

Using Localized Features for Analyzing College Students' Imagination

Chuan-Ming Liu ^{1*}, Kuan-Teng Liao ¹

¹ Department of Computer Science and Information Engineering, National Taipei University of Technology, Taipei, TAIWAN

Received 1 February 2018 • Revised 21 October 2018 • Accepted 30 November 2018

ABSTRACT

The analysis of imagination has become popular in recent years because imagination is one of the key components of creativity and innovation. For extracting students' implicit degrees and thought processes of imagination, we use frequent pattern mining and association rule extraction to localize the features and explain the deep meanings of imagination in the study. By our observations, these two methods may sometimes explore meaningless frequent patterns and rules on a global sparse dataset. In order to eliminate such phenomena when mining with these two methods, we use a localized feature approach called forecast with clustering and integration (FCI) to improve the drawbacks of two methods on a sparse dataset. The approach consists of two strategies. One is clustering and the other is the prediction based on integration from (1) frequent patterns, (2) association rule pruning with correlation, and (3) forecast with linear regression. The former strategy can reduce the number of samples and highlight the features of imagination and the latter strategy can prune meaningless information and predict the trend of scores from imagination input data. Experimental results show both proposed approaches can localize special features, thereby providing supervisors with meaningful information about students' degrees and thought processes of imagination.

Keywords: imagination, frequent pattern mining, association rule extraction, DBSCAN, correlation analysis, linear regression

INTRODUCTION

Imagination analysis has become popular in recent decades (Safar and Alkhezzi, 2016; Huang et al, 2017) because imagination is a major driving force of human progress in stimulating creativity and innovation, and a critical factor promoted by education researchers (Eisner, 2001). For boosting imagination, the primary task is comprehending imagination cognition by measuring imagination with an imagination scale. By evaluating the degrees of students' imagination, the implicit information can be obtained, such as the exploration of inspired processes as they answer the scale. To comprehend students' imagination cognition, statistical analyses such as correlation, linear regression, may be used at the initial stage. However, compared to statistical analyses, mining technique may be an approach to support supervisors to obtain the deep meanings behind imagination input data. The objective of this paper thus is to adopt mining techniques with statistical analyses to analyze imagination for widely comprehending and evaluating students' imagination as well as predicting students' imagination degrees and students' imaginative thought processes. The derived results can assist supervisors in various aspects of exploration.

In mining technique, frequent pattern mining (FPM) and association rule exploration (ARE) (Ge & Xia, 2016) are two common mining approaches to show frequent items and frequent rules from a dataset. According to the properties of using frequent items and inferring rules, supervisors can easily reveal various imaginative phenomena from input datasets. However, in some situations, the size of datasets is usually small, and then results in sparse datasets. Unfortunately, in the case of sparse input datasets, the results obtained using these two approaches may be usually imprecise and inaccurate because of low relative supports. To avoid the effect of low

Contribution of this paper to the literature

- We contribute strategies for analyzing correlations and association rules to reduce meaningless relationships when analyzing imagination.
- We evaluate students' imagination and predict the degree of imagination using FCI as well as provide representative features for supervisors to conclude what the students need to improve the imagination.
- Our proposed model can reveal the special features efficiently and be widely applied to explore the representative relationships from spare datasets.

relative supports, we concentrate on localized feature strategies to evaluate frequent items and, thereby, infer the kernel association rules for enhancement.

The first localized feature strategy is top-K clustering. By top-K clustering, the data can be separated into several groups, and the representative features can be explored from the largest K groups individually. Generally, clustering, an unsupervised learning, can aggregate similar data into several groups and use in constructing the data model from a dataset with unknown categories. The clustered groups can be expressed as the localized features of all input data because clustering can separate global data into several local data. To localize students' imagination, in this study, we use the density-based clustering scheme, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), for aggregating input data. Unlike centroid-based clustering approaches such as *k*means approaches, DBSCAN aggregates input data without restricting circle-shaped clusters because it considers the relationships such as Euclidean distance or cosine similarity between two data points. This feature of relationship results in grouping the similar input data into the same clusters. By contrast, the results obtained using a centroid-based clustering approach may be different due to the locations of the initial centroids. Hence, clustering using DBSCAN can yield more stable results than that with any centroid-based clustering approach. After clustering, researchers can choose the largest K clusters for analysis. These K clusters contain the features that most students have. Therefore, the results of frequent pattern mining and association rules by using top-K clustering reflect most of the degrees and relationships of imagination.

The second localized feature strategy is integrated prediction (IP). This approach combines three different types of information for prediction. The first type of information is the correlated association rules. Generally, the association rules imply frequent relationship among data with different variables. However, in terms of the relationships among imaginative thought processes, the derived association rules cannot offer any information about the variables. For considering the relationships among imaginative thought processes, we employ a correlated concept along with association rules to prune meaningless association rules.

The second type of information is regarding the forecast using linear regression. Generally, association rules cannot cover all possible worlds, which are the combinations of all questions. As a result, we can use linear regression for forecasting the score of a question by the previous highly correlated questions. The third type of information comprises the frequent scores obtained by DBSCAN. The frequent scores are representative guides for displaying imagination. If a question cannot be predicted from the previous two information entries, we use the frequent scores as a substitute.

In order to cultivate the imagination of college students and give the necessary training courses, we focus on measuring freshmen's and sophomores' imagination for analysis by using forecast with clustering and integration (FCI). The imaginative evaluations are performed with an imagination scale proposed by Lin and Tsau (2012). The scale has been validated with high reliability for college students, and contains four components, namely initiation, fluency, flexibility, and originality. Lin and Tsau (2012) considered that imagination belongs to divergent thinking and can be defined seriously as four portions derived from the theoretical basis of IERG's five understandings (Cant, 2012). Therefore, in accordance with the sequential processes of initiation, fluency, flexibility, and originality, the degrees, relationship, and thought processes can be easily revealed if students answer the questions on the scale sequentially.

Our contributions include the following. First, we adapted mining technologies to concentrate on the special features of imagination in a spare dataset. Second, we contributed strategies for co-analyzing correlations and association rules to reduce meaningless relationships. Third, we proposed a model to comprehend the imaginative processes of questions that students answered for sequential input data. Finally, we evaluated students' imagination and predicted the degree of imagination with the aforementioned evaluation of students' imagination by using FCI.

In this paper, we introduce localized feature concepts for analyzing imagination to determine the degrees and relationships of each component of imagination and predict the trend of the students' thought processes in a sparse dataset. The balance of this paper is organized as follows. We first briefly review the literature on imagination, approaches for evaluating imagination, and mining algorithms. Then, an overview of the proposed approach and the design of the scale used herein are presented. After that, the experiments and related results are discussed and

Table 1. Summary of the use of notations

Symbol	Descriptions
eps	The neighborhood radius for DBSCAN clustering
$minpts$	The minimum number of data required to form a dense region for DBSCAN clustering
N	The number of data in the dataset
$Sup(Pat)$	The frequency/count of pattern Pat
$RSup(Pat)$	The ratio of the frequency/count of pattern Pat to the number of data N
$Pat1 \Rightarrow Pat2$	The association rule regarding $Pat1$ implies $Pat2$
$C_{Pat1 \Rightarrow Pat2}$	The confidence of the association rule regarding $Pat1$ implies $Pat2$
$\eta_{MinAsso}$	The minimum threshold defined by the supervisors for association rules
X_i	A variable X_i in the dataset
$cov(X_1, X_2)$	The covariance between variables X_1 and X_2
$\sigma(X_i)$	The standard deviation of variable X_i
ρ_{X_1, X_2}	The Pearson's product moment coefficient from variables X_1 and X_2
R_{X_3, X_1, X_2}	The multiple correlation coefficient of the independent variables X_1 and X_2 , and the dependent variable X_3
$h_{\theta}(X_i)$	The dependent variable in linear regression
θ_i	The coefficient of the variable X_i
Q_i	A question in the scale where $1 \leq i \leq 20$
$Q_i(\text{score})$	A presentation of the score that students get in the question Q_i
$(Q_i, Q_{(i+1)}, \dots)$	The scores modeled in a vector form

compared to those obtained without using localized features. Finally, we provide the conclusion. In the following, we use the same notations for the consistency in the study. The summary of the notations is shown in [Table 1](#).

