Industrial Engineering
& Management Systems
Vol 11, No 4, December 2012, pp.371-384
ISSN 1598-7248 | EISSN 2234-6473 |
http://dx.doi.org/10.7232/iems.2012.11.4.371
© 2012 KIIE

Toward Successful Management of Vocational Rehabilitation Services for People with Disabilities: A Data Mining Approach

Yong Seog Kim*
Management Information Systems Department, Jon M. Huntsman School of Business,
Utah State University, Logan, UT, USA

(Received: October 23, 2012 / Revised: November 21, 2012 / Accepted: November 21, 2012)

ABSTRACT

This study proposes a multi-level data analysis approach to identify both superficial and latent relationships among variables in the data set obtained from a vocational rehabilitation (VR) services program of people with significant disabilities. At the first layer, data mining and statistical predictive models are used to extract the superficial relationships between dependent and independent variables. To supplement the findings and relationships from the analysis at the first layer, association rule mining algorithms at the second layer are employed to extract additional sets of interesting associative relationships among variables. Finally, nonlinear nonparametric canonical correlation analysis (NLCCA) along with clustering algorithm is employed to identify latent nonlinear relationships. Experimental outputs validate the usefulness of the proposed approach. In particular, the identified latent relationship indicates that disability types (i.e., physical and mental) and severity (i.e., severe, most severe, not severe) have a significant impact on the levels of self-esteem and self-confidence of people with disabilities. The identified superficial and latent relationships can be used to train education program designers and policy developers to maximize the outcomes of VR training programs.

Keywords: Vocational Rehabilitation Services, Data Mining, Classification, Association Rule, Clustering, NLCCA

* Corresponding Author, E-mail: yong.kim@usu.edu

1. INTRODUCTION

The vocational rehabilitation (VR) service program is a sponsored program with support from Congress, U.S. Department of Education, and state and local governments to help people with physical, mental, or emotional disabilities for vocational evaluation, counseling, or job placement assistance. People with disabilities in VR service programs are encouraged to develop individualized rehabilitation plans with the help of counselors who provide job market information and find available training and education programs. Ultimately, the VR program is oriented to help people with disabilities in VR service programs accomplish the employment outcome by entering or retaining full-time employment or any other types of employment. Typically, the VR program in each state is operated in compliance with the federal Rehabilitation Act of 1973, and most people with disabilities are eligible to apply for this program regardless of gender, age, race, or type of disability. Note that because people with disabilities often suffer from severe vocational impediments due to their physical or mental conditions or prejudice from people without disabilities, VR services must be delivered effectively and efficiently toward VR program trainees to overcome these impediments. However, not all people

with disabilities accomplish desired employment outcome even after they complete VR service programs. Therefore, it is an interesting research question to ask what individual characteristics and VR service program factors affect the effectiveness of VR service programs and result in different employment outcome.

Nevertheless, not many studies have been conducted to provide a framework for the use of intelligent systems with managerial and organizational implications for people with disabilities. Therefore, the immediate objecttive of this research is to encourage practitioners and researchers to collaborate, interact, and exchange research ideas about the development of services and education programs for people with disabilities. Note that the design and development of services is strongly dependent on the types, knowledge, and skill levels of the users. As an initial step toward this new direction, we analyze the data sets obtained from VR services program of people with significant disabilities so that the findings from our research can be used for training education program designers and policy developers who can efficiently and effectively manage VR programs and maximize the outcomes of VR training programs.

Overall, this study takes a multi-level analysis approach to better understand VR data set and ultimately provide education program designers and policy developers with insights on superficial and latent relationships among various factors that can be used to maximize the outcomes of VR training programs. One of the most popular VR programs for people with disabilities is to provide an on-site job training including IT trainings and other ongoing supports. Note, however, that not all trainees can find and keep their jobs. Therefore, the immediate objective of our first level analysis is to develop predictive models that can accurately predict or profile VR trainees who are most likely to secure a job after completing a VR service program. Typically, the well known classifiers from data mining community and statistical predictive models from statistics and mathematics community can be adopted for this purpose.

At the second level, we pay more attention to managerial insights that can be obtained from other descriptive rather than predictive models based on the fact that the most predictive model is already obtained at the first level. For example, administrative managers and education program developers would like to know how VR outcomes would change when they change or control a certain set of VR or demographic related instruments. For this purpose, we employ association analysis models which have been successfully applied in marketing and e-commerce community.

Finally, at the third level, we like to focus on identifying the latent relationships between self-perception (e.g., self-confidence and self-esteem) and physical and psychological hindrance factors, which in turn affect social activities and presumably the outcomes of VR services program. Numerous studies have reported that children and adolescents with a physical disability are very likely to be socially isolated (Thomas *et al*., 1989), and maintain the low levels of self-confidence and self-esteem in their adulthood, resulting in physical and psychological hindrance on social relationships and outcomes of VR services program. To discover these latent relationships, we segment the participants based on their perceptions of themselves (i.e., self-esteem and self-confidence level, and the subjective weights they assign to physical and psychological hindrance factors on their social activities) and relate the segmentation characteristics to personal factors (i.e., gender and marriage status) and disability-related factors (i.e., disability type and severity). As a tool for this kind of analysis, *k*-means clustering and nonlinear nonparametric canonical correlation analysis (NLCCA) are chosen. Ultimately, findings from this research will help us better understand social cognitive factors and personal characteristics of people with disabilities and their effects on the education outcomes, which, in turn, many IT educators can utilize to develop and maximize the outcomes of IT-intensive education programs.

The remainder of this paper is organized as follows. Section 2 presents our research model and VR data set in detail. In Section 3, the predictive performance of single classifiers and ensemble classifiers are compared and discussed. Section 4 briefly introduces association rule algorithms, and outputs of two different association rule algorithms are presented. Section 5 starts with a brief introduction of clustering and NLCCA. Then the characteristics of clustering outputs based on social cognitive factors are interpreted. After that, managerial implications and the relationships among multiple sets of key variables based on NLCCA are presented and discussed. Finally, we conclude the paper with some suggestions for possible future research in Section 6.

2. RESEARCH METHODOLOGY AND DATA DESCRIPTION

2.1 Research Methodology

As shown in Figure 1, this study employs a multi-level approach to fully identify superficial and latent relationships among variables in VR data sets. The main task in the first level analysis is classification, identifying the relationships between dependent and independent variables for prediction purpose. We consider the identified relationships between dependent and independent variables “superficial” because in-depth understanding on possibly complex and hidden relationships among variables is not required as long as the identified relationships are useful for predicting the value of the dependent variable. The classification task considered in this study is to accurately predict and profile VR trainees who are most likely to find and keep a job after completing a VR service program. Then the resulting

predictive model can be used by developers and administrators of VR training programs to estimate the success rate in advance for a new pool of VR trainees and pay more attention to trainees during the training program who are less likely to find and keep jobs. Furthermore, program administrators may consider offering different VR service programs in terms of program contents and length depending on VR training applicants' likelihood of finding a job. Any predictive classifiers and ensembles (i.e., an ensemble is a classifier that combines multiple classifiers) from data mining and artificial intelligence community, and statistical predictive models can be used for this purpose. These methods can be evaluated in terms of several evaluation criteria such as predictive accuracy, computational complexity, and performance robustness.

