Behavior Research Methods
2007, 39 (2), 259-266

ARTICLES

An introduction to association rule mining: An application in counseling and help-seeking behavior of adolescents

DION H. GOH AND REBECCA P. ANG
Nanyang Technological University, Singapore

Association rule mining (ARM) is a technique used to discover relationships among a large set of variables in a data set. It has been applied to a variety of industry settings and disciplines but has, to date, not been widely used in the social sciences, especially in education, counseling, and associated disciplines. This article thus introduces ARM and presents aspects of existing work that will be relevant and useful to researchers and practitioners in the social sciences. Definitions and concepts are presented, and examples of ARM applications are highlighted to strengthen these ideas. We also discuss an example from our existing research to show that ARM can be used to investigate help-seeking behavior in a sample of secondary school students in Singapore. We also present some guidelines and recommendations for using ARM.

First proposed by Agrawal, Imielinski, and Swami (1993), association rule mining (ARM) refers to the discovery of relationships among a large set of variables. That is, given a database of records, each containing two or more variables and their respective values, ARM determines variable-value combinations that frequently occur. Similar to the idea of correlational analysis (although they are theoretically different), in which relationships between two variables are uncovered, ARM is also used to discover variable relationships, but each relationship (also known as an association rule) may contain two or more variables. ARM has been extensively employed in business decision-making processes, and a typical example of its use is in market basket analysis (Agrawal et al., 1993). Here, ARM analyzes the buying habits of customers to identify associations between the different items that the customers place in their "shopping baskets." In such analyses, each record in the database represents a shopping basket, and the variables represent items. Variable values will typically be binary and indicate whether the respective item was purchased (true) or not (false). Put differently, ARM discovers what items customers typically purchase together. These associations can then be used by a supermarket, for example, to place frequently co-purchased items in adjacent shelves to increase sales. Thus, if bread and cereal are often purchased together, placing these items in close proximity may encourage customers to buy them within single visits to the supermarket.

ARM is a technique that is part of the field of data mining. Also known as knowledge discovery in databases, data mining attempts to discover useful information or patterns in large databases containing thousands to millions of records, where conventional statistical analysis is not feasible. Data mining is a broad field and encompasses many disciplines, including statistics, machine learning, databases, and artificial intelligence, among others. For this reason, a variety of techniques are available and are described in Fayyad, Piatetsky-Shapiro, and Smyth (1996) and Han and Kamber (2001). Data-mining techniques have been extensively employed in a number of industries, such as medical diagnosis, marketing and sales, and finance (Bose & Mahapatra, 2001; Klösgen & Żytkow, 2002; Langley & Simon, 1995). In particular, specific examples where ARM has been applied include the following: marketing and sales (Anand, Patrick, Hughes, & Bell, 1998; Wang, Chuang, Hsu, & Keh, 2004), medical diagnosis (Doddi, Marathe, Ravi, & Torney, 2001; Pendharkar, Rodger, Yaverbaum, Herman, & Benner, 1999), hospital administration (Brosette et al., 1998), education (Ma, Liu, Wong, Yu, & Lee, 2000; Zaïane & Luo, 2001), and law (Ivkovic, Yearwood, & Stranieri, 2002). A common theme in these examples is that they employ ARM to discover relationships among variables—that is, which variable-value combinations frequently occur. Such variables include items purchased, medical procedures, diagnosis codes, and examination grades.

Although ARM has broad applicability, as is demonstrated in the list of applications, it has not been widely used in the social sciences, especially in education, coun-

D. H. Goh, ashlgoh@ntu.edu.sg

seling, and associated disciplines. For example, a literature search in major databases, such as PsycINFO, Web of Science, and EBSCOhost, yielded a dearth of publications in which ARM was employed in these areas. In fact, the two ARM articles on education (Ma et al., 2000; Zaïane & Luo, 2001) were authored by computer scientists.

The focus of this article is thus to introduce ARM and demonstrate how it may be applied in the social sciences (e.g., education and counseling). Realizing that the ARM literature is huge and diverse, we attempt to synthesize and present aspects of existing work that will be relevant and useful to social science researchers and practitioners. Furthermore, since ARM is still a relatively new area of research in the social sciences, we also present guidance and recommendations for its use. The remainder of this article is organized as follows. The next section will review ARM and related concepts. We then will provide an example based on our existing research, to demonstrate how ARM can be applied in school-based counseling settings. We will conclude by discussing issues related to ARM and data mining.