LITERATURE REVIEW

For clarity, we briefly review the literature on imagination and then review mining and statistical technologies sequentially. The definition of imagination is first discussed in this section. Before analysis, imagination shall be defined clearly because it can be misinterpreted easily as other concepts such as fantasy and creativity. Second, we recall the DBSCAN and FPM technology for extracting localized features. Finally, for localized features and predictions of students' imagination, a few statistical approaches—correlations and linear regressions—are presented.

Imagination

Imagination is a native ability of human beings, the basis of creative activities, and the result of cognitive and emotional processes. By the operation and exercise of imagination, Human beings not only construct scientific theories but also create new inventions to improve life through the process of constant thinking, trial and error. For example, in the 19th century, some people imaged eagerly they could fly in the sky. This imagination promoted artists, scientists and engineers to improve with various of past experiences and trial experiments to invent airplanes. Therefore, imagination is the key of innovation.

The definition of imagination is broad and vague. Generally, imagination is usually misinterpreted by the public because of its similarity to creativity. However, the meanings of imagination are not the same as the ones of creativity. The gyroscope theory proposed by Lin and Tsau (2012), who extended and redefined the activities related to the development of new ideas about physical objects or psychological feelings from the Vygotsky theory (Lindqvist, 2003; Vygotsky, 2004), demonstrated the difference between creativity and imagination. Its concept is shown in [Figure 1](#). The gyroscope figure shows the definition of creativity. It has eight elements, namely, stimuli from database, fluency, flexibility, originality, crystallized imagination, techniques, elaboration, and evaluation. The top four elements belong to convergent thinking, and the bottom four elements belong to divergent thinking. When human beings desire to create new concepts or objects, they first diverge to various lines of thinking and then arrive at the target concept by using their special insights. Such divergent thinking is called imagination. After a series of improvements and considerations of experience and intelligence, divergent thinking can converge into an extreme idea for new concepts or objects. The process of combining divergent and convergent thinking is creativity. For example, if students consider improvements and techniques, the thoughts are already limited and reduced with realistic considerations. The thoughts belong the creativity.

Nowadays, there exist popular imagination models for training and defining students' imagination such as IDEAL training models (Wang et al., 2010), IDV (Ho et al., 2013), and IFFO (Lin & Tsau, 2012). In the first model, the researchers considered the imagination can be formed in four components, including initiation, development, alternatives and links. The initiation component is the same as the stimuli from database that we mentioned in the

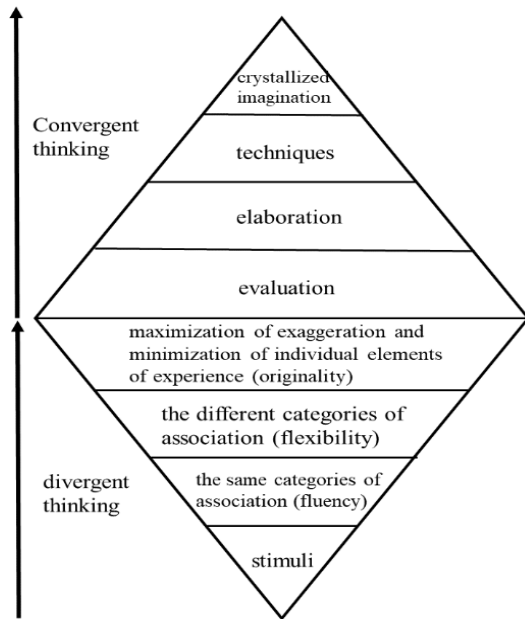


Figure 1. Gyroscope theory (Lin & Tsau, 2012): Mechanisms arousing imagination and creative activity

previous paragraph. Development implies the association from any initiative thoughts and then expands thoughts. Alternatives implies reversed thoughts from the goal. Generally, the thoughts from the development may be suspended due to the size of students' databases. To stimulate more thoughts for students, the reversed thoughts are proposed. By considering the thoughts in the different way, the imagination can be inspired. The component of links is the thoughts that connect from initiative/developed thoughts and alternative thoughts. By these four components, imagination can be trained.

In the IDV model, the imagination can be described in the three stages, namely initiation stage, dynamic adjustment stage and virtual implementation stage. In initiation stage, the property of the thoughts is the same as the ones that we mentioned in IDeAL. The next stage is the dynamic adjustment stage. The key concept of the stage focuses on the association which can be briefly divided in two parts. One part is related to how to envision relationships among thoughts; that is, students should link related thoughts, extend the concepts behind thoughts, or identify the contradictions between thoughts and reorganize them accordingly (Ho et al, 2013); the other part is the quality which involved in conferring new meanings on a thought within an associative network to transform it into a novel thought. As for the virtual implementation stage, the key concept focuses on refining particular thoughts.

In the last model, IFFO, the imagination can be combined by four components. Those are initiation, fluency, flexibility, and originality. The first one component is the same as the one in IDeAL. Both of them emphasize the stimuli from database. As for the fluency, it is partially similar to the second component in IDeAL. In fluency, the authors stressed on the thoughts which belong to the same property (e.g., students inspired clocks from watches) or category (e.g., students inspired flowers from trees). As for the thoughts with different properties and categories, the authors think that such thoughts should be classified into the component, flexibility. By distinguishing the fluency and flexibility, the thoughts can easily be defined. The last component, called originality, is the special point in imagination because it is an ability to generate a product or thought with unique or unusual, unexpected, and first of its kind. In comparison with IDeAL and IDV, the last component, originality, emphasizes that students should combine all experience in initiation and think in association to create new thoughts which are different from others. With originality, the thoughts will be crystallized in imagination.

According to the above imagination models, for promoting students' imagination and precisely measuring imagination, we adopt the model, IFFO, proposed by Lin and Tsau because the model disassembles imagination more meticulous than the other models. We also consult some creativity scales such as creativity assessment packet (Williams, 1980). Unfortunately, questions in creativity scales should be verified with the imagination aspects and reliability tests and may not be suitable for IFFO. We thus use the scale proposed by Lin and Tsau who designed evaluation questions related to divergent thinking and roughly generalized the questions into IFFO. The scale concentrates on the degrees in divergent thinking and does not limit students' thoughts because the questions in the scale are designed in an open-ended manner. By analyzing the scores of these four components, the degree of imagination can be evaluated.