At the second level, managerial insights that can be extracted from the predictive and other descriptive models become more important, assuming that the most predictive model is already obtained at the first level. For example, administrative managers, VR program developers, and state and federal officers may want to know what factors (i.e., VR program related or demographic related) are influential on post VR employment status. In addition, a new analysis at the second level can provide important (associative) relationships among independent variables. With additional relationships and insights, VR service program managers and developers may consider changing or controlling a certain set of instruments toward better VR outcomes. For this purpose, a new type of analysis tools, association algorithm, is introduced and applied. Note that various association algorithms and variants have been actively studied and developed to analyze the associative relationships among commodities in consumers' market baskets in marketing community (Silverstein *et al*., 1998; Brijs *et al*., 1999) and among navigation patterns of Web surfers in e-commerce community (Anandarajan, 2002; Spiliopoulou *et al*., 2000; Mobasher *et al*., 2002).

Finally, at the third level, the main objective of the analysis is to identify the relationships between psychological and societal characteristics of VR trainees and external outcomes of VR training. For analysis at this level, clustering analysis and NLCCA are employed. NLCCA is a form of canonical correlation analysis in which categorical variables are optimally scaled as an integral component in finding linear combinations of variables with the highest correlations between them. In our analysis, clustering analysis with a non-hierarchical algorithm (e.g., *k*-means) is used to segment participants into seven clusters and find out meaningful characteristics. Then, NLCCA is conducted with three sets of variables: a seven-category segmentation variable from clustering analysis, disability characteristics (disability type and severity), and personal variables (gender and marriage status).

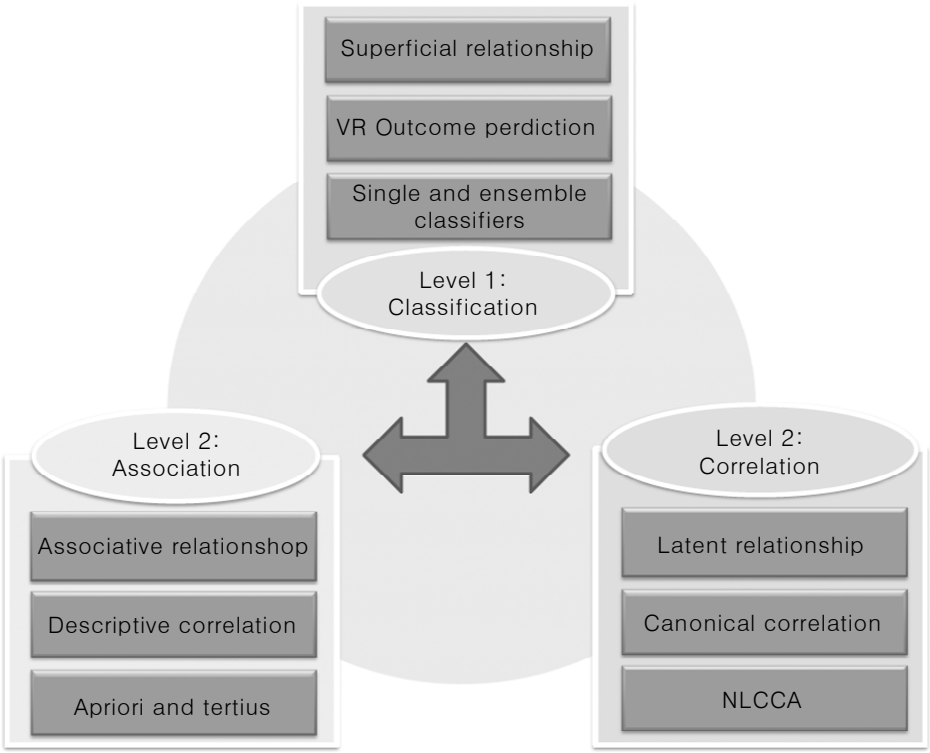

Figure 1. Research model. VR: vocational rehabilitation, NLCCA: nonlinear nonparametric canonical correlation analysis.

2.2 Data Description

We started with the data set from the Longitudinal Study of the Vocational Rehabilitation Services Program (LSVRSP). This data set is publicly available at http://www.ilr.cornell.edu/edi/lsvrsp/ and contains a total of 8,818 records with 951 input variables from eight data sets: status file (STATUS), demographics and disability characteristics (CDF1), vocational interests and goals (CDF2), quality of services factors (CDF3), applicant/client function interview (CF1), applicant work history interview (WH1), and two follow-up interviews (F126 and FU).

The input variables selected for our study include demographic information (e.g., age, gender, race, marriage status), disability related variables (e.g., type and severity), trainees' perceived importance of physical and psychological hindrance factors, self-esteem and self-confidence on their social activities, and post VR employment status. We selected self-esteem and self-confidence related variables mainly based on social cognitive theory (SCT) (Bandura, 1986). According to SCT, one's behavior is constantly influenced by cognitive and environmental influences, and self-esteem and self-confidence are one of the most important factors that affect the performance of academic learning and vocational services (Zimmerman *et al*., 1992; Edwards and Hardy, 1996). For example, it has been shown that people with high self-esteem are more likely to work on a task if they believe they can succeed (Vogt *et al*., 2007), and maintain healthy interpersonal interactions and prosocial behavior such as helping others, sharing, being kind and cooperative (Bandura *et al*., 2003). However, it has been also known that self-esteem is related to beliefs about appearance, and hence it is important to observe the level of self-esteem for a person with disabilities. Self-confidence is "derived upon judgments of one's capabilities to organize and execute courses of action to attain specific goals" (Bandura, 1986) and hence negative comments and discouraging words negatively affect the self-confidence of individuals, which eventually deteriorates their performance (Cox, 1998). Therefore, we conjectured that self-confident people might perform better than people with lack of self-confidence in VR service program because their positive attitude has them work harder to accomplish their success (i.e., obtaining job) through VR service program and maintain it by building healthy social relationships.

We also identified the primary eight disability types based on Cornell's recoding: orthopedic including amputation, mental illness, non-orthopedic physical, mental retardation, hearing, vision impairment, substance abuse, and traumatic brain injury. In fact, one more disability type, learning disability, was identified but there was no matching record after we removed all records with missing values. The final data set includes 1,895 records with employment outcome and 1,200 records without employment outcome along with 15 input variables. Further information regarding the LSVRSP including data dictionaries and user's guide can be found at http://www.LSVRSP.org.