AN INTRODUCTION TO ASSOCIATION RULE MINING

To better explain the concepts behind ARM, an education-related example adapted from market basket analysis will be used. Consider a university department that maintains a database of courses taken by its students. A hypothetical list of course combinations taken by nine students is shown in Table 1. Each row is often referred to as a *transaction*, each comprising a combination of *items* or *variables*. Suppose the department wants to discover popular combinations of courses taken by its students and finds, for example, that the counseling and the classroom management courses tend to be taken together. Such information could be used for curriculum planning to drop low-demand courses or introduce courses related to popular ones. It could also be used to determine whether course combinations taken by students meet the learning objectives set by the department and, also, to make recommendations for students needing suggestions.

Basic Concepts and Definitions

The analysis of frequently occurring variable combinations does not imply causality nor allude to any theories about education or student behavior. Rather, these relationships are found by analyzing the co-occurrences of variables in the database of transactions. Such relationships are known as association rules and may be defined as an implication of the form $A \Rightarrow B$, where A (also known as the antecedent) and B (also known as the consequent) are conjunctions of variable-value pairs. An association rule may be interpreted as meaning that *when the variables represented by A occur in a database, the variables represented by B also occur*. For example, the rule "Counseling $\wedge$ ClassroomManagement $\Rightarrow$ Statistics" suggests that when counseling and classroom management courses are taken (Counseling $\wedge$ ClassroomManagement), the statistics course is also taken ($\Rightarrow$ Statistics).

Table 1
Hypothetical Set of Course Combinations Taken by Students

No.	Course Combination
1	Counseling, statistics, classroom management
2	Statistics, creative thinking
3	Statistics, developmental psychology
4	Counseling, statistics, creative thinking
5	Counseling, developmental psychology
6	Statistics, developmental psychology
7	Counseling, developmental psychology
8	Counseling, statistics, developmental psychology, classroom management
9	Counseling, statistics, developmental psychology

An issue with ARM is that there is an exponential growth in the number of association rules as the number of variables used increases. In ARM, two measures are commonly used to help a researcher decide the usefulness of an association rule: *support* and *confidence*. The support of an association rule $A \Rightarrow B$ is the percentage of transactions that contain $A \cup B$. The confidence of an association rule $A \Rightarrow B$ is the ratio of the number of transactions that contain $A \cup B$ to the number of transactions that contain A.

Support measures how frequently an association rule occurs in the entire set of transactions, whereas confidence measures the strength/reliability of a rule. In ARM, rules are selected only if they satisfy both a minimum support and a minimum confidence threshold. Table 2 lists some examples of association rules, together with their support and confidence values generated from Table 1. For example, in Rule 3, the courses in the set co-occur in approximately 20% of the entire set of transactions, and this rule holds all the time: Every time counseling and classroom management are taken, statistics is also taken. This is a stronger rule than Rule 5, since only 17% of the time when counseling is taken is creative thinking also taken.

Generating Association Rules

The *Apriori* algorithm (Agrawal & Srikant, 1994) is an influential algorithm in ARM for generating association rules and is found in most commercial data-mining software. The basic algorithm requires variables to be categorical, and numeric variables will need to be discretized, although more sophisticated variants do not have such a restriction. The Apriori algorithm is so named because it is based on the fact that it uses prior knowledge of *frequent itemset* properties known as the *Apriori property*. These terms are defined as follows. An *itemset* is any subset of all the items in the database of transactions. An itemset that contains k items is known as a k-itemset. A *frequent* (or *large*) itemset is one that satisfies a minimum support threshold. The Apriori property states that all nonempty subsets of a frequent itemset must also be frequent. That is, if an itemset satisfies the minimum support threshold, so do all of its subsets. Conversely, if an itemset does not satisfy the minimum support threshold, any superset of it will not be frequent as well.

The Apriori algorithm is iterative. In each iteration i, it generates candidate itemsets C_i of size i from the database of transactions and then counts these to see whether they are

Table 2
Support and Confidence for Some Association Rules

No.	Association Rule	Support	Confidence
1	Counseling ⇒ DevelopmentPsychology	4/9 = .44	4/6 = .67
2	DevelopmentPsychology ⇒ Statistics	4/9 = .44	2/2 = .67
3	Counseling ∧ ClassroomManagement ⇒ Statistics	2/9 = .22	2/2 = 1.00
4	Counseling ∧ Statistics ∧ DevelopmentPsychology ⇒ ClassroomManagement	1/9 = .11	1/2 = .50
5	Counseling ⇒ CreativeThinking	1/9 = .11	1/6 = .17

frequent (i.e., satisfy the minimum support threshold). Only those candidates that are frequent (denoted as L_i) are used to generate candidate itemsets C_{i+1} for the next iteration. To generate the next set of C_{i+1} candidates of size $i+1$, joins are made of frequent itemsets, L_i, found in the previous iteration. Here, a join takes place if two itemsets have $i-1$ items in common. Duplicate candidates are discarded after the join process is completed. The process stops when all C_{i+1} candidates in the next iteration are not frequent.

