. The requirements are:

one survey can have many questions;
one question can be used in many surveys;
one answer can be offered for many questions;
one question can have many answers offered; and one respondent can participate in many surveys.

2. Design of Proposed GNH Survey Application

2.1 Entity-relationship Model

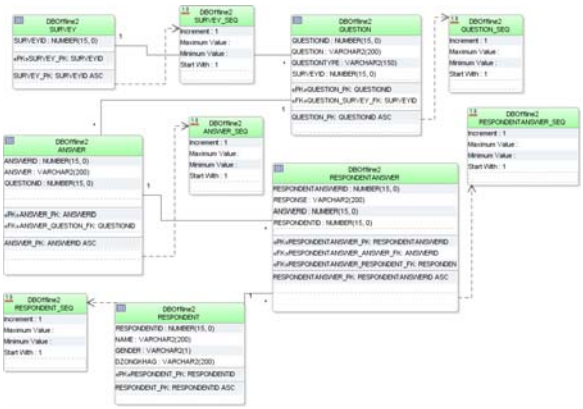

Fig. 1: GNH survey database diagram with sequence.

2.2 Business Services

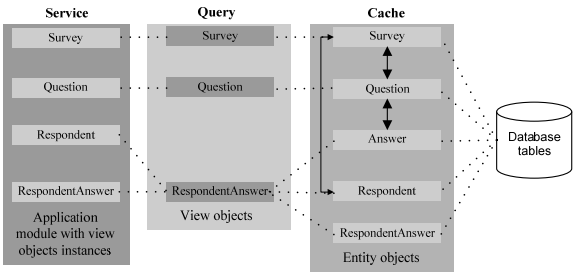

Fig. 2: Entity objects, view objects, and application modules of a business service.

The entity objects map database tables and act as an application cache for records from those tables. The view objects defines an application-specific view of records queried into the underlying entity objects. The application module, a collection of instances of view of objects define the data model and transaction for a particular business task.

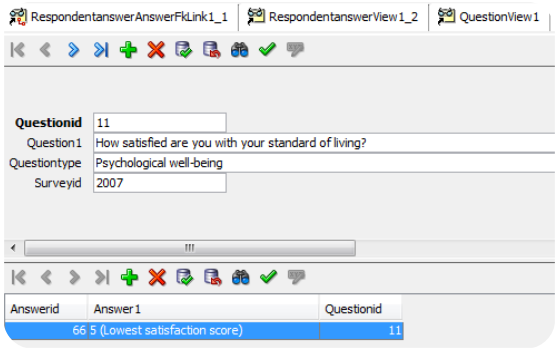

Fig. 3: A test result of business service.

2.3 User Interface

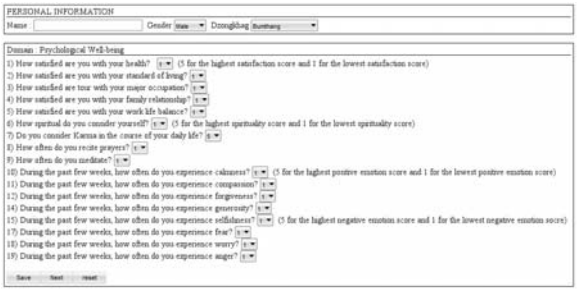

Fig. 4: A JavaServer Page (JSP) for psychological well-being.

Eight such JSP pages were created for all the remaining eight domains of GNH. These forms have the potential to collect a large amount of data in a relatively short span of time. Also, a survey doesn't have to worry about issues of missing or out-of–range responses.

2.4 Missing Data Values in a Database

Even though the data collection forms have been made automatic there have been missing data in a database. Missing data are unavoidable for reason like non-response and so on and so forth [1]. Researchers rely on a variety of ad hoc techniques to reconstruct the missing data by discarding incomplete cases or by filling in the missing values. Unfortunately, these techniques require strict assumptions about the cause of missing data, otherwise they are prone to substantial bias. These assumptions are known by the name of missing data mechanism.

3. Missing Data Mechanism

Rubin presented the standard definition of missing data mechanism which are classified into three categories, namely MCAR, MAR, and MNAR. Suppose (Y, R) is a data matrix with complete data; R being the missing data indicator matrix is defined as:

$$R = \begin{cases} 1, \text{ if } Y \text{ is observed} \\ 0, \text{ otherwise} \end{cases}$$

Y^o and Y^m are the respective observed and missing parts of Y.

An observation is said to be MCAR if the missingness is independent of all observed and unobserved values, hence:

$$P(R \mid Y) = P(R) \text{ for all } Y$$

$$R \text{ is independent of both } Y^{(o)} \text{ and } Y^{(m)}$$

An observation is said to be MAR if the missingness is independent of unobserved values but dependent on the observed values, therefore:

$$P(R \mid Y) = P(R \mid Y^{(o)}) \text{ for all } Y$$

$$R \text{ is independent of } Y^{(m)}$$

MNAR is a missingness mechanism that is neither MCAR nor MAR. It occurs when the missingness depends not only on observed values but also on the unobserved values, thus:

$$P(R \mid Y) \text{ depends on } Y^{(m)}$$

Based on this concept we have previously came up with a method to identify missing data mechanism.