Data Analysis

In terms of data analysis, data mining analyses and statistical research are two popular techniques. In general, statistical research uses the cores/algorithms of data mining. The only difference is the behavior of data analysis. Data mining is the process of automatically discovering implicit information from masses of data, whereas statistical research focuses on the verification of hypotheses defined by researchers. For exploring data, data mining techniques may serve as a guide for supervisors to provide interesting information and situations before creating a hypothesis.

Generally, there exist many mining approaches, such as classification, clustering, FPM and ARE (Aher & Lobo, 2011; Slater et al., 2017). However, using only one single approach for data analysis cannot provide accurate results. For obtaining realistic results, the improvement with the combination from two or more approaches is necessary. For example, Hahsler and Karpienko (2017) extracted association rules with hierarchical clustering. For revealing accurate students' imagination in the study, we adopt some mining approaches and statistical approaches, and thereby combine these approaches to analyze and predict students' imagination. To accurately predict students' imagination, we review some mining and statistical approaches which can assist with concentrating on localized features.

The clustering approaches fall under the category of mining analysis, and they are usually adopted to extract preliminary difference of data information for supervisors. The main purpose of clustering is to differentiate the degrees of data from unknown categories by aggregating them into several groups. Supervisors can then focus on clusters of interest.

DBSCAN (Khan et al., 2014) is a well-known clustering approach, and it uses the density model to aggregate data. Unlike the centroid-based clustering model, DBSCAN provides stable clustering results based on two factors because it does not consider the initial locations of centroids. The two factors are called neighborhood radius (*eps*) and minimum number of data required to form a dense region (*minpts*), where *eps* represents the degree of data similarity in a region and *minpts* denotes the lower bound of cluster size. By setting the values of these two factors, clustering results can be generated steadily.

Example 1: Consider the students' scores on two questions, A and B, and then model the scores as a vector (*a*, *b*). The vectors of scores are as follows: (12, 8), (12, 9), (4, 3), (4, 3), (4, 3), (4, 3), (3.5, 3), (2, 2), (2, 2), (2, 2), (2, 2), (2, 4), (1, 5), and (1.5, 6). Assume the values of *eps* and *minpts* are 5 and 3, respectively and the Euclidean distance is used for the similarity calculation. After DBSCAN clustering, the results can be aggregated into two clusters, CLUSTER1 and CLUSTER2. CLUSTER1 contains two vectors, (12, 8) and (12, 9), because the distance between these two vectors is less than *eps*. The remained vectors belong to the CLUSTER2 because in each vector, the distance between the vector and its nearest vector is smaller than *eps*. After clustering, the verification would be done. Obviously, CLUSTER1 will be treated as a noise because the number of data in CLUSTER1 is smaller than *minpts*. As a result, CLUSTER2 is the only on cluster after DBSCAN clustering.

In addition to DBSCAN, FPM is used to look for interesting information in data analysis. To filter meaningless information, supervisors can extract interesting information from patterns/itemsets by comparing the support of patterns (*Sup(.)*) with a user-defined threshold, called the minimum support (*MinSup*). These patterns, which are very frequent in a dataset, are called frequent patterns, and they can be described as follows. Assume a pattern *Pat* is generated from any combinations of all variables, *X* where $X = \{X_i \mid i = 1, \dots, m, \text{ and } m \text{ is the number of variables}\}$. As a result, $Pat \subseteq X$ and $\exists X_i \in Pat$. *Pat* will become a frequent pattern if the support of *Pat*, (*Sup(Pat)*), that is the frequency(/count) of *Pat* in the dataset, is equal to or larger than *MinSup*. Many approaches have been proposed for obtaining frequent patterns by using possible worlds such as the Apriori approach (Inokuchi et al., 2000) or by using tree-based approaches such as the frequent pattern tree approach (Han et al., 2004). These approaches can yield the same frequent patterns.

ARE is a rule discovery technique to extract frequent interesting relationships of appearances from frequent patterns. Based on the confidence value, rules that appear very frequently can be treated as association rules. The association rules can be calculated using relative support *RSup(.)*, which is the proportion of the number of data in the dataset that contains the item, and can be expressed as follows. Consider patterns *Pat1*, and *Pat2* in a dataset and the number of data is *N*, and $Pat1 \neq Pat2$. The rule of $Pat1 \Rightarrow Pat2$ implies that if *Pat1* appears, *Pat2* appears as well. To verify whether the rule $Pat1 \Rightarrow Pat2$ is an association rule, the confidence of the rule ($C_{Pat1 \Rightarrow Pat2}$) should be calculated with relative support (*RSup(.)*) by the using following formula:

$$C_{Pat1 \Rightarrow Pat2} = \frac{RSup(Pat1 \cup Pat2)}{RSup(Pat1)} \quad (1)$$

$$RSup(Pat1) = \frac{Sup(Pat1)}{N} \quad (2)$$

If $C_{Pat1 \Rightarrow Pat2}$ is higher than the minimum threshold ($\eta_{MinAsso}$) defined by the supervisors, the rule is an association rule.

However, in the real world, the process of ARE may be complicated for the variable-oriented data. As mentioned previously, all patterns only combine from variables. If a variable is a random variable, patterns will be difficult to be obtained. To simplify the process of ARE, researchers used a concept of the quantitative association rule (Salleb-Aouissi, Vrain, & Nortet, 2007) for categorical and quantitative data. The kernel of the concept is that all values in a variable are discrete and countable. Therefore, a pattern can be organized from any variables with values and variable X_i will be yield in a pattern at most once.

Example 2: Consider the situation from Example 1, and supervisors set the values of the *MinSup* and the $\eta_{MinAsso}$ are 3 and 0.6 respectively. According to the description of the frequent patterns, the frequent patterns in *Pat1* which is yield by the variable A are scores 2 and 4, and the ones in *Pat2* which is yield by the variable B are scores 2 and 3. After calculating the frequent patterns, the association rules can be derived from these frequent patterns. According to Formulas (1), (2) and the concept of the quantitative association rule, the rules are $X_A(4) \Rightarrow X_B(3)$, $X_A(2) \Rightarrow X_B(2)$, $X_B(3) \Rightarrow X_A(4)$, and $X_B(2) \Rightarrow X_A(2)$ where the confidences of the rules are 1, 0.8, 1 and 1, respectively, and all of the confidences are larger than 0.6.

In addition to mining analysis, statistical analysis is also widely used in data analysis. Correlation analysis is used to quantify the relationships between two variables. It has two differences compared to association rule analysis. Causality is the main difference between these two analysis methods. Generally, correlation analysis is concerned only with the relationships among variables. That is, if variable X_1 and variable X_2 have high correlation, the relationship only objectively expresses the degree between X_1 and X_2 . By contrast, association rule analysis strengthens causality in variables such as $X_1 \Rightarrow X_2$, which implies that if X_1 appears, X_2 appears as well. Besides causality, purpose is the other difference between the two analysis methods. Association rules refer to more generalized terms from a dataset, whereas correlation can be considered as a special case of association that implies a linear degree of dependence between the variables.

Correlations between/among variables are evaluated by determining Pearson's product moment coefficient or the correlation coefficient R , which are common approaches for providing correlation coefficients, ρ , to measure the relationship between two variables and among multiple variables, respectively. According to the values of ρ and R , the relationship can be revealed easily. The definitions of ρ for two variables and multiple correlation coefficient of R for multiple variables are given in Definitions 1 and Definitions 2, respectively.