3. LEVEL 1 ANALYSIS: CLASSIFICATION WITH SINGLE AND ENSEMBLE CLASSIFIERS

In our first experiment, we tested whether or not many well known data mining algorithms can successfully predict post VR employment status based on trainees' self-esteem, self-confidence, and physical and psychological hindrance factors on their social activities. Using Weka, a free data mining tool (http://www.cs.waikato.ac.nz/ml/weka/), four well-known classifiers: ZeroR, logistic regression, artificial neural network (ANN), and decision tree algorithm (C4.5) were implemented. The ZeroR classifier in this study was included to serve as a basis algorithm because of its simple classification rule, predicting all observations as points in the majority class. The logistic regression is one of the most popular statistical classifiers. The ANNs is a nonlinear classifier that has been known to be robust and accurate, but it is difficult to understand classification rules from ANNs because of its black-box algorithm characteristics and structural complexities with many subjective parameter settings. Unlike ANNs, the C4.5 is relatively free from subjective parameter setting, and it is faster and provides much more interpretable decision rules while providing a comparable performance with ANNs. In our implementations of these algorithms, we used all default settings in Weka for easy replications of our results except the number of hidden layers (which was set to three to replicate the most popular ANN structure) in an ANN. To fairly evaluate these algorithms, we took a 10-fold cross validation scheme in which the entire data set is divided into 10 equal size blocks and each block is in turn used as a test set while the classifier is built on the remaining blocks. We summarized the performance of these classifiers in Table 1.

The predictive accuracy of ZeroR was expected to be 61.21% because the records with the majority class, trainees with a job after VR program, consist of 61.21% of the data set (1,895 trainees out of a total 3,096 trainees). The accuracy of ANN model (80.07%) was acceptable considering the fact that we did not try to find the best performing parameter values such as the number of epoch (training time), learning rate, momentum rate, and most of all, the number of hidden layers. The accuracy of a Logistic regression model (82.39%) was significantly better than ZeroR, but only slightly better than ANN. The best performance was recorded by C4.5 with an 83.49% of accuracy. We also observed that the ZeroR was the fastest, followed by C4.5, logistic regression, and ANN. Based on predictive accuracy, speed, and easy interpretability, C4.5 was chosen to be the best (we will discuss in detail about the interpretation of

C4.5 tree model in the following sections).

Table 1. Summary of single classifier performance

Classifiers	ZeroR	Logistic	ANN	C4.5
Accuracy (%)	61.21	82.39	81.07	83.49
Speed (sec)	<0.01	0.41	5.75	0.1
Interpretability	Good	Good	Bad	Good

In addition to single classifiers, ensemble classifiers (or meta-classifiers) that combine multiple classifiers were also tested to see if the performance of a single classifier can be improved. Bagging (Breiman, 1996) and boosting (Freund and Schapire, 1996) are the most popular methods for creating a meta-classifier. Bagging builds each component classifier on a randomly drawn data set from the original data set, and combines the prediction of multiple classifiers with equal weight for the final prediction. Boosting produces a series of classifiers, with each data set based on the performance of the previous classifiers so that new classifiers are constructed to better predict observations for which the current meta-classifier's performance is poor. This is accomplished using adaptive resampling (i.e., observations that are incorrectly predicted by previous classifiers are sampled more frequently) and AdaBoost weights the predictions based on the classifiers' training error. Since C4.5 was the best single classifier in aforementioned experiment, we combined 25 C4.5 tree classifiers to form the AdaBoost and bagging ensembles, respectively. To our surprise, the performance of the AdaBoost and bagging models (78.75% and 83.46%) was not significantly better than that of single C4.5 tree, while they took almost 25 times longer to build an ensemble prediction model than a single C4.5 tree. Note that in many cases, it is possible for ensemble models to significantly perform better than single classifier by promoting diversity through multiple classifiers trained on different parts of records. However, the performance of AdaBoost has been known to be unstable, which can explain a relatively low performance on VR data set. Although it is possible to further improve the prediction performance of ensemble models, we leave it to other researchers because fine tuning the performance of classifiers is not our main objective.

4. LEVEL 2 ANALYSIS: ASSOCIATION ANALYSIS

4.1 Preliminaries of Association Rule

A simple introduction to association rule is necessary. Typically, an association rule, R_i, is represented in the form [A => B] where each A and B represents an itemset (e.g., a set of products) in a transaction record where $A \cap B = \varnothing$. For convenience, we refer to A and B as the assumption (or antecedent) and the consequent of the rule, respectively. In addition, we denote D as a set of transactions, while ||D|| and count (A) denote the number of transactions in D and the number of transactions containing A, respectively. Then, the support and confidence of R_i is defined as count $(A \cup B)/||D||$ and count$(A \cup B)$/count(A), respectively. Note that the support of R_i measures the probability of observing the antecedent and the consequent together out of entire transaction records, while the confidence of R_i measures the conditional probability of the consequent (B) given the antecedent (A). Intuitively, the higher the support of the rule, the more prevalent the rule is, and the higher the confidence of the rule, the more reliable the rule is (Brijs *et al*., 1999). When rules are different in both support and confidence and have to be compared, a predictive accuracy measure can be used in which a larger support is traded against a higher confidence to maximize the expected accuracy (Scheffer, 2005). Ultimately, the main objective of association analysis is to generate all the association rules that have support and confidence greater than the user-specified minimum support and minimum confidence. Readers who are interested in additional metrics to measure the usefulness of association rules are referred to (Aumann and Lindell, 2003; Padmanabhan and Tuzhilin, 1999; Silberschatz and Tuzhilin, 1996; Kim, 2009).

To illustrate the usefulness of association analysis, we presented a C4.5 tree (74 leaves and 28 nodes) obtained from the analysis at the first level on the VR data set in Figure 2. Note that a rule-based classifier (e.g., C4.5) that consists of *if–then* rules improves comprehensibility and boost managers' trust in the classifier itself. According to the tree structure shown in Figure 2, the most important variable for predicting post VR employment status was the self-esteem, which is located at the root node. Although *if–then* decision rules from a C4.5 decision tree classifier can be very useful, they become very complex as a decision tree grows big with many nodes and leaves as shown in Figure 2. In fact, Figure 2 shows only a part of the whole decision tree structure due to its complex structure. Further, these decision rules are useful mainly for classification purposes, and provide little information about relationships among variables and how their relationships can be linked to post VR employment status or other variables, encouraging the usage of another type of data analysis model.

4.2 Association Rule Analysis

In Weka, three different types of association algorithms are available: Apriori (Agrawal and Srikant, 1994), Predictive Apriori (Scheffer, 2005), and Tertius (Flach and Lachiche, 2001). Note that it is not our main goal to compare all these algorithms for prediction accuracy as in (Mazid *et al*., 2008), but rather to extract associative relationships among variables in VR data set as many as possible. Note that Apriori and Predictive Apriori are