Following this, association rules are generated for every frequent itemset l for all the itemsets, L_i. Here, for every l, all nonempty subsets of l are generated. Next, for each nonempty subset s belonging to l, a rule of the form $s \Rightarrow (l - s)$ is generated only if it satisfies the minimum confidence threshold. All the rules are guaranteed to satisfy the minimum support threshold since they are derived from itemsets that already satisfy this requirement. For example, Table 3 shows the association rules generated from the frequent three-itemset {*Counseling*, *Statistics*, *ClassroomManagement*}. Only those rules meeting the minimum support and confidence thresholds are accepted.

A STUDY OF HELP-SEEKING BEHAVIOR AMONG SECONDARY SCHOOL STUDENTS

This section will demonstrate how ARM can be applied to school-based counseling research. Different cultural groups have their own characteristic and preferred ways of handling personal problems (Sue & Sue, 1999). Most studies on help-seeking behavior have consistently shown that adolescents rarely seek professional help (Offer, Howard, Schonert, & Ostrov, 1991), preferring instead to approach informal helping agents, such as parents or friends (Dubow, Lovko, & Kausch, 1990; Tishby et al., 2001). This preference for informal helping agents also extends to Chinese students (Cheung, 1984). In addition, Asian families in the United States have been found to underutilize mental health services and to overutilize informal sources of support, in comparison with other ethnic groups (Suan & Tyler, 1990). There is limited published research on adolescent help-seeking behavior in Asia. To date, fewer than five studies on adolescent help-seeking attitudes and behavior based in Asia have been identified (Ang, Lim, Tan, & Yau, 2004; Fukuhara, 1986; Yeh, 2002). More empirical work is urgently needed to better understand Asian adolescents who seek professional help versus those who do not. In addition, among adolescents who have not sought professional help, it would be crucial to differentiate those who are open to this possibility in the future versus those who are not.

The purpose of the present study was to investigate help-seeking behavior in a sample of secondary school students in Singapore. The study is part of a larger collaborative research study with the Tampines Family Service Center (TFSC), a voluntary welfare organization in Singapore, to better understand secondary students' attitudes toward counseling and help-seeking behavior. A voluntary welfare organization in Singapore is a nonprofit organization that works with government authorities, private organizations, and the community to provide social and community services. Specifically, in this study, the following research question was examined: What are the characteristics of students who are open to seeking counseling and of those who are not?

Since the focus of this article is on introducing ARM and demonstrating its use in school-based counseling research, we will present only the methodology and results of our study that are relevant to these goals. Hence, the analyses presented that pertain to counseling and adolescent help-seeking behavior are not meant to be comprehensive. Readers interested in the substantive counseling-based issues of this research may refer to Ang and Yeo (2004).

Method

Participants. Four hundred forty-eight secondary school students from one secondary school in Singapore

Table 3
Association Rule Generation Example for
{Counseling, Statistics, ClassroomManagement}

Rule	Support	Confidence	Accepted?
Counseling ∧ Statistics ⇒ ClassroomManagement	.22	2/4 = .50	Yes
Counseling ∧ ClassroomManagement ⇒ Statistics	.22	2/2 = 1.00	Yes
Statistics ∧ ClassroomManagement ⇒ Counseling	.22	2/2 = 1.00	Yes
Counseling ⇒ Statistics ∧ ClassroomManagement	.22	2/6 = .33	No
Statistics ⇒ Counseling ∧ ClassroomManagement	.22	2/7 = .29	No
ClassroomManagement ⇒ Counseling ∧ Statistics	.22	2/2 = 1.00	Yes

participated in the study. There were 245 male students (54.7%), 195 female students (43.6%), and 8 students (1.7%) who did not provide any information about gender. The age of the students ranged from 12 to 18 years, with a mean age of 14.28 years (*SD* = 1.77). Self-reported ethnic identification for the sample was as follows: 51.1% of the participants were Chinese, 39.5% were Malay, 3.6% were Indian, 1.8% endorsed "Others" (which includes all other ethnic groups not listed), and 4% did not provide information on ethnic identification.