3.1 Missing Data Mechanism Identification Method

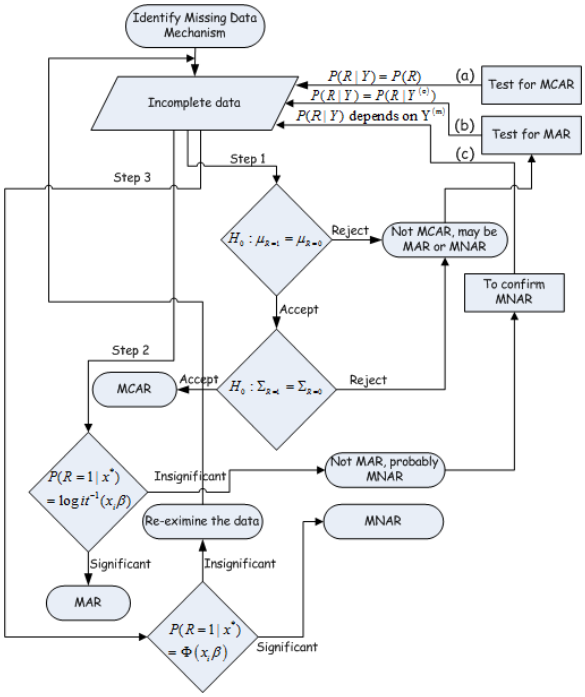

Fig. 5. A flow chart of missing data mechanism identification method.

The flowchart consists of three steps: One, in which an incomplete data is tested for MCAR by employing a hypothesis test to see if the means and covariances of the observed and unobserved groups are homogeneous or not. Acceptance of the means' null hypothesis leads to further testing of covariances' null hypothesis while the rejection terminates MCAR testing, and recommends performing MAR test. Acceptance of the null hypothesis of the covariance test confirms MCAR test. Second, following the result of MCAR test the incomplete data is then tested for MAR. A significant coefficient of logistic regression provides evidence in favour of MAR data while insignificant coefficients are by default MNAR data. This is so because MNAR is a missingness mechanism that is neither MCAR nor MAR. Additionally, a latent variable model is employed to reconfirm it.

We've coded the missing data identification method in R-Statistical language. And understandably, it takes certain programming skills to use R-language, hence the missing data identification method which has a wide application in social sciences remained inaccessible to many researchers. To make it accessible to wider researchers, we've built a web service for it.

4. Web Services

"A Web service is a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (WSDL). Other systems interact with the Web service in a manner prescribed by its description using SOAP-messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related Standards."

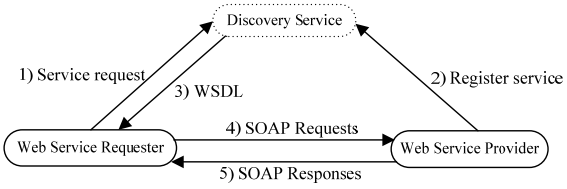

Fig. 6. Web services architecture with discovery service.

A requester sends a service request to a discovery service, the discovery service finds the available service provider which matches the request criteria, and in turn responds with the Web Service Definition Language (WSDL) associated with the provider. From the WSDL specification the requester understands the location of the provider and semantics for communication with provider and places Simple Object Access Protocol (SOAP) requests.

We've created a missing data mechanism identification web service from a Java class in Oracle Database 11g. Calling Oracle's DBMS_STAT_FUNCS, SQL statements for missingness mechanisms are defined as:

MCAR

```
SELECT respondent_id,
      AVG(answer_recode) group_mean,
      STATS_T_TEST_ONE(answer_recode 'STATISTICS')
      STATS_T_TEST_ONE(answer_recode,) two_sided_p_value
FROM answer;
SELECT respondent_id,
      COVAR_SAMP(missing_data, observed_data) AS covar_samp
FROM answer;
```

MAR and MNAR

```
SELECT respondent_id,
      REGR_R2(missing_data, observed_data) REGR_R2
FROM answer;
```

The web service is then created from this java class using the method of JAX-WS web service development. The web service is then tested on WebLogic server and later deployed to it.

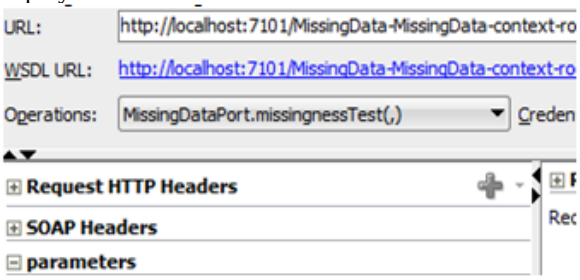

Fig. 7. Web service tested and deployed on WebLogic server.

Finally, we've used the WSDL URL in GNH survey application to utilize the web service in GNH application.