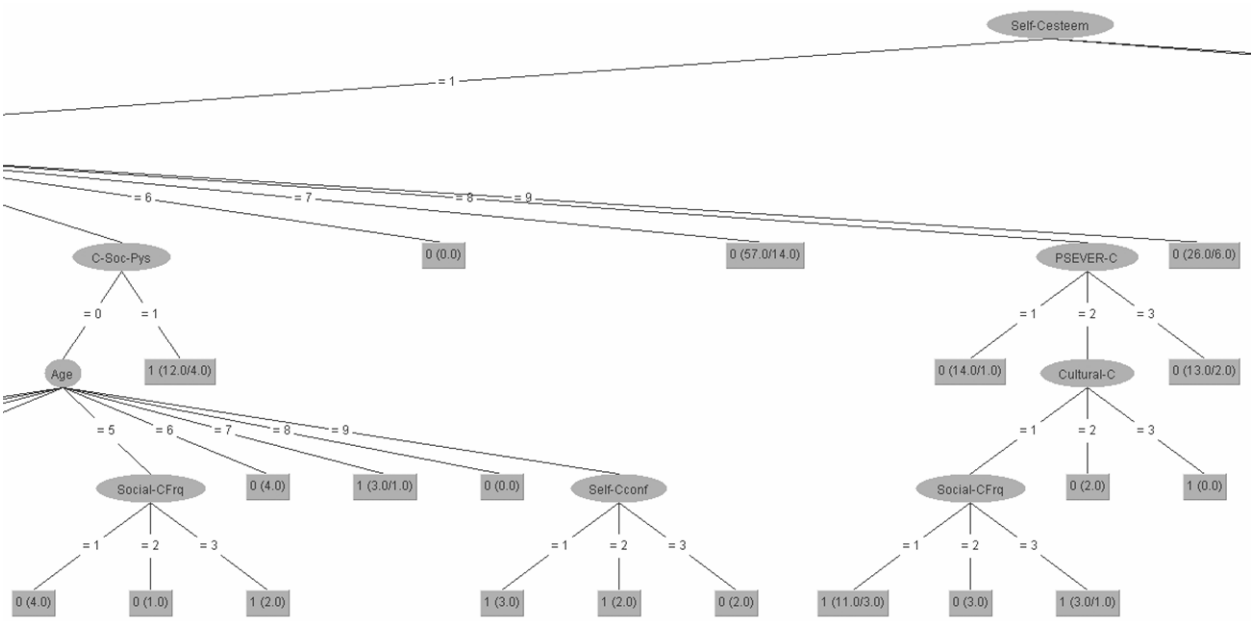

Figure 2. Partial snapshot of a C4.5 decision tree structure.

very similar with very comparable predictive power (Mazid *et al*., 2008), and hence we only present the output of Apriori. We also show the output of Tertius mainly because its association rules are more descriptive than predictive, satisfying the needs of our analysis at the second level. We ran both algorithms with the default setting in Weka, and show their outputs in Tables 2 and 3, respectively.

The rules shown in Table 2 were subjectively chosen out of 1,000 rules that satisfied the minimum support and confidence criteria to present only rules with different implications. Let's take the first rule, A1, which means that trainees who believe their disability does not prevent them from socializing with friends outside (Social = 2) also believe that their physical, sight, and psychological impairments are not critical factors for their social activity (C-Soc-Pys = 0, C-Soc-See = 0, C-Soc-Psy = 0). Note that the numbers in A1 indicate the number of records in VR data set that satisfy the assumption (1,925 trainees with Social = 2) and consequent (1,925 trainees with C-Soc-Pys = 0, C-Soc-See = 0, C-Soc-Psy = 0) of A1, respectively. We found that most rules shown in Table 2 include the same consequent, referring to the trainees who feel their physical, sight, and psychological impairments are not important for their social activity (C-Soc-Pys = 0, C-Soc-See = 0,

Table 2. Association rules among socio-psychographic variables

Rule index	Sup. (%)	Conf. (%)	Association rules		
			Antecedent	=>	Consequent
A1	62.18	100	Social = 2	=>	C-Soc-Pys = 0 and C-Soc-See = 0 and C-Soc-Psy = 0
A20	52.78	100	Social = 2 and Cultural-C = 1	=>	
A39	62.18	84.52	Race = 1 and Social = 2	=>	
A58	44.70	100	Social = 2 and Social-SMeeting = 2	=>	
A191	33.66	100	Social = 2, Cultural-C = 1, and I26 = 1	=>	
A286	30.49	100	Social = 2 and PSEVER-C = 2	=>	
A305	29.30	100	Gender = 2 and Social = 2	=>	
A324	29.20	100	Marriage = 5 and Social = 2	=>	
A554	24.74	100	Social = 2 and Self-Cesteem = 3	=>	
A573	24.71	100	Social = 2, Self-Cesteem = 3, and I26 = 1	=>	
A918	21.19	100	Gender = 2, Race=1, Social = 2, and Cultural-C = 1	=>	
A975	20.67	100	Social = 2, Cultural-C = 1, and Self-Cesteem = 1	=>	
A994	20.54	100	Social = 2, Cultural-C = 1, and Self-Cesteem = 3	=>	

C-Soc-Psy = 0). According to rules, A20, A39, and A58, these trainees are those who keep social activities even with their disability (Social = 2), and do not bother with cultural background (A20; Cultural-C = 1), or who are Caucasian (A39; Race = 1), or who ever participated in social meeting designed for people with disability (A58; Social-SMeeting = 2).

Similarly, A191, A286, A305, and A324 also describe trainees who consider physical, sight, and psychological impairments are a minor factor for their social activity. These people are who do not bother their social activities because of their disability while they are either people who think cultural factor is not very import and secure post VR employment (A191; Social = 2, Cultural-C = 1, and I26 = 1); or who are severely, but not most severely, disabled (A286; Social = 2 and PSEVER-C = 2); or either female or never married trainees who do not bother their social activities because of their disability (R305, Social = 2 and Gender = 2; R324, Marriage = 5 and Social = 2). Two rules (A554 and A573) are about 765 trainees who maintain their social activity (Social = 2) and regard self-esteem as very important on social activity (Self-Cesteem = 3). Whether these trainees have a job after VR training (A554) or not (A573), they expect their physical, sight, and psychological impairments to have only marginal effect on their social activity (C-Soc-Pys = 0, C-Soc-See = 0, C-Soc-Psy = 0). We leave the interpretations of other rules to the readers.

The outputs of another association rule algorithm, Tertius, are more intuitive because it is related to the post VR employment index. For example, the first rule (T1) indicates that trainees who believe that self-esteem is very important for their social activity are likely to find a job after VR training (confidence, 60%; support, <1%) while trainees who believe self-esteem is not important for their social activity are less likely to have a job after VR training (T2; confidence, 59.88%; support, 10.85%). Other rules specify that trainees who consider self-esteem important but physical (T5) or psychological impairments (T6) are not important are likely to obtain post VR employment (confidence 51% and 50%, respectively), while those who consider self-esteem, cultural factors, and sight impairment unimportant for their social activity (T7) are not likely to find a job after VR training.

Clearly, the outputs of association rule algorithms at the second level are different from the outputs of classifiers at the first level in the sense that they are more descriptive and provide additional insights for VR program administrators and education program developers to better understand trainees with disability or maximize the outcomes of VR service programs.

5. LEVEL 3 ANALYSIS: CLUSTERING AND NLCCA

Although the outputs from association rule algorithms provide additional managerial insights, they are still limited in the sense that they do not provide insights on unobserved (or latent) relationships among variables. For example, many researchers want to know whether or not (and if so, how far) strong self-confidence can create positive will power, encouraging people with disabilities to overcome their psychological and physical hindrances to accomplish the goals of activities. In addition, it is interesting to understand that, through which psychological mechanism, respondents' physical disability and psychological disturbance affect their social activities. To gain such insights, we need to utilize another layer of analysis and we suggest using clustering and NLCCA methods for this purpose.