Questionnaire. All the participating students completed a short questionnaire consisting of a demographic sheet, questions relating to the students' preferences for counselor characteristics, such as gender and ethnicity, and questions relating to the students' help-seeking behavior in general. An additional question that was asked of the students concerned to whom they might talk when they encountered problems. For this question, the students could select one or more of the following options: friends, parents, brothers/sisters, other relatives, teachers, counselors, or others (e.g., religious leaders or medical doctors). Finally, questions were asked about the students' opinions on those who spoke to counselors and the problems that they worry about most.

Procedure. We wanted to find relationships between the students' attitudes, opinions, and behaviors and whether they sought or would seek counseling help or not. ARM was selected for this exploratory study because it allowed us to examine combinations of variables that characterize students in relation to our research question. Specifically, the goal was to discover frequently occurring attitudes, opinions, and behaviors that describe help-seeking behavior. In the general form of ARM, there is no restriction on the variables that can appear as the antecedent and consequent. Our research question, however, required that the consequents were restricted to one variable: "If you were to have problems in the future, would you see a counselor about it?" ("yes" or "no"). Stated differently, this variable divided the participants into two groups on the basis of their responses. All the other variables were used as antecedents and served to characterize the participants who were open to the possibility of seeing a counselor and those who were not. An alternative to using ARM in this way, in which a single variable serves as a consequent, is to employ supervised-learning techniques (Dunham, 2003) that are also a part of the field of data mining. These techniques are used to classify or predict dependent variables, which may be categorical or numeric, and examples include decision trees and neural networks. Employing different criteria for mining patterns, such techniques may be used to complement or confirm the results of ARM.

In consultation with a domain expert (one of the authors) familiar with the larger study and with counseling and adolescent help-seeking behavior, variables not relevant were removed, to prevent the generation of too many rules and making their interpretation difficult. For example, variables denoting gender and racial preferences for counselors were removed. The Apriori algorithm implemented by Clementine (SPSS, 2005), a commercial data-mining software product, was used in the study. Due to the relatively small size of the data set, a low support threshold of 1% was used to ensure that as many rules were generated as possible. Sliding confidence values were used, starting initially at 60%. At each stage, the rules were inspected by the domain expert to determine their interpretability. Confidence values were reduced by 10% until the final threshold value of 40% was reached. The number of rules generated for each of the three stages was 1,161 (60% confidence value), 1,186 (50% value), and 1,207 (40% confidence value).

In addition, twofold cross-validation was performed to ensure stability of the association rules. The data set was randomly split into two subsets, A and B. Subset A (*N* = 184) was used to generate association rules, employing the Apriori algorithm, and for further analyses, where Subset B (*N* = 182) was used to verify the stability of those rules, again employing the Apriori algorithm. Cross-validation results indicated that the association rules generated across Subsets A and B were similar, with support for many rules varying by only a few percentage points (refer to Tables 4

Table 4
Association Rules for Adolescents Open to the Possibility of Seeing a Counselor

No.	Rule	Support (%)	Confidence (%)	Cross-Validated Support (%)	Cross-Validated Confidence (%)
1	Normal people ∧ Counseling time ∧ Talk to teachers ⇒ See counselor	12.5	100	11.54	90.48
2	Talk to parents ∧ Talk to teachers ⇒ See counselor	11.41	100	9.34	88.24
3	School work ∧ Talk to counselors ⇒ See counselor	9.78	100	7.14	100.00
4	Self-referral ∧ Talk to parents ∧ Talk to teachers ⇒ See counselor	9.78	100	7.69	92.86
5	Counseling time ∧ Loneliness ∧ Talk to teachers ⇒ See counselor	5.43	100	6.59	100.00

Note—"See counselor," adolescents who are open to seeing a counselor; "Normal people," adolescents' opinions that those who speak to counselors are normal people who need some guidance and advice; "Counseling time," adolescents' opinion that counseling does not take up too much time; "Self-referral," the knowledge that adolescents can see a counselor for help on their own without a formal referral; "Talk to parents," "Talk to teachers," "Talk to counselors," when adolescents responded to the question "When you have problems, who would you usually talk to?"; "School work," "Loneliness," when adolescents were asked about the problems that they worry about most.