Definition 1 (Pearson's Product Moment Coefficient) Given variables X_1 (independent variable), and X_2 (dependent variable), the Pearson's product moment coefficient ρ_{X_2, X_1} is equal to

$$\rho_{X_2, X_1} = \frac{cov(X_1, X_2)}{\sigma_{X_1} \sigma_{X_2}} \quad (3)$$

where $cov(X_1, X_2)$ is the covariance between X_1 and X_2 , and $\sigma(\cdot)$ is the standard deviation.

Definition 2 (Multiple Correlation Coefficient) Given variables X_1, X_2 (independent variables), and X_3 (dependent variable), the multiple correlation coefficient $R_{Z, XY}$ is equal to

$$R_{X_3, X_1 X_2} = \sqrt{\frac{R_{X_1 X_3}^2 + R_{X_2 X_3}^2 - 2R_{X_1 X_3} R_{X_2 X_3} R_{X_1 X_2}}{1 - R_{X_1 X_2}^2}} \quad (4)$$

and the correlation coefficients $R_{X_1 X_2}, R_{X_1 X_3}, R_{X_2 X_3}$ are equal to $\rho_{X_1, X_2}, \rho_{X_1, X_3}$, and ρ_{X_2, X_3} , respectively.

Regression analysis is a statistical process for estimating the relationships among dependent variables and independent variables. It is the best function to summarize associations among data and then easily predict the value of the dependent variable from the model. In simple linear regression, the model can be denoted as follows:

$$h_\theta(X_i) = \sum_{i=0}^n \theta_i X_i, X_0 = 1 \quad (5)$$

where $h_\theta(X_i)$ is the dependent variable, θ_i is the coefficient of the variable X_i , and X_0 is a constant. By calculating the linear regression and then resolving the coefficients θ_i , the predicted function can be obtained easily, and thus, supervisors can predict the dependent value by using independent variables.

PROPOSED MODEL AND SCALE DESIGN

For analyzing imagination, we measure imagination by exploring frequency scores and implicit generalized association rules of imagination components, namely, initiation, fluency, flexibility, and originality. To weaken the effect of noise on frequent patterns and association rules, we proposed localized feature strategies to highlight common features and to display some interesting relationships among different variables. Clustering is the first

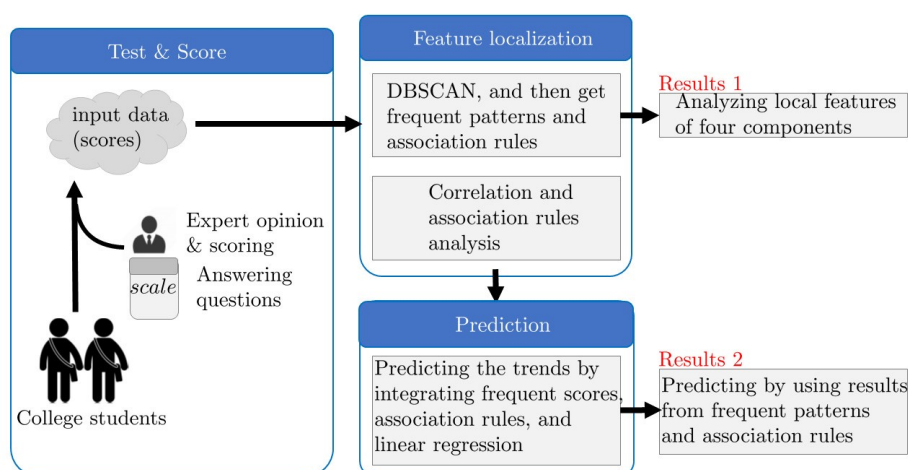


Figure 2. Diagram of proposed model

Table 2. Questions in scale

	Format	Questions (What do you think when you see.....)	Purpose
Initiation (Q ₁ —Q ₄)	Literature	- a starfish - a cloud - a leaf in autumn - a tree	Verify the degree of student’s initiation by making them answer all questions within 10 min.
Fluency (Q ₅ —Q ₈)	Geometry graphics	- two parallel lines - a triangle - a square - a diamond	Verify fluency based on the number of symbolic graphs drawn within 10 min.
Flexibility (Q ₉ —Q ₁₂)	Geometry graphics	- the number 9 - the number 3 - letter Y - letter D	Verify flexibility based on drawings of geometric graphs of different categories within 10 min.
Originality (Q ₁₃ —Q ₂₀)	Geometry graphics	Eight questions with abstract geometric graphics	Verify the degree of originality based on geometric graphs drawn within 10 min.

strategy in this study because it prevents noise data from having any effect on frequent patterns and association rules. In addition to clustering, we also pruned association rules by using correlations to explore meaningful rules for thought processes. In the following, we explain the proposed model in detail.

For the sake of clarity in the proposed model, we analyze imagination in three phases, namely, the test and score phase, feature localization phase, and prediction phase, according to the flowchart is shown in Figure 2. In the test and score phase, the purpose is to collect students’ scores by having them answer the questions on the imagination scale. For comprehending the degrees and thought processes of students’ imagination, students were asked to answer the questions one by one and within a short period (e.g., 150 s). The questions, Q_is, where $i \in \mathbb{Z}^+ \mid 1 \leq i \leq 20$, on the scale are shown in Table 2. In the scale, questions in initiation are represented in literature, such as “What do you think when you see a starfish?”. As for the remainder questions in the other three components, the questions are illustrated in geometric graphics. After the students finished answering all questions, all scores were determined using the guide given in Table 3, which contains the scoring criteria for each question defined by referring to experts’ opinions.

To comprehend students’ imaginative qualities for development in the future, we first used two concepts of localized features to analyze scores that students obtained in the second phase. The first concept is regarding the use of the clustering technique to classify the data into several degrees. Generally, frequent patterns and association rules show the frequency of patterns and rules in a dataset. However, without the use of localized features, this information may confuse supervisors, leading them to misjudge. For strengthening the features of various groups and avoiding the effect of noise, the clustering technique shall be executed before analysis.

Clustering approaches can rapidly aggregate students with similar degrees of imagination into clusters. For decreasing the drawback in determining the number and locations of clusters in advance, we adopted the density-based approach, DBSCAN, for clustering. As explained previously, the DBSCAN approach uses *eps*, and *minpts* to

Table 3. Scoring criteria for imagination

	Format	Scoring criteria
Initiation	Literature	The initiation score depends on the number of answers filled in by the student. If a student filled in many answers without duplicates, s/he would have a high initiation score.
Fluency	Geometric graphics	The fluency score is arrived at based on the number of answers drawn by a student. Akin to initiation, the fluency score is proportional to the number of answers.
Flexibility	Geometric graphics	The degree of flexibility score relies on the number of categories drawn by a student. The greater the number of categories (i.e., different macro taxonomies) drawn by a student, the higher is her/his score. (e.g., if a student drew a flower and a car, both should be grouped into different categories. If a student drew a flower and a tree, both should be grouped into the same category— plant.)
Originality	Geometric graphics	For measuring originality, we pre-investigated the answers of 800 other students with ages similar to the students in the present study and then used the answers as an originality guide. The degree of originality score depends on the uniqueness of the answers of a student, which is decided based on whether they are in the guide. Therefore, students would get high scores if they drew the items that are not in the guide.

aggregate data into clusters. To determine the optimal value of *eps*, we selected a “knee” for *eps* by using a *k* nearest neighbor distance histogram (Kurumalla & Rao, 2016a), where a knee corresponds to a threshold at which a sharp change occurs along the *k*-distance curve.