5.1 Clustering Analysis

Formally, clustering is defined as the process of partitioning transaction records of customers into a fixed number of groups (or clusters) based on heuristic metrics such as intra-cluster compactness (how similar the elements of each cluster are) and inter-cluster separability (how dissimilar the clusters are) (Jain *et al*., 1999). In

Table 3. Association rules for vocational rehabilitation employment status

Rule index	Sup. (%)	Conf. (%)	Association rules		
			Antecedent	=>	Consequent
T1	60.00	0.1	Self-Cesteem = 3	=>	I26 = 1
T2	59.88	10.85	Self-Cesteem = 1	=>	I26 = 0
T3	55.76	9.46	Self-Cesteem = 1 and C-Soc-See = 0	=>	I26 = 0
T4	52.91	9.69	Cultural-C = 1 and Self-Cesteem = 1	=>	I26 = 0
T5	51.33	0.06	Self-Cesteem = 3 and C-Soc-Pys = 0	=>	I26 = 1
T6	50.20	0.03	Self-Cesteem = 3 and C-Soc-Psy = 0	=>	I26 = 1
T7	49.57	8.33	Cultural-C = 1, Self-Cesteem = 1, and C-Soc-See = 0	=>	I26 = 0
T8	47.66	9.11	Race = 1 and Self-Cesteem = 1	=>	I26 = 0
T9	47.49	6.27	Social-SMeeting = 2 and Self-Cesteem = 1	=>	I26 = 0
T10	46.75	7.82	Self-Cesteem = 1 and C-Soc-Pys = 0	=>	I26 = 0

general, clustering algorithms try to find a clustering where the elements within the same cluster are as similar as possible and the clusters are as dissimilar as possible. In this study, cluster analyses were performed in the space of the four social-activity hindrance and encouragement factors using a *k*-means clustering algorithm to find homogenous groups. The first factor, self-esteem, is a social-activity encouragement variable, and it is expected that a person with higher self-esteem is more inclined to having social activity and hence the outcomes of vocational training will be positive. The second factor, self-confidence, is another social-activity encouragement variable and hence is expected to be positively related to social activity and the outcomes of vocational training. In particular, strong self-confidence can create positive will power that people with disabilities overcome many psychological and even physical hindrances to accomplish the goals of activities. The two remaining factors, social-physical and social-psychological factors, reflect respondents' own judgments on how significantly their physical disability and psychological disturbance affect their social activity.

After trying several clustering analyses with between 5 and 10 clusters, we found that *k*-means with 7 clusters satisfy our subjective criterion, at least 100 records and at most 1,000 records in each cluster to draw reliable cluster characteristics. Based on cluster centers, we drew qualitative characteristics of each segment and summarized them in Table 4. One encouraging observation is that trainees with disabilities in the largest segment (C4, 28.9%) maintain a very high level of self-esteem and self-confidence, and they believe that their physical disabilities and psychological hindrance factors do not significantly affect their social activity. In addition, trainees in C5 (13.7%) also do not regard their physical disabilities and psychological disturbances as a significant factor to limit their social activity. However, trainees in segment C6 (3.6%) and C2 (13.5%) maintain a low level of self-esteem and self-confidence, and keep limited social activity because of physical and psychological factors originated from their disabilities. Other trainees (C1, 9.6%) have a high level of self-esteem and self-confidence, but maintain limited social activity due to their physical and psychological reasons related to their disability. Finally, the remaining trainees in C3 and C7 (30.7%) maintain a low level of self-esteem and self-confidence, but are actively involved in social activity.

5.2 NLCCA Model Introduction

To introduce NLCCA, we first briefly review a conventional linear canonical correlation analysis (CCA). The main objective of CCA is to find linear combinations of the variables in each set, a set of dependent variables and a set of independent variables, so that the canonical correlation between the linear combinations is maximized. That is, CCA finds and extracts the linear combination of independent variables that produces maximum correlation with the dependent variables. Then the process is repeated for the residual data, with the constraint that the second linear combination (defined in a new dimension) of variables must not correlate with the first one. The process is repeated until a successive linear combination is no longer significant. The linear combinations are commonly called canonical variates and all canonical variates are mutually orthogonal (i.e., independent). Once we obtain canonical variates, we can assess how strongly each of them is related to measured variables in its own set, the set for the other canonical variate, or how much percent of variance in the dependent set is explained by the independent set of variables along a given dimension.

NLCCA corresponds to categorical CCA with optimal scaling. By "categorical" CCA, we mean that NLCCA is designed to explain the relation between multiple variable sets that include categorical or ordinal (nonlin-

Table 4. Cluster characteristics

Cluster	Social activity hindrance and encouragement factors	Cases	
		N	%
C1	**Very high motivation and very high hindrance:** Maintain very high level of self-esteem and self-confidence. Believe that physical and psychological factors are significant for their social activity.	298	9.6
C2	**Low motivation and very high hindrance:** Maintain very low level of self-esteem and somewhat low self-confidence. Believe that physical and psychological factors are significant for their social activity.	417	13.5
C3	**Very low motivation and very low hindrance:** Maintain very low level of self-esteem and self-confidence. Believe physical and psychological factors are not significant for their social activity.	612	19.8
C4	**Very high motivation and very low hindrance:** Maintain very high level of self-esteem and self-confidence. Believe physical and psychological factors are not significant for their social activity.	895	28.9
C5	**Average motivation and very low hindrance:** Maintain a median level of self-esteem and self-confidence. Believe physical and psychological factors are not significant for their social activity.	424	13.7
C6	**Low motivation and very high hindrance:** Maintain somewhat low level of self-esteem and very low level of self-confidence. Believe physical and psychological factors are significant for their social activity.	110	3.6
C7	**Low motivation and very low hindrance:** Maintain very low level of self-esteem and somewhat low self-confidence. Believe physical and psychological factors are not significant for their social activity.	339	10.9

ear or nonparametric) variables. By categorical CCA with "optimal scaling," we mean that, in the process of finding the best-fitting model, NLCCA simultaneously search for both optimal re-scaling of the categories of all variables and corresponding weights (i.e., component loadings) in such a way that a canonical variate from weighted re-scaled variables in one set has the maximum possible correlation with another canonical variate from weighted re-scaled variables in the second set.

We utilized OVERALS procedure available in categories module of SPSS to conduct NLCCA. In OVERALS, categorical variables are quantified using optimal scaling and treated as numerical variables. For nominal variables, OVERALS creates values for each category while ignoring the order of the categories so that the goodness-of-fit is maximized. For ordinal and interval variables, OVERALS retains the order of the categories while creating values to maximize the goodness-of-fit of a model. OVERALS output includes several measures of goodness-of-fit, component loadings, optimal category scores, and plots including component loadings plots, category centroids plots, and transformation plots. Component loadings in NLCCA and factor loadings in principal component analysis (PCA) are similar in the sense that they represent correlations between the optimally scaled variables and the canonical variates. Therefore, we can infer how much of the variable was explained by the canonical variates in total by computing the sum of squared loadings, the distance between the origin and the component loadings of a given variable in the orthogonal space of the canonical variates (ter Braak, 1990). It is also possible to estimate the contribution of a specific canonical variate by computing the square of the projections onto it.