Table 5
Association Rules for Adolescents Not Open to the Possibility of Seeing a Counselor

No.	Rule	Support (%)	Confidence (%)	Cross-Validated Support (%)	Cross-Validated Confidence (%)
1	Not self-referral $\wedge$ Not talk to counselors $\wedge$ Not talk to teachers $\Rightarrow$ Not see counselor	10.11	67.27	9.89	81.82
2	Not self-referral $\wedge$ Not talk to counselors $\Rightarrow$ Not see counselor	10.33	65.65	9.89	85.71
3	Not self-referral $\wedge$ Not talk to teachers $\Rightarrow$ Not see counselor	10.33	65.52	9.89	85.71
4	Not self-referral $\wedge$ Counseling time $\Rightarrow$ Not see counselor	10.87	60.61	10.44	70.37
5	Not self-referral $\Rightarrow$ Not see counselor	10.87	60.61	10.44	70.37

Note—"Not see counselor," adolescents who are not open to seeing a counselor; "Not self-referral," adolescents who responded that they did not know they could see a counselor for help on their own without a formal referral; "Not talk to teachers," "Not talk to counselors," when adolescents responded negatively to the question "When you have problems, who would you usually talk to?"; "Counseling time," adolescents' opinion that counseling does not take up too much time.

and 5). Larger variations in confidence values for some rules can be accounted for by the small subset sizes.

During association rule generation, the domain expert had to sift through the rules to extract those that yielded useful information. The cross-validation procedure assisted in this, since only rules that appeared in both subsets were considered for further evaluation by the domain expert. Next, rules that were redundant or contradictory were removed. An example of redundant rules is the different permutations of variables in a set of antecedents that characterized adolescent students who had not seen a counselor but were open to the possibility of seeing one in the future. Here, commonly occurring variables in the association rule antecedents included a willingness to speak to parents, counselors, and teachers, as well as the students' positive attitudes toward counseling. In general, the number of association rules rapidly increases with the increase in the number of variables. The domain expert thus had to decide which set of antecedents constituted useful information in the present study. On the other hand, contradictory rules involve sets of rules with permutations of opposing antecedents but the same consequent. For example, one rule might indicate that bullying in school is not a major concern among students open to seeing a counselor, whereas another rule might indicate otherwise. Here, the domain expert had to examine the other variables in the antecedents to determine which rule should be considered as useful information. Note that there is no substitute for domain expertise and, hence, manual inspection. Therefore, with larger data sets that yield tens of thousands of association rules, it is advisable that support and confidence thresholds be set higher initially and then gradually decreased.

Results and Discussion

Association rules were generated to determine the characteristics of adolescent students who had not seen a counselor but were open to the possibility of seeing a counselor in the future and of those who had not seen a counselor and were not open to the possibility of seeing a counselor in the future. For clarity, these two sets of rules are presented separately in Tables 4 and 5, respectively. Only a selection of high-support and high-confidence rules relevant to the present study are shown for brevity, and implications for further research on adolescent help-seeking behavior will be discussed.

Two distinct patterns of association rules emerged across the two groups of adolescent students. For those open to seeing a counselor, problems such as school work and loneliness surfaced (see Rules 3 and 5 in Table 4). Positive attitudes toward seeing a counselor were also found among all the adolescents (100% confidence), who felt that students who spoke to counselors are normal people who need some additional help and guidance and that counseling was not a time-consuming process (see Rules 1 and 5 in Table 4). In addition, the knowledge that adolescents could see a counselor on their own without being formally referred was important (see Rule 4 in Table 4). Finally, these adolescents seemed open to talking to adults and authority figures, in that they expressed a willingness to talk to counselors, parents, and teachers when they encountered problems (see Rules 1–5 in Table 4).

In contrast, adolescents not open to seeing a counselor exhibited a different set of characteristics. In particular, they were not aware that they could see a counselor on their own without being formally referred (see Rules 1–5 in Table 5). Although the confidence values were within the 60%–80% range, this response was still in the majority. Another major difference was that adolescents in this group did not appear to be open to talking to adults and authority figures when they encountered problems (see Rules 1–3 in Table 5). However, they did have a similar positive attitude toward counseling, in that they felt that counseling was not a time-consuming process (see Rule 4 in Table 5).

Chi-square tests were used to determine whether these association rules could accurately characterize and differentiate students who were open to counseling and those who were not. In the present study, the phi coefficient was used to report nonparametric effect size estimates (Kline, 2004). For brevity, only a few examples will be described here. Chi-square tests yielded statistically significant results between the two groups of adolescents for the following association rules: "self-referral" [$\chi^2(1, N = 184) = 7.79, p = .005, \varphi = .21$], "talk to parents and talk to teachers" [$\chi^2(1, N = 184) = 15.24, p < .001, \varphi = .29$], and "not self-referral and not talk to counselors" [$\chi^2(1, N = 184) = 10.09, p = .001, \varphi = .23$].