The other concept in the feature localization phase pertains to the integration of association rules and variable correlations for preparations in the prediction phase. The association rules mainly highlight frequent relationships among variables by using conditional probability, called confidence, to reveal the implicit common information in a dataset. However, in a sparse dataset, plenty of association rules are occasionally confusing supervisors determining policies in the real world. For decreasing the number of association rules, we use correlation analysis to indirectly determine meaningful association rules.

For increasing the number of meaningful localized features of imagination, we use the strength of correlation to prune association rules by co-analyzing correlations and association rules. That is, if variables in an association rule have a strong correlation, the association rule will become a confidence rule for predictions in the future. By contrast, if two variables have non-correlation or weak correlation, the association rules containing these two variables describe only the generalized features with frequency and cannot be confidence rules for predictions. Therefore, these association rules can be ignored, even though there exists a generalized relationship between the two variables.

For calculating the correlations between two variables from association rules such as $X_1 \Rightarrow X_2$, the correlation coefficient can be calculated easily by Pearson’s ρ . As for complex association rules that contain variables in the leftmost of the right arrow, such as $X_1X_2 \Rightarrow X_3$ and $X_1X_2X_3 \Rightarrow X_4$, the correlation coefficient can be calculated by multiple correlation analysis. For simplifying the processes of multiple correlation, we evaluate the *R* value as correlation coefficient from the value of R^2 in multiple linear regression.

According to the strength of correlation, the association rules can be distinguished into two types, namely, (1) non-correlation or weak correlation ($0 \leq R \leq 40\%$), and (2) moderate or strong correlation ($R > 40\%$). If the variables in an association rule fall under type (1), the implicit information displays that the association rule is meaningless because the variables in the rule may be “independent” or there might be “similar amounts of negative and positive data.” As for type (2), these associations will be meaningful and confident association rules because the variables have strong correlations. Therefore, the meaningless association rules with the correlation coefficient $R \leq 40\%$ can be pruned.

In the last phase, we predict the degrees and thought processes of imagination by integrating previous association rules, frequent scores, and results of linear regression analysis. Prediction is the kernel phase that occurs after the feature localization phase to forecast imagination-related information. However, this prediction requires activation by the most principal starting point. Here, we used the question Q_1 , which belongs to the initiation, as the starting point. The reasons for this selection are given below. First, owing to extension in thinking (i.e., thinking would become rich in the same component when students answered more of the same type of questions.), most degrees and thought processes were determined when the students answered the first Q_1 . By contrast, obtaining a frequent score on Q_1 would be representative of imagination. Second, given that the students answered the questions sequentially, Q_1 is the first question that the students came across. Finally, Q_1 belongs to the initiation, which is the base of the three components. Therefore, all predictions will start from Q_1 .

Table 4. Example: frequent scores, association rules and correlations after feature localization

Localized features	Values
Frequent Scores in Q_1 – Q_4 after using DBSCAN	(4, 3, 4, 3)
Rules and Correlations	Q_1 and Q_2 belong to the high correlation, Q_1 and Q_3 belong to the high correlation, Q_2 and Q_4 belong to the low correlation. $Q_1(3) \Rightarrow Q_2(4)$, $Q_1(4) \Rightarrow Q_3(6)$, $Q_2(4) \Rightarrow Q_4(5)$

According to remainder of the association rules after correlated pruning, the degrees and thought processes of imagination can be predicted from Q_1 with the following priorities:

1. Predict the score of Q_i by checking the association rules from the component to which Q_i belongs
2. Predict the score of Q_i by checking the association rules from two components, where one of the components contains Q_i
3. Predict the score of Q_i by using linear regression analysis with $Q_j(s)$, where $j < i$ and Q_i and Q_j belong to the same component
4. Set the score from the frequent scores.

For example, consider four questions $Q_1 - Q_4$, where $Q_1 - Q_3$ belong to a component, and Q_4 belongs to another component. For analyzing the imaginative thought processes, students answered these questions sequentially, and the score of each question is denoted as $Q_i(\text{score})$. After the analysis of localized features, information about frequent scores, association rules, and correlation is shown in **Table 4**. The number of predictions of scores from $Q_1 - Q_4$ is three because three association rules are involved in Q_1 . These three predictions can be calculated as follows. In the first prediction, the score of Q_1 is set to three based on priority because there exists an association rule $Q_1(3) \Rightarrow Q_2(4)$ with strong correlation. According to the previous association rule, the scores of Q_1 and Q_2 are three and four, respectively. Q_3 is the special case in the prediction because it cannot be predicted by the third priority according to the rule $Q_1(4) \Rightarrow Q_3(6)$, where the score of the Q_1 is four. As a result, we will use the third priority to calculate the score of Q_3 by multiple linear regression with Q_1 and Q_2 . If R corresponds to weak or no correlation, the score of Q_3 would be equal to four based on the fourth priority; otherwise, the score of Q_3 would be equal to m , which is the value predicted by using multiple linear regression with the scores of Q_1 and Q_2 . As for the last question, Q_4 , the score cannot be obtained from the association rule $Q_2(4) \Rightarrow Q_4(5)$ because Q_2 and Q_4 have weak correlation. Therefore, the score of Q_4 will be determined using the fourth priority, and the prediction of $Q_1 - Q_4$ in vector form can be denoted as $(3, 4, 4 \text{ or } m, 3)$.

In the second prediction, the scores of Q_1 and Q_3 can be predicted and are equal to four and six, respectively by using the association rule $Q_1(4) \Rightarrow Q_3(6)$. The score of Q_2 can be determined using the third priority because (1) Q_2 cannot be supposed from any association rule and (2) Q_1 and Q_2 are strongly correlated. As for Q_4 , its score can be obtained by using the fourth priority. As a result, the prediction of the $Q_1 - Q_4$ in vector form can be denoted as $(4, m, 6, 3)$.

In the third prediction, the score of Q_1 is set to four according to the fourth priority because the score of Q_1 cannot be obtained from the first three priorities. As for the score of Q_2 , even though $Q_1(3) \Rightarrow Q_2(4)$ and Q_1 and Q_2 are strongly correlated, it cannot be predicted from the association rule $Q_1(3) \Rightarrow Q_2(4)$ because there exists no appropriate score of Q_1 in the association rules to predict the score of Q_2 . However, the score of Q_2 can still be predicted from Q_1 by linear regression because Q_1 and Q_2 are strongly correlated. The score of Q_3 can be obtained easily and is equal to six from the association rule, $Q_1(4) \Rightarrow Q_3(6)$ because the score of Q_1 equals to four, and Q_1 and Q_3 are strongly correlated. As for the score of Q_4 , it follows the fourth priority and is equal to three which is the frequent score of Q_4 . The first reason is Q_4 is the first question in another component; therefore, the first priority is not suitable for Q_4 . The second reason is that Q_2 and Q_4 are weakly correlated, and therefore, the score of Q_4 cannot be predicted from the association rule $Q_2(4) \Rightarrow Q_4(5)$. The last reason is Q_4 and $Q_1, Q_2,$ and Q_3 belong to different components. In the light the last reason, the score of Q_4 cannot be predicted by multiple linear regression from $Q_1, Q_2,$ and Q_3 . Therefore, the prediction of Q_1 to Q_4 in vector form is denoted as $(4, m, 6, 3)$.