5.3 NLCCA Model Specification

First, a non-hierarchical clustering algorithm (e.g., *k*-means) is used to segment participants into seven clusters and find out meaningful characteristics. Then, NLCCA is conducted with three sets variables: a seven-category segmentation variable from clustering analysis, disability characteristics (disability type and severity), and personal variables (gender and marriage status). While the first two analyses are mainly for data analysts, the last analysis at the lowest level is mainly for education program designers and policy developers who like to maximize the outcomes of their training programs.

To explain how strongly the disability-related and personal characteristics affect the level of social activities of trainees with disabilities, an NLCCA model was specified with three sets of variables as shown in Table 5. Note that we cannot use CCA because all of the variables are categorical and we expect some relationships to be nonlinear. For example, one set of variables in our study is comprised of a multiple categorical variable defining clusters, while the other set of variables is composed of categorical variables describing disability char-

Table 5. Variables in nonlinear canonical correlation analysis

Set	Variable	Type	Categories	Cases	
				N	%
1	Cluster indexes based on social activity hindrance and encouragement factors	Multiple nominal	Refer to Table 4	Refer to Table 4	
2	Disability type	Single nominal	1 = Orthopedic including amputation	876	28.3
			2 = Mental illness	633	20.5
			3 = Non-orthopedic physical	586	18.9
			4 = Mental retardation	274	8.9
			5 = Hearing	273	8.9
			6 = Vision impairment	256	8.3
			7 = Substance abuse	142	4.6
			8 = Traumatic brain injury	55	1.7
	Severity of disability	Single nominal	1 = Most severely disabled	604	19.5
			2 = Severely disabled	1,614	52.1
			3 = Not severely disabled	877	28.4
3	Gender	Single nominal	1 = male	1,585	51.2
			2 = female	1,510	48.8
	Marriage status	Single nominal	1 = Married	954	30.8
			2 = Widowed	158	5.1
			3 = Divorced	508	16.4
			4 = Separated	155	5.0
			5 = Never married	1,320	42.7
	Race	Single nominal	1 = White	2,600	84.0
			2 = Black	436	14.1
			3 = American Indian/Alaskan Native	23	0.7
			4 = Asian or Pacific Islander	36	1.2

acteristics (e.g., disability type and severity) and social activity hindrance factor (e.g., physical and psychological hindrance). The first variable set consists of the seven-category segmentation variable based on social activity hindrance and encouragement factors. In our analysis, this variable was considered as ''multiple nominal'' having different optimal category quantifications for each canonical dimension (i.e., different contribution to the canonical variates). The second set consists of two disability characteristics (disability type and severity of disability), while two personal demographic variables (gender and marriage status) were assigned to the third set. Each variable in the second and third sets was considered as "single nominal" with a single optimal quantification for all canonical dimensions.

5.4 NLCCA Model Fit

A two-dimensional NLCCA solution was chosen, and Table 6 shows the overall fit (= eigenvalues) of this two-dimensional solution in terms of the variance accounted for within each set of variables by each of the two dimensions (canonical variates). Note that the maximum fit value equals the number of dimension, indicating the perfect relationship. The overall fit of our model was 0.893, a sum of two eigenvalues from the first variate (0.528) and the second variate (0.365). Therefore, 0.528/0.893 = 59.1% of the actual relationship among each set of variables is explained by the first dimension. The canonical correlation, a measure of the correlations among the three sets of variables, for each of the canonical variates can also be computed from eigenvalues as follows:

$$\rho_d = \frac{((K \times E_d) - 1)}{K - 1}$$

where d is the dimension number, K is the number of sets, and E is the eigenvlaue. Using $d = 2$ and $K = 3$, we obtained the canonical correlation for each dimension, 0.292 and 0.048, respectively, and hence the first dimension is approximately 6 times more effective than the second at capturing the relationships among the three sets.

The weights defining the two dimensions in terms of the optimally scaled variables are also shown in Table 6. Note that weights for multiple nominal variables (i.e., the segmentation variable in set one) are not unique and hence are not presented. According to Table 6, in terms of the variables of sets 2 and 3, the first canonical variate primarily relates disability type to marriage status in set 3, while the second variate mainly relates the severity of disability to gender. Note that this is consistent with the findings in Clark *et al*. (1977) that boys (gender) with (severe) physical disabilities are more likely to have difficulties in interpersonal relationships.

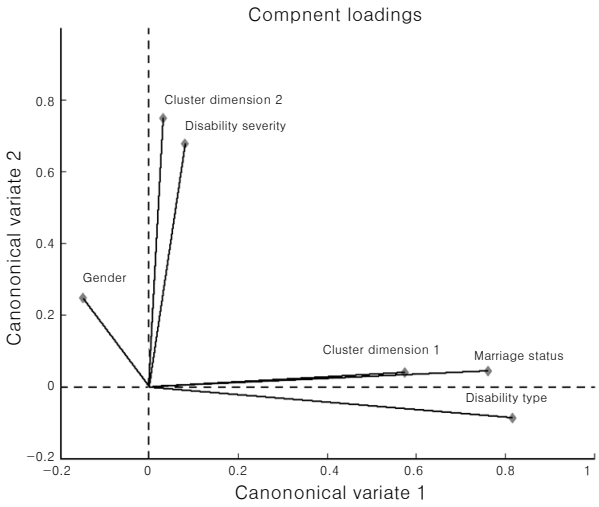

Figure 3. Component loadings.

The component loadings of each variable are measures of the correlations between the optimally scaled variables and the two orthogonal canonical variates. These are similar to factor loadings in PCA. The loadings for all variables are plotted in Figure 3. Note that the first dimension is measured along the abscissa and

Table 6. Weights of variables comprising canonical variates

Set	Variable	Dimension		
		1	2	R^2
1	Segmentation by social activity hindrance and encouragement factors	* –	* –	
2	Disability type	0.815	-0.090	0.672
	Severity of the disability	0.075	0.679	0.467
3	Gender	-0.048	0.255	0.067
	Marriage status	0.755	0.045	0.572
Summary of analysis	Eigenvalues (= Fit)	0.528	0.365	
	Canonical correlation	0.292	0.048	

* Weights are not unique for variables treated as multiple nominal.

the second along the ordinate. The length of the vector from the origin to the coordinates of each variable indicates the extent to which the variable is explained by the two canonical variates (i.e., the square of the length being equal to the percent of variance explained by all the other variables). The segmentation variable has two locations in the canonical space because it is allowed to have a different quantification for each dimension. The scalar (dot) product between any two variable vectors is indicative of the correlation between the two optimally scaled variables (ter Braak, 1990).

The components loadings plot shows that disability type and marriage status are highly related to differences among one of the disability perception segments (that most closely aligned with the first and the most powerful canonical dimension), while the severity of disability is correlated with the other less powerful dimension (i.e., the second dimension). Contribution to its explanation of gender is derived almost equally from each of the two canonical deviates variables, although gender is the least well-explained (R^2 = 0.067) by the two canonical variates. Disability type, on the other hand, is the best-explained disability characteristic (R^2 = 0.672), and contributions to its explanation are derived almost entirely from the first canonical deviate. Overall, Figure 3 supports the claim that trainees (including children and adolescents) with physical disabilities are less competent in their physical abilities and social life (Miyahara and Piek, 2006).