Some researchers have suggested that for Asians, the underutilization of professional psychological help has included lack of knowledge or cultural stigmas (Leong, 1986; Leong & Lau, 2001; Mau & Jepsen, 1988; Yeh, 2002). The present study of adolescent students in Singapore supports such work, since the data indicate that lack of knowledge is one of the reasons why adolescents underutilize formal sources of help. In addition, the analyses suggest that adolescents who refuse to seek professional help very likely consist of those who choose not to talk to adults or authority figures, especially those outside their immediate family.

The discussion above illustrates how an application of ARM to adolescent help-seeking data may be of interest to educators and counselors. Mental health researchers and practitioners may benefit substantially from a better understanding of the characteristics associated with the adolescent help-seeking process: The association rules generated provide an indication of various adolescent help-seeking characteristics that are likely to occur together. In addition, the TFSC was able to use the profiles generated by the association rules to assist in the identification of groups of students who might be open to counseling versus those who were not. The provision of counseling services was then tailored specifically to address the barriers to adolescent help-seeking and to further enhance outreach programs. Finally, these association rules provide a sound basis for the generation of future research hypotheses that can then be empirically tested.

GENERAL DISCUSSION

Addressing the Data-Mining Controversy

ARM, being classified as a data-mining technique, is subject to the same criticisms as those directed at statisticians and researchers who rely heavily on statistical methods. For example, data mining has had negative connotations in the statistics literature, and the term is sometimes synonymous with data dredging or fishing (Chatfield, 1995; Selvin & Stuart, 1966)—the process of trawling through data in the hope of discovering interesting patterns. Glymour, Madigan, Pregibon, and Smyth (1997) have argued, however, that data-mining techniques have their place, especially in situations involving large numbers of variables and records. Here, computational efficiency and scalability may take precedence over statistical consistency. Hand and Blunt (2001) concurred and drew parallels between data mining and exploratory data analysis, which has gained respectability due to the work of Tukey (1977). A major difference, however, is the size of the data sets involved in data mining, thus requiring storage and manipulation techniques that are not addressed in statistics. Note, however, that there is no agreement on the optimal sizes of the databases used in data mining, and interesting and useful patterns have been obtained from databases as small as a few kilobytes (e.g., Brosette et al., 1998, and our present study).

Guidelines for ARM

Domain expertise. Since ARM, like many data-mining techniques, is exploratory, it is important that an expert in the area being investigated should be available for consultation, especially because there are no concrete rules establishing a "good" set of association rules. Glymour et al. (1997) also have pointed out that data mining is often pattern focused, rather than model focused, hence requiring a domain expert at hand for interpretation. Expertise is required for a variety of tasks, including variable selection, evaluation of association rules, and selection of association rules.

Variable selection. Although ARM is able to generate association rules with data sets containing hundreds or even thousands of variables, the number of association rules will grow exponentially as well. Reducing the number of variables by discarding irrelevant ones, with the consultation of a domain expert, will reduce the number of rules. Furthermore, the generated association rules will be easier to interpret, since there will be less noise introduced by these irrelevant variables. Since ARM is an iterative process, a researcher can always start with a smaller set of variables deemed relevant to the research objectives and then increase the number of variables in future iterations if the association rules obtained are not considered satisfactory.

Generating association rules. The generation of association rules is dependent on the minimum support and confidence threshold values: the higher these values, the smaller the number of rules generated, holding other parameters constant. No single set of guidelines exists for selecting appropriate support and confidence values. However, a practical suggestion is to employ sliding support and confidence values, beginning either with large values, as is done in our work, or with small values (e.g., Chen, Chou, & Hwang, 2003).

Selecting association rules. Once a candidate set of association rules is generated, given support and confidence parameters, the next step is the selection of the "best" rules for use in decision making. Once again, no standard set of procedures is available for selecting association rules. One possible approach is to use *n*-fold cross-validation, where the data set is randomly subdivided *n* times and ARM is performed on each of these subdivisions to obtain association rules. The generated rules are then empirically compared across the subdivisions to assess their stability and replicability. In the context of the present study, this was accomplished using chi-square tests to determine whether the association rules could characterize and differentiate adolescents who were open to counseling and those who were not.