In the prediction phase, there may be plenty of trends connected to the scores of the questions because many correlated association rules that contain Q_1 may generate different trends. For providing more information for analysis, we illustrate each trend with a probability as a weight. Unfortunately, the probabilities of the scores from frequent patterns and linear regression are difficult to measure. For simplicity, we are concerned only with the probabilities of confidence that arise from the association rules. The probability of each trend can be modeled from the product confidence. By using these probabilities, supervisors can make any decision using different trends.

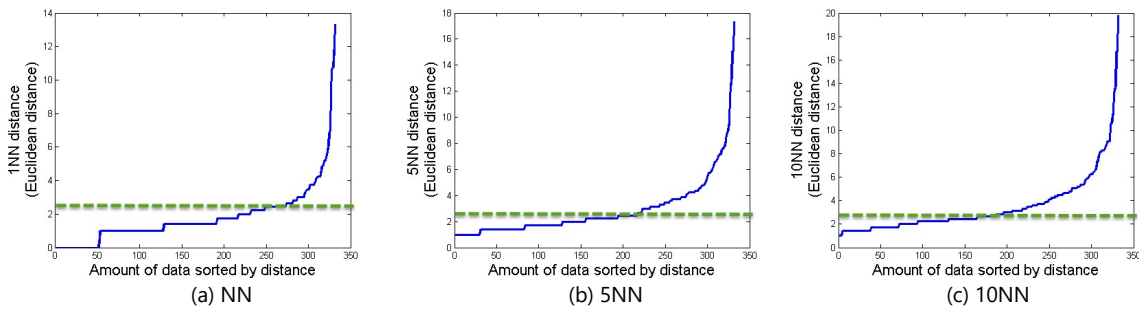


Figure 3. Values of eps in initiation component for different values of k

Table 5. Optimal eps values of different components

	Initiation	Fluency	Flexibility	Originality
eps	2.4	1.5	1.7	1

EXPERIMENTS

For revealing students' imagination and reducing misinterpretations owing to the various degrees in the data, we used localized features, clustering, and co-analysis between correlations and association rules for analyzing imagination components before extracting frequent scores and association rules. Owing to the use of clustering, the imagination scores will show the frequencies of similar degrees, and thereby supervisors can easily obtain less impure frequent scores and association rules. In addition to clustering, we used the mechanism of supporting the frequent scores with predictions. By using association rules pruned by correlation analysis of components, linear regression analysis, and frequent scores, the trends of students' imagination can be obtained easily. For simplifying the tasks, we experimented with students' imagination input data in MATLAB 6.2 on the Windows 7 platform.

Datasets

We performed a proficiency assessment of imagination on freshmen and sophomores at two universities in Taiwan. One consists of the engineering institute and the college of design; the other consists of the college of arts. The number of data points was 332, and the backgrounds of the students included computer science, civil engineering, electrical engineering, fine arts, business management, industrial design & creative design, architecture, and electronics engineering. In these data, the gender ratio was unbalanced; therefore, an analysis of the gender aspect was beyond the scope of our study.

Experimental Setup

DBSCAN requires two parameters, namely, $minpts$ and eps . We set $minpts=5$ as the lower bound for a cluster for less computing cost and avoiding the cases of generating lot of noise data (In our experiments where $K=1$, when $minpts < 229$, the results are the same); clusters having fewer than $minpts$ data items would be viewed as noise. With respect to eps , we used the kNN distance histogram (Kurumalla & Rao, 2016b) which is a popular approach determining the optimal values of eps in DBSCAN. The different values of k do not affect the eps because the value of eps determined this way depends on k , but it does not change dramatically as k changes. For example, the values of eps in the initiation component are shown in Figure 3. According to the kNN distance histogram, the optimal values of eps for the different components listed in Table 5 can be obtained. For simplifying the imagination analysis, we only listed results from top-K clustering, where $K=1$.

Localized Features and Predictions

For highlighting the frequent scores and association rules of students that have similar degrees of imagination, we first considered two types of strategies – using DBSCAN (localized aspect) and non-localized features (global aspect) to analyze each component and the relationship between two components. In the frequent score analysis, we defined various levels for displaying the degrees of the components according to the scores. The levels include zero (score =0), low ($1 \leq \text{score} \leq 3$), moderate ($4 \leq \text{score} \leq 7$), high ($8 \leq \text{score} \leq 12$), and rich (>12). By revealing the levels of frequent scores and association rules, supervisors can obtain (1) the degrees of the imagination component and the processes of students' imaginative thinking along with the order of questions to describe the students' imagination. After the DBSCAN strategy, we integrated (1) the frequent scores, (2) association rules obtained from correlation

Table 6. Number of clusters and data in various components where K=1

	Initiation	Fluency	Flexibility	Originality
Number of clusters	1	1	2	4
Number of data in the largest cluster	228	290	298	273

Table 7. Comparisons of frequent scores and association rules in initiation ($MinSup=6$ and $\eta_{MinAsso}=0.6$)

	Using DBSCAN	Using non-localized features
Frequent scores (Q_1-Q_4)	(5, 4, 4, 4)	(5, 4, 4, 4)
Association rules (confidence)	$Q_3(2) \Rightarrow Q_4(3)$ (61.9%) $Q_2(1) \Rightarrow Q_3(1)$ (61.5%) $Q_1(4), Q_3(2) \Rightarrow Q_4(3)$ (77.8%) $Q_1(4), Q_3(3) \Rightarrow Q_4(4)$ (75%)	$Q_3(0) \Rightarrow Q_4(0)$ (77.8%) $Q_1(4), Q_3(2) \Rightarrow Q_4(3)$ (77.8%) $Q_1(4), Q_3(3) \Rightarrow Q_4(4)$ (75%)

pruning, and (3) conjecture by linear regression of each component as the degrees and the processes of students' imagination.

To localize the features, we first use the DBSCAN to fetch similar data. **Table 6** shows the information of the number of clusters and number of data in the largest cluster (i.e., displaying the number of data where $K=1$) in various components after DBSCAN. According to the clustering, some data will be pruned from the data set. **Table 7** shows the moderate levels of frequent scores and the most popular association rules in initiation from the data in **Table 6**. Both strategies yielded the same frequent scores and similar association rules, and the association rules $Q_1(4), Q_3(2) \Rightarrow Q_4(3)$ and $Q_1(4), Q_3(3) \Rightarrow Q_4(4)$ in the two strategies can be thought to represent the students' initiation. This is because the scores of the questions in the itemsets related to both strategies are close to the frequent scores in initiation. However, the association rules related to the two strategies are slightly different. In terms of the localized aspect, the association rules reflect some additional information such as $Q_2(1) \Rightarrow Q_3(1)$ and $Q_3(2) \Rightarrow Q_4(3)$ that indirectly reveals the students' initiation strengthened gradually when answering the questions. On the contrary, in terms of the global aspect, the association rules show the students had constant initiation because the additional information was hidden by other students having different degrees. In terms of the localized aspect, our strategy can provide more information to supervisors.