Figure 4a shows that disability type is almost entirely explained by the first canonical variate. Note that all physical types of disability (e.g., vision, orthopedic, and hearing disability) belong to the negative domain of the first variate, while all mental types of disability (e.g.,

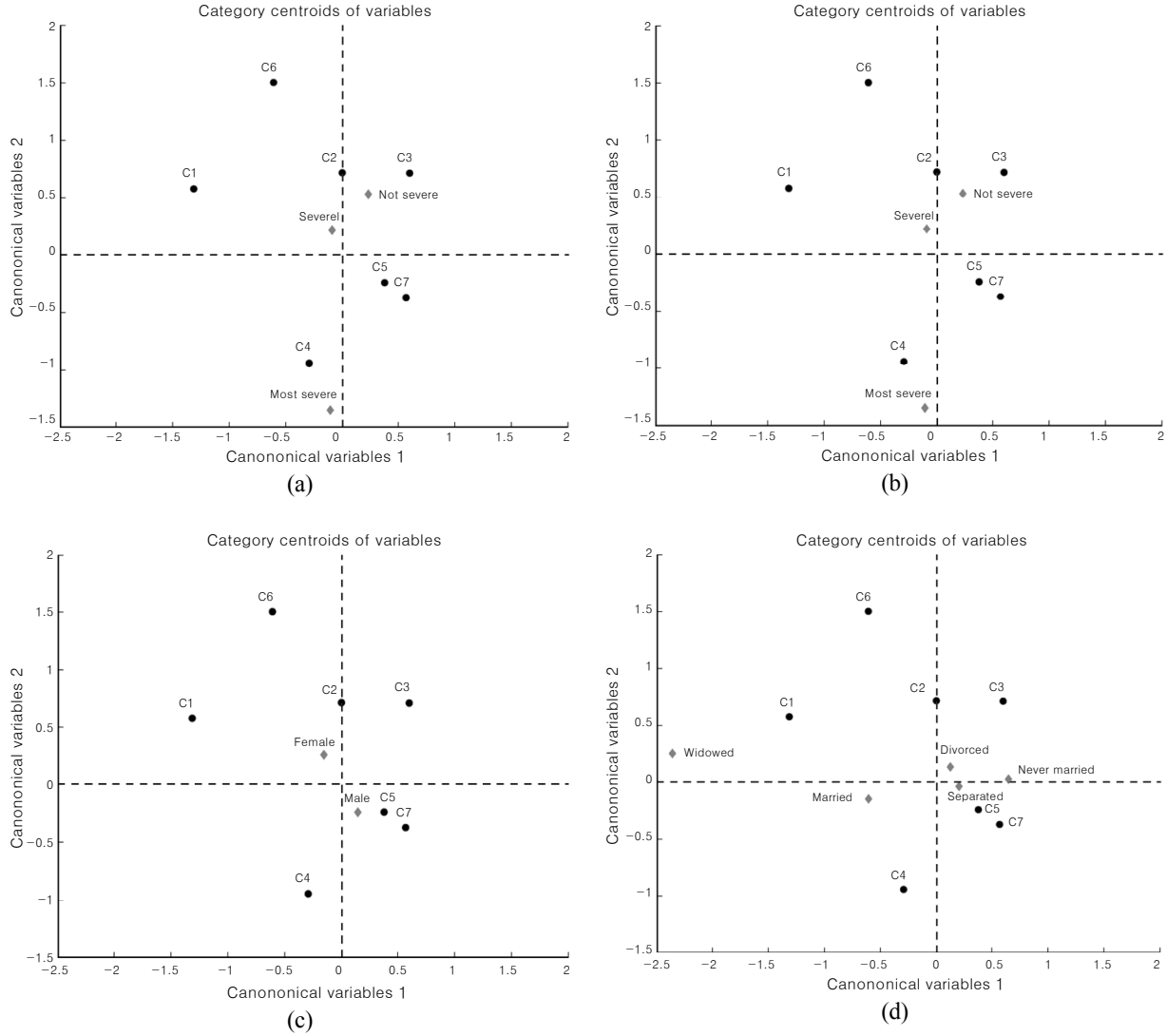

Figure 4. Category centroids for clusters and disability type (a), clusters and disability severity (b), clusters and gender (c), and clusters and marriage status (d).

brain injury, mental retardation, mental illness, non-orthopedic, and substance abuse) are located in the positive domain of the first variate. We also note that trainees with mental type of disability are aligned with segments (C3, C5, and C7) with a low level of self-esteem and self-confidence. These trainees typically do not consider their disability a significant hindrance factor of social activity. Trainees with vision disability are located closely to the C1 segment in which many trainees maintain a high level of self-esteem and self-confidence, but their disability imposes a significant hindrance on their social activity.

Figure 4b shows that the severity of the disability is mainly explained by the second canonical variate. It is interesting to observe that most severe disability is located in the negative domain while severe and non-severe disability is located in the positive domain of the second variate. This insinuates that trainees with most severe disability (located close to a segment C4) maintain a high self-esteem and self-confidence, but maintain a low level of physical and psychological hindrance factor. Trainees with non-severe disability are closely located to a segment C3 that shows a low level of self-esteem and self-confidence, and a low level of physical and psychological hindrance factor. We also observe that trainees with severe (but not most severe) disability are located in a close range to a segment C2 in which trainees maintain a low level of self-esteem and self-confidence, and suffer from physical and psychological hindrance factor. In short, most trainees with disabilities believe that their disability does not impose physical and psychological burden on their social activities, but it affects their self-esteem and self-confidence.

To our surprise, trainees with non-severe disability maintain a lower level of self-esteem and self-confidence than trainees with severe disability. We attribute this finding to the fact that trainees with non-severe disability often tend to lament after comparing their situations to those of trainees whom they consider normal. In contrast, trainees with severe physical disabilities have developed skills to accept themselves with physical limitation and find themselves self-confident by overcoming their disability and accomplishing little things (Miyahara and Piek, 2006). Note also that severe physical disabilities are highly visible and many people empathize them, while minor physical disability is often not noticed and hence people do not empathize when trainees with minor physical disabilities fail to complete required tasks (Miyahara and Register, 2000). When other trainees do not empathize their poor performance that partly comes from their minor physical disability and blame them for a lack of effort, trainees with minor disability often lose their competence and cannot develop desired self-esteem and self-confidence (Miyahara and Piek, 2006).

According to Figure 4c, gender variable was explained by both variates, but the alignment with either variate was not significant. In Figure 4d, the category centroids of marriage status in the canonical variates showed almost identical pattern as those of disability type in Figure 4a. That is, marriage status was mainly captured by the first variate. In particular, widowed and never married categories have most significantly aligned with the first variate, while divorced and separated categories are not tightly coupled with either variate.