4.1NH Application and a Web Service

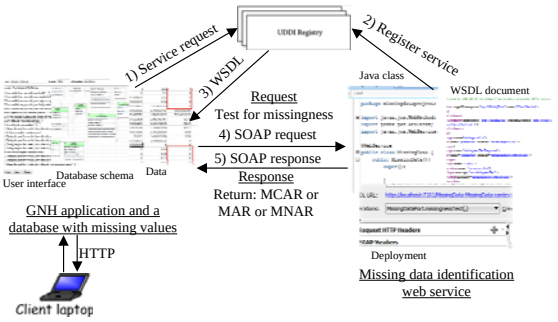

Fig. 8. The components and communication flow of GNH survey application and missing data identification web service.

A user's browser sends a HTTP request for GNH application and executes 'test missing data mechanism' command. Consequently, GNH application sends a service request to an UDDI registry, the UDDI registry finds the available service provider that matches the request criteria, and provides the WSDL associated with the provider to the GNH application. From the WSDL specification the requester understands the location of the provider and semantics for communication with the provider and places SOAP request and the provider responds with appropriate service.

5. Conclusion

While our proposed web service integrated GNH survey application has the advantages of identifying missing data mechanism, reducing the cost of carrying out survey, a wider reach within a short span of time, automatic validation, branching and skipping over the current method, it requires a robust backup system, a reliable and secure internet access, and technicians to administer the survey application. In the light of these shortcomings, an organization responsible for GNH data collection and analysis may consider building in-house technical capacity to handle the system, linkup with media to create awareness, come up with back-up policies, etc. for smooth operation of the system.

References

[1] "guidelines.pdf (application/pdf Object)." [Online]. Available: http://missingdata.lshtm.ac.uk/downloads/guidelines.pdf. [Accessed: 09-Jun-2012].

[2] G. Fitzmaurice, "Missing data: implications for analysis.," Nutrition (Burbank, Los Angeles County, Calif.), vol. 24, no. 2, pp. 200-2, Feb. 2008.

[3] "MissingDataFinal.pdf (application/pdf Object)." [Online]. Available: http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/MissingDataFinal.pdf. [Accessed: 09-Jul-2012].

[4] "Missing Data dij1461.pdf (application/pdf Object)." [Online]. Available: http://ferran.torres.name/download/material_comun/general/Missing Data dij1461.pdf. [Accessed: 19-Jun-2012].

[5] R. Schlittgen, "Analysis of incomplete multivariate data," Computational Statistics & Data Analysis, vol. 30, no. 4, pp. 478-479, Jun. 1999.

[6] "Roderick J. A. Little - Professor of Biostatistics: Missing Data." [Online]. Available: http://sitemaker.umich.edu/rlittle/missing_data. [Accessed: 09-Jun-2012].

[7] Business Survey Methods (Google eBook). John Wiley & Sons, 2011, p. 752.

[8] H. Toutenburg, "Little, R.J.A. and D.B. Rubin:Statistical analysis with missing data," Statistical Papers, vol. 32, no. 1, pp. 70-70, Dec. 1991.

[9] "aaps_schafer.pdf (application/pdf Object)." [Online]. Available: http://sites.stat.psu.edu/~jls/aaps_schafer.pdf. [Accessed: 09-Jun-2012].

[10] T. Erl, Service-Oriented Architecture.

[11] "Web Services Architecture." [Online]. Available: http://www.w3.org/TR/2004/NOTE-ws-arch-20040211/. [Accessed: 09-Jan-2012].

[12] R. Monson-Haefel, "J2EE Web Services," Oct. 2003.

[13] "The Standard Implementation for JAX-RPC — Java.net." [Online]. Available: http://java.net/projects/jax-rpc/. [Accessed: 02-Jun-2012].

[14] "code conventions for the java programming language: contents." [Online]. Available: http://www.oracle.com/technetwork/java/codeconvtoc-136057.html. [Accessed: 09-Aug-2012].

[15] [[Scott Guthrie]] and [[Scott Guthrie]], Releasing the Source Code for the NET Framework. 2007.

[16] "No Title." [Online]. Available: http://www.bhutanstudies.org.bt/. [Accessed: 05-Feb-2012].

Sonam Tshering is a PhD candidate at the University of the Ryukyus, Japan. He received Master's Degree in Information Engineering, Bachelor's Degree in Computer Application, and Diploma in Information Management Systems from the Ryukyus University, Japan; Vinayaka Missions University, India; and Royal Institute of Management, Bhutan respectively. He was working as a system administrator at Ministry of Labour and Human Resources, Bhutan prior to coming to Japan for studies. His research interests include handling missing data, design and analysis of QoL studies, constructing socio-economic indicators, causal modeling, web services and predictive analytics.

Takeo Okazaki took B.Sc., M.Sc. from Kyushu University in 1987 and 1989, respectively. He had been a research assistant at Kyushu University from 1989 to 1995. He has been a lecturer at University of the Ryukyus since 1995. His research interests are statistical data normalization for analysis, statistical causal relationship analysis.

Satoshi Endo took M.E., D.E. from Hokkaido University in 1990 and 1995, respectively. He had been a research assistant at Hokkaido University from 1990 to 1995. He has been a professor at University of the Ryukyus since 2005. His research interests are intelligent informatics and sensitivity informatics/soft computing.