In the fluency analysis, the frequent scores and association rules are similar. In Q_5 , the two strategies reveal that a student has a high level when starting to answer the first question related to fluency. The reasons for this include: (1) the thinking accumulated in initiation explodes when using it for derivation, and (2) the two parallel lines are easy to answer. After the students answered Q_5 , the frequent scores reflected that most students had moderate fluency levels, except on Q_7 .

Although both frequent scores on Q_7 are low after accumulated thinking exploded, the declines of the scores on Q_7 have different meanings. According to the score pertaining to the localized aspect, we infer that most students showed a short burnout period for repaying the thinking because they asked for an advance on their thinking in Q_5 . However, the information about frequent scores may imply students' thought processes were unstable. On the contrary, the scores pertaining to the global aspect show that thinking becomes steady after Q_5 , with the change being unclear because the value of the differences from Q_6 to Q_7 , and Q_7 to Q_8 are equal to 1. Therefore, compared to the scores obtained using non-localized features, those obtained using DBSCAN can represent students' thought status more accurately and precisely.

With respect to the association rules in fluency, both strategies provided two common facts, which are the same to those obtained using non-localized features. One is that the students achieved high levels in any question from Q_6 to Q_8 , and they achieved the high levels in the remainder of the questions pertaining to fluency; the other is that the students achieved low levels on Q_6 and Q_8 or Q_7 and Q_8 , and they achieved low levels in Q_7 and Q_6 , respectively. These two facts indicate that the students' partial thought processes diverge into two levels after answering Q_5 . However, these association rules cannot explain the moderate levels on Q_6 and Q_8 . Therefore, we can infer that most of the students' thought processes remained unsteady after answering Q_5 . The aforementioned inference from the association rules corresponds to the one from the strategy using DBSCAN. Therefore, compared to the strategy that uses non-localized features, the strategy that uses DBSCAN can explain the students' thought processes in the real world. All frequent scores and association rules in fluency are explored from data in **Table 6** and shown in **Table 8**.

Table 9 shows the same frequent scores and similar association rules for students' flexibility in both strategies from data in **Table 6**. After the unsteady thinking process in fluency, the frequent scores converged toward stable status, where the scores are approximately equal to 3. We infer the changes from different types of questions are the main reason for this because different types of questions would stop students from thinking in the same way and break the model of thinking from an infinite thinking cycle.

Table 8. Comparisons of frequent scores and association rules in fluency ($MinSup=6$ and $\eta_{MinAsso}=0.6$)

	Using DBSCAN	Using non-localized features
Frequent scores (Q_5-Q_8)	(8, 4, 2, 4)	(8, 4, 3, 4)
Association rules (confidence)	$Q_6(1) \Rightarrow Q_7(1)$ (64.7%) $Q_7(8) \Rightarrow Q_8(8)$ (100%) $Q_5(8), Q_7(8) \Rightarrow Q_8(8)$ (100%) $Q_6(8), Q_7(8) \Rightarrow Q_8(8)$ (100%) $Q_5(8), Q_6(8), Q_7(8) \Rightarrow Q_8(8)$ (100%) $Q_5(8), Q_6(2) \Rightarrow Q_7(2)$ (77.8%)	$Q_7(8) \Rightarrow Q_8(8)$ (73.3%) $Q_5(8), Q_7(8) \Rightarrow Q_8(8)$ (81.8%) $Q_6(8), Q_7(8) \Rightarrow Q_8(8)$ (72.7%) $Q_5(8), Q_6(8), Q_7(8) \Rightarrow Q_8(8)$ (88.9%) $Q_7(0) \Rightarrow Q_8(0)$ (77.8%) $Q_5(8), Q_6(2) \Rightarrow Q_7(2)$ (70%)

Table 9. Comparisons of frequent scores and association rules in flexibility ($MinSup=6$ and $\eta_{MinAsso}=0.6$)

	Using DBSCAN	Using non-localized features
Frequent scores (Q_9-Q_{12})	(3, 3, 4, 4)	(3, 3, 3, 4)
Association rules (confidence)	$Q_9(2), Q_{10}(2) \Rightarrow Q_{12}(2)$ (83.3%) $Q_{11}(0) \Rightarrow Q_{12}(0)$ (88.9%) $Q_9(2), Q_{10}(1) \Rightarrow Q_{12}(2)$ (66.7%)	$Q_9(2), Q_{10}(2) \Rightarrow Q_{12}(2)$ (83.3%) $Q_{11}(0) \Rightarrow Q_{12}(0)$ (72.7%) $Q_{10}(0) \Rightarrow Q_{11}(0)$ (85.7%) $Q_9(2), Q_{10}(1) \Rightarrow Q_{12}(2)$ (66.7%)

Table 10. Comparisons of frequent scores and association rules in terms of originality ($MinSup=6$, and $\eta_{MinAsso}=0.6$)

	Using DBSCAN	Using non-localized features
Frequent scores ($Q_{13}-Q_{20}$)	(0, 0, 0, 0, 0, 0, 0, 0)	(0, 0, 0, 0, 0, 0, 0, 0)
Association rules (confidence)	$Q_{15}(0) \Rightarrow Q_{16}(0)$ (96.6%) $Q_{15}(0) \Rightarrow Q_{17}(0)$ (96.2%) $Q_{15}(0) \Rightarrow Q_{20}(0)$ (96.2%) $Q_{16}(0) \Rightarrow Q_{17}(0)$ (97.0%) ... (omitting 38 association rules) $Q_{15}(1) \Rightarrow Q_{16}(0)$ (87.5%) $Q_{16}(1) \Rightarrow Q_{17}(0)$ (77.8%) $Q_{15}(1) \Rightarrow Q_{19}(0)$ (75%) $Q_{15}(1) \Rightarrow Q_{20}(0)$ (75%)	$Q_{15}(0) \Rightarrow Q_{16}(0)$ (94.2%) $Q_{15}(0) \Rightarrow Q_{17}(0)$ (93.6%) $Q_{16}(0) \Rightarrow Q_{17}(0)$ (93.9%) $Q_{15}(0) \Rightarrow Q_{20}(0)$ (93.0%) ... (omitting 39 association rules) $Q_{15}(4) \Rightarrow Q_{16}(0)$ (87.5%) $Q_{14}(2) \Rightarrow Q_{17}(0)$ (85.7%) $Q_{14}(2) \Rightarrow Q_{19}(0)$ (85.7%) $Q_{15}(4) \Rightarrow Q_{17}(0)$ (75%)

The association rules in the two strategies also support the aforementioned main reason obtained from the frequent scores because the scores on the itemsets in the association rules are approximately equal to the ones in the frequent scores. They show the students had similar frequent scores and also imply that the students answered questions with steady thought processes. Although both strategies reveal the same information about the students' flexibility, the strategy that uses DBSCAN provides richer information than the one using non-localized features.

The frequent scores and association rules in terms of originality, as given in Table 10, are more different than the ones in the three components because the input data are approximately equal to zero. Clearly, in terms of originality, with both strategies, the frequent scores are zero and the association rules are similarly zero. We infer that the lack of originality is the main factor leading to the zero scores. According to the arousing mechanisms of imagination and creative activity, originality is the last component of imagination. It relies very heavily on the other three components, and a long time is required to crystalize and converge thoughts into unique ideas. Therefore, based on (1) the decreasing trend from the scores of three components and (2) the zero scores in originality, we infer that these students lacked originality.