6. CONCLUSION

In this study, we propose a multi-level analysis approach to better understand VR service data sets. When the outcomes of analyses at each level are integrated, a comprehensive understanding of VR data set becomes feasible. First, from the analysis at the first level, we obtain a decision tree classifier that successfully identifies trainees who are likely to find a job after completing a VR service program with an accuracy of 83.49% based on only 15 input variables. While having an accurate predictive model is useful, the final decision tree model consists of too many *if-then* rules, making it almost impossible to understand relationships among input variables and why and how some variables are associated with other variables. Therefore, two association algorithms are employed at the second level in an effort to provide additional insights from the VR data set. As a result, numerous associative relationships are extracted. Most of the associative rules can be used for profiling trainees, providing additional insights. For example, according to a few association rules, trainees who believe self-esteem is very important for their social activity are likely to find a job after VR training while trainees who believe self-esteem is not important for their social activity are not likely to have a job after VR training. It is also found that trainees who consider self-esteem, cultural factors, and sight impairment unimportant for their social activity are not likely to find a job after VR training. Finally, at the third level, clustering and NLCCA are employed to understand psychological and societal characteristics of VR trainees and to analyze the relationships between their internal psychological factors and their social activity. Our analysis confirms that many trainees with mental types of disability maintain a low level of self-esteem and self-confidence, but they do not believe that their disability imposes physical and psychological burden on their social activities. However, trainees with physical types of disability (e.g., vision disability) maintain a high level of self-esteem and self-confidence, but their disability imposes a significant hindrance on their social activity. Our analysis also finds a negative relationship between the severity of disability and the level of self-esteem and self-confidence (e.g., trainees with non-severe disability maintain a lower level of self-esteem and self-confidence).

As an extension of the current study, we are currently exploring a new NLCCA model that includes the post VR employment status variable itself as another set

of variables to explain the relationship between canonical variates and VR employment status variables along with clustering indexes, disability-related variables, personal characteristics, and social cognitive factors. Furthermore, we will draw insights on how we should develop and organize training programs to maximize the effectiveness of VR services for trainees with disabilities. For example, we would like to investigate relationships between job training indicator and the levels of self-esteem and self-confidence of trainees with disabilities. This is important to note because if trainees who have job-training (regardless of the fact that they actually have a job after VR training program) maintain much higher self-esteem and self-confidence level, the local and state governments may revise their VR programs to reach out to more trainees with disabilities to boost their self-esteem and self-confidence, which will lead to a higher quality of life for people with disability.

REFERENCES

Agrawal, R. and Srikant, R. (1994), Fast algorithms for mining association rules, *Proceedings of the 20th Very Large Data Bases Conference*, Santiago, Chile, 487-499.

Anandarajan, M. (2002), Profiling Web usage in the workplace: a behavior-based artificial intelligence approach, *Journal of Management Information Systems*, **19**(1), 243-266.

Aumann, Y. and Lindell, Y. (2003), A statistical theory for quantitative association rules, *Journal of Intelligent Information Systems*, **20**(3), 255-283.

Breiman, L. (1996), Bagging predictors, *Machine Learning*, **24**(2), 123-140.

Bandura, A. (1986), *Social Foundations of Thought and Action: A Social Cognitive Theory*, Prentice-Hall, Englewood Cliffs, NJ.

Bandura, A., Caprara, G. V., Barbaranelli, C., Gerbino, M., and Pastorelli, C. (2003), Role of affective self-regulatory efficacy in diverse spheres of psychosocial functioning, *Child Development*, **74**(3), 769-782.

Benbasat, I. and Zmud, R. W. (2003), The identity crisis within the is discipline: defining and communicating the discipline's core properties, *MIS Quarterly*, **27**(2), 183-194.

Brijs, T., Swinnen, G., Wanhoof, K., and Wets, G. (1999), Using association rules for product assortment decisions: a case study, *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, 254-260.

Clark, D. B., Moss, H. B., Kirisci, L., Mezzich, A. C., Miles, R., and Ott, P. (1997), Psychopathology in preadolescent sons of fathers with substance use disorders, *Journal of the American Academy of Child & Adolescent Psychiatry*, **36**(4), 495-502.

Cox, R. H. (1998), *Sport Psychology: Concepts and Applications*, McGraw-Hill, Boston, MA.

Edwards, T. and Hardy, L. (1996), The interactive effects of intensity and direction of cognitive somatic anxiety and self-confidence upon performance, *Journal of Sport and Exercise Psychology*, **18**(3), 296-312.

Flach, P. A. and Lachiche, N. (2001), Confirmation-guided discovery of first-order rules with Tertius, *Machine Learning*, **42**(1-2), 61-95.

Freund, Y. and Schapire, R. E. (1996), Experiments with a new boosting algorithm, *Proceedings of the 13th International Conference on Machine Learning*, Bari, Italy, 148-156.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999), Data clustering: a review, *ACM Computing Surveys*, **31**(3), 264-323.

Kim, Y. S. (2009), Streaming association rule (SAR) mining with a weighted order-dependent representation of Web navigation patterns, *Expert Systems with Applications*, **36**(4), 7933-7946.

Mazid, M. M., Shawkat Ali, A. B. M., and Tickle, K. S. (2008), Finding a unique association rule mining algorithm based on data characteristics, *Proceedings of the International Conference on Electrical and Computer Engineering*, Dhaka, Bangladesh, 902-908.

Miyahara, M. and Piek, J. (2006), Self-esteem of children and adolescents with physical disabilities: quantitative evidence from meta-analysis, *Journal of Developmental and Physical Disabilities*, **18**(3), 219-234.

Miyahara, M. and Register, C. (2000), Perceptions of three terms to describe physical awkwardness in children, *Research in Developmental Disabilities*, **21**(5), 367-376.

Mobasher, B., Dai, H., Luo, T., and Nakagawa, M. (2002), Discovery and evaluation of aggregate usage profiles for Web personalization, *Data Mining and Knowledge Discovery*, **6**(1), 61-82.

Padmanabhan, B. and Tuzhilin, A. (1999), Unexpectedness as a measure of interestingness in knowledge discovery, *Decision Support Systems*, **27**(3), 303-318.

Scheffer, T. (2005), Finding association rules that trade support optimally against confidence, *Intelligent Data Analysis*, **9**(4), 381-395.

Silberschatz, A. and Tuzhilin, A. (1996), What makes patterns interesting in knowledge discovery systems, *IEEE Transactions on Knowledge and Data Engineering*, **8**(6), 970-974.

Silverstein, C., Brin, S., and Motwani, R. (1998), Beyond market baskets: generalizing association rules to dependence rules, *Data Mining and Knowledge Discovery*, **2**(1), 39-68.

Spiliopoulou, M., Pohle, C., and Faulstich, L. C. (2000), Improving the effectiveness of a web site with web usage mining, *Web Usage Analysis and User Pro-*

filing, Lecture Notes in Computer Science, **1836**, 142-162.

ter Braak, C. J. F. (1990), Interpreting canonical correlation analysis through biplots of structure correlations and weights, *Psychometrika*, **55**(3), 519-531.

Thomas, A. P., Bax, M., and Smyth, D. P. L. (1989), *The Health and Social Needs of Young Adults with Physical Disabilities*, MacKeith, London, UK.

Vogt, C. M., Hocevar, D., and Hagedorn, L. S. (2007), A social cognitive construct validation: determining women's and men's success in engineering programs, *Journal of Higher Education*, **78**(3), 337-364.

Zimmerman, B. J., Bandura, A., and Martinez-Pons, M. (1992), Self-motivation for academic attainment: the role of self-efficacy beliefs and personal goal setting, *American Educational Research Journal*, **29**(3), 663-676.