An alternative approach, suggested by Ivkovic, Yearwood, and Stranieri (2003), is to formulate and test hypotheses by grouping association rules with common consequents and with the antecedents containing the "dependent" variable of interest. For example, consider the following two rules "See counselor = Self referral ^ Talk to parents (confidence = 60%)" and "Not see counselor = Self referral ^ Talk to parents (confidence = 30%)." Ivkovic, Yearwood, and Stranieri (2003) proposed that chi-square tests be used to select association rules by determining whether the deviation between confidence values in these association rules is statistically significant. Another technique for rule selection, proposed by Bay and Pazzani (1999), detects differences between contrasting groups. Here, conjunctions of variable-values with deviations of support values that are statistically significant (using the chi-square test) are sought. In their work, contrasting groups are characterized by association rules with different antecedents but sharing the same consequent. In sum, a variety of techniques have been proposed to assist in the selection of association rules, but to date, none are standard features in commercial data-mining software products. The domain expert will thus have to perform these computations separately, using such products as statistical software packages.

CONCLUSION

This article has introduced ARM and has shown how it can be used in social-science-related fields such as education and counseling. We began by explaining important concepts in ARM and describing how an important ARM algorithm, Apriori, generates association rules. Instances of ARM usage in the literature were also highlighted to illustrate its broad applicability, and this was followed by a counseling-related example from our research. As its name suggests, ARM is used to find relationships or associations between frequently occurring variables. An important point to note is that the discovered rules in ARM should not be taken as a definitive model that describes the solution to the problem at hand, because association rules are generated on the basis of frequency counts of variables in the data sets. No knowledge about the domain being investigated is incorporated into the rule generation process. Consequently, a rule may or may not make sense, even if it has high support and confidence values. The ARM process is thus exploratory and requires further validation of the rules in consultation with a domain expert. Our goal is not to claim that ARM is a superior technique and a universal solution, but rather to inform researchers of a valuable resource that could be incorporated into their arsenal of data analysis tools.

AUTHOR NOTE

The authors thank the Tampines Family Service Center for assistance with data collection. Address correspondence to D. H. Goh, Division of Information Studies, School of Communication and Information, Nanyang Technological University, 31 Nanyang Link, Singapore 637718 (e-mail: ashlgoh@ntu.edu.sg).

REFERENCES

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In P. Buneman & S. Jajodia (Eds.), *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data* (pp. 207-216). New York: ACM Press.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In J. B. Bocca, M. Jarke, & C. Zaniolo (Eds.), *Proceedings of the 20th International Conference on Very Large Data Bases* (pp. 487-499). San Francisco: Morgan Kaufmann.

Anand, S. S., Patrick, A. R., Hughes, J. G., & Bell, D. A. (1998). A data mining methodology for cross-sales. *Knowledge-Based Systems*, **10**, 449-461.

Ang, R. P., Lim, K. M., Tan, A.-G., & Yau, T. Y. (2004). Effects of gender and sex role orientation on help-seeking attitudes. *Current Psychology*, **23**, 203-214.

Ang, R. P., & Yeo, L. S. (2004). Asian secondary school students' help-seeking behaviour and preferences for counsellor characteristics. *Pastoral Care in Education*, **22**, 40-48.

Bay, S. D., & Pazzani, M. J. (1999). Detecting change in categorical data: Mining contrast sets. In U. Fayyad, S. Chaudhuri, & D. Madigan (Eds.), *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 302-306). New York: ACM Press.

Bose, I., & Mahapatra, R. K. (2001). Business data mining—a machine learning perspective. *Information & Management*, **39**, 211-225.

Brosette, S. E., Sprague, A. P., Hardin, J. M., Waites, K. B., Jones, W. T., & Moser, S. A. (1998). Association rules and data mining in hospital infection control and public health surveillance. *Journal of the American Medical Informatics Association*, **5**, 373-381.

Chatfield, C. (1995). Model uncertainty, data mining, and statistical inference. *Journal of the Royal Statistical Society: Series A*, **158**, 419-466.

Chen, T.-J., Chou, L.-F., & Hwang, S.-J. (2003). Application of a data-mining technique to analyze coprescription patterns for antacids in Taiwan. *Clinical Therapeutics*, **25**, 2453-2463.

Cheung, F. M. (1984). Preferences in help-seeking among Chinese students. *Culture, Medicine, & Psychiatry*, **8**, 371-380.

Doddi, S., Marathe, A., Ravi, S. S., & Torney, D. C. (2001). Discovery of association rules in medical data. *Medical Informatics & The Internet in Medicine*, **26**, 25-33.

Dubow, E. F., Lovko, K. R., & Kausch, D. F. (1990). Demographic differences in adolescents' health concerns and perceptions of helping agents. *Journal of Clinical & Child Psychology*, **19**, 44-54.

Dunham, M. H. (2003). *Data mining: Introductory and advanced topics*. Upper Saddle River, NJ: Prentice Hall/Pearson Education.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, **39**, 27-34.