In addition to discussions on the degrees and thought processes of each component, we discuss the effects of interactions of two components from among initiation, fluency and flexibility. Owing to zero scores in originality, we ignore the interactions regarding originality. Table 11 displays the interactions of two components in different strategies. In the DBSCAN strategy, the scores of the association rules in Table 11 are close to the frequent scores in aforementioned three components in Tables 7, 8, and 9. By contrast, these association rules honestly reflect the relationship between two components. According to the number of association rules, we infer that initiation is strongly related to fluency, and fluency is related to flexibility; otherwise, the relationship between initiation and flexibility is weak.

Generally, between the two strategies, there is a difference in these relationships between initiation and fluency, initiation and flexibility, and fluency and flexibility. The difference is regarding the scores on association rules. In the DBSCAN strategy, the association rules in three relationships show similar scores to the frequent scores in the three aforementioned components. As a result, these association rules reflect the relationships that are strongly related to most of the scores of a student. By contrast, when using non-localized features, the association rules

Table 11. Comparisons of association rules between two components in different strategies ($MinSup=6$, and $\eta_{MinAsso}=0.6$)

	Using DBSCAN	Using non-localized features
Initiation vs fluency	$Q_2(3) \Rightarrow Q_6(3)$ (84.2%)	$Q_1(8) \Rightarrow Q_5(8)$ (74.4%)
	$Q_1(5) \Rightarrow Q_5(5)$ (80%)	$Q_1(5) \Rightarrow Q_5(5)$ (61.2%)
	$Q_1(4) \Rightarrow Q_5(4)$ (84.6%)	$Q_3(1) \Rightarrow Q_7(1)$ (68.2%)
	$Q_3(4) \Rightarrow Q_7(4)$ (90.9%)	$Q_1(5), Q_2(3) \Rightarrow Q_5(5)$ (90.9%)
	... (omitting 19 association rules)	... (omitting 23 association rules)
	$Q_3(2), Q_4(3) \Rightarrow Q_7(2)$ (85.7%)	$Q_1(8), Q_4(8) \Rightarrow Q_5(8)$ (85.7%)
$Q_4(3), Q_5(4) \Rightarrow Q_8(3)$ (85.7%)	$Q_2(3), Q_3(2) \Rightarrow Q_7(2)$ (85.7%)	
$Q_1(4), Q_4(3) \Rightarrow Q_5(4), Q_8(3)$ (85.7%)	$Q_2(6), Q_3(3) \Rightarrow Q_6(6)$ (75%)	
$Q_1(5), Q_2(3) \Rightarrow Q_5(5), Q_6(3)$ (75%)	$Q_2(6), Q_4(6) \Rightarrow Q_7(3)$ (66.7%)	
Initiation vs flexibility	$Q_2(3) \Rightarrow Q_9(2)$ (100%)	none
	$Q_4(3) \Rightarrow Q_9(2)$ (100%)	
Fluency vs flexibility	$Q_5(4) \Rightarrow Q_{10}(3)$ (60.7%)	$Q_6(2), Q_8(3) \Rightarrow Q_{12}(2)$ (68.8%)
	$Q_8(3) \Rightarrow Q_{12}(2)$ (60.7%)	$Q_5(4), Q_7(3) \Rightarrow Q_{10}(3)$ (62.5%)
	$Q_6(2) \Rightarrow Q_{12}(2)$ (71.4%)	$Q_6(5), Q_7(3) \Rightarrow Q_{10}(4)$ (81.82%)
	$Q_6(4) \Rightarrow Q_{10}(3)$ (68.4%)	$Q_6(2), Q_7(1) \Rightarrow Q_{12}(2)$ (66.7%)
	... (omitting 12 association rules)	$Q_5(3), Q_6(3) \Rightarrow Q_{11}(2)$ (77.8%)
	$Q_5(5), Q_8(3) \Rightarrow Q_{12}(2)$ (100%)	$Q_5(3), Q_6(2) \Rightarrow Q_{12}(2)$ (66.7%)
	$Q_6(4), Q_8(4) \Rightarrow Q_{10}(3)$ (75%)	$Q_5(3), Q_8(2) \Rightarrow Q_{10}(2)$ (85.7%)
	$Q_6(2), Q_7(1) \Rightarrow Q_{12}(2)$ (75%)	$Q_6(2), Q_8(2) \Rightarrow Q_{11}(2)$ (66.7%)
	$Q_6(4), Q_7(3) \Rightarrow Q_{10}(3)$ (85.7%)	$Q_6(2), Q_8(2) \Rightarrow Q_{12}(2)$ (66.7%)
		$Q_6(4), Q_7(2) \Rightarrow Q_{10}(2)$ (66.7%)

Table 12. Association rules with correlation pruning

	Using DBSCAN and correlation pruning
+Initiation	$Q_2(1) \Rightarrow Q_3(1); Q_3(2) \Rightarrow Q_4(3)$
	$Q_1(4), Q_3(2) \Rightarrow Q_4(3); Q_1(4), Q_3(3) \Rightarrow Q_4(4)$
Fluency	$Q_6(1) \Rightarrow Q_7(1); Q_7(8) \Rightarrow Q_8(8)$
	$Q_5(8), Q_6(2) \Rightarrow Q_7(2); Q_5(8), Q_7(8) \Rightarrow Q_8(8)$ $Q_6(8), Q_7(8) \Rightarrow Q_8(8); Q_5(8), Q_6(8), Q_7(8) \Rightarrow Q_8(8)$
Flexibility	$Q_{11}(0) \Rightarrow Q_{12}(0)$
	$Q_9(2), Q_{10}(1) \Rightarrow Q_{12}(2); Q_9(2), Q_{10}(2) \Rightarrow Q_{12}(2)$
Initiation vs Fluency	$Q_1(3) \Rightarrow Q_5(3); Q_1(4) \Rightarrow Q_5(4); Q_1(5) \Rightarrow Q_5(5)$
	$Q_2(3) \Rightarrow Q_6(3); Q_2(4) \Rightarrow Q_6(4); Q_3(1) \Rightarrow Q_7(1)$
	$Q_3(2) \Rightarrow Q_7(2); Q_3(3) \Rightarrow Q_7(3); Q_3(4) \Rightarrow Q_7(4)$
	$Q_4(2) \Rightarrow Q_6(3); Q_4(2) \Rightarrow Q_8(2); Q_4(3) \Rightarrow Q_5(4)$
	$Q_4(3) \Rightarrow Q_8(3); Q_4(4) \Rightarrow Q_8(4); Q_4(5) \Rightarrow Q_8(5)$
	$Q_1(4), Q_4(3) \Rightarrow Q_5(4); Q_1(5), Q_2(3) \Rightarrow Q_5(5)$
	$Q_1(4), Q_2(3) \Rightarrow Q_6(3); Q_1(4), Q_3(2) \Rightarrow Q_5(4)$
	$Q_1(4), Q_4(3) \Rightarrow Q_8(3); Q_1(5), Q_2(3) \Rightarrow Q_6(3)$
	$Q_2(3), Q_3(2) \Rightarrow Q_7(2); Q_3(2), Q_4(3) \Rightarrow Q_7(2)$
Initiation vs Flexibility	$Q_4(3) \Rightarrow Q_9