Fukuhara, M. (1986). The attitude of students towards consultation/counseling. *School Psychology International*, **7**, 76-82.

Glymour, C., Madigan, D., Pregibon, D., & Smyth, P. (1997). Statistical themes and lessons for data mining. *Data Mining & Knowledge Discovery*, **1**, 11-28.

Han, J., & Kamber, M. (2001). *Data mining: Concepts and techniques*. San Francisco: Morgan Kaufmann.

Hand, D. J., & Blunt, G. (2001). Prospecting for gems in credit card data. *IMA Journal of Management Mathematics*, **12**, 173-200.

Ivkovic, S., Yearwood, J., & Stranieri, A. (2002). Discovering interesting association rules from legal databases. *Information & Communications Technology Law*, **11**, 35-47.

Ivkovic, S., Yearwood, J., & Stranieri, A. (2003). Visualizing association rules for feedback within the legal system. In G. Sartor (Ed.), *Proceedings of the 9th International Conference on Artificial Intelligence and Law* (pp. 214-223). New York: ACM Press.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Klösgen, W., & Żytkow, J. M. (2002). *Handbook of data mining and knowledge discovery*. Oxford: Oxford University Press.

Langley, P., & Simon, H. A. (1995). Applications of machine learning and rule induction. *Communications of the ACM*, **38**, 54-64.

Leong, F. T. L. (1986). Counseling and psychotherapy with Asian Americans: Review of the literature. *Journal of Counseling Psychology*, **33**, 196-206.

Leong, F. T. L., & Lau, A. S. L. (2001). Barriers to providing effective mental health services to Asian Americans. *Mental Health Services Research*, **3**, 201-214.

Ma, Y., Liu, B., Wong, C. K., Yu, P. S., & Lee, S. M. (2000). Targeting the right students using data mining. In R. Ramakrishnan, S. Stolfo, R. Bayardo, & I. Parsa (Eds.), *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 457-464). New York: ACM Press.

Mau, W.-C., & Jepsen, D. A. (1988). Attitudes towards counselors and counseling processes: A comparison of Chinese and American graduate students. *Journal of Counseling & Development*, **67**, 189-192.

OFFER, D., HOWARD, K. I., SCHONERT, K. A., & OSTROV, E. (1991). To whom do adolescents turn for help? Differences between disturbed and nondisturbed adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*, **30**, 623-630.

PENDHARKAR, P. C., RODGER, J. A., YAVERBAUM, G. J., HERMAN, N., & BENNER, M. (1999). Association, statistical, mathematical and neural approaches for mining breast cancer patterns. *Expert Systems With Applications*, **17**, 223-232.

SELVIN, H. C., & STUART, A. (1966). Data dredging procedures in survey analysis. *American Statistician*, **20**, 20-23.

SPSS (2005). *Clementine* [Computer software]. Retrieved July 2, 2005, from www.spss.com/clementine/.

SUAN, L. V., & TYLER, J. D. (1990). Mental health values and preference for mental health resources of Japanese-American and Caucasian-American students. *Professional Psychology: Research & Practice*, **21**, 291-296.

SUE, D. W., & SUE, D. (1999). *Counseling the culturally different: Theory and practice* (3rd ed.). New York: Wiley.

TISHBY, O., TUREL, M., GUMPEL, O., PINUS, U., BEN LAVY, S., WINOKOUR, M., & SZNAJDERMAN, S. (2001). Help-seeking attitudes among Israeli adolescents. *Adolescence*, **36**, 249-264.

TUKEY, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

WANG, Y.-F., CHUANG, Y.-L., HSU, M.-H., & KEH, H.-C. (2004). A personalized recommender system for the cosmetic business. *Expert Systems With Applications*, **26**, 427-434.

YEH, C. J. (2002). Taiwanese students' gender, age, interdependent and independent self-construal, and collective self-esteem as predictors of professional psychological help-seeking attitudes. *Cultural Diversity & Ethnic Minority Psychology*, **8**, 19-29.

ZAÏANE, O. R., & LUO, J. (2001). Towards evaluating learners' behavior in a Web-based distance learning environment. In T. Okamoto, R. Hartley, M. Kinshuk, & J. Klus (Eds.), *IEEE International Conference on Advanced Learning Technologies (ICALT'01): Issues, achievements and challenges* (pp. 357-360). Los Alamitos, CA: IEEE Computer Society Press.

(Manuscript received April 24, 2005;
revision accepted for publication January 10, 2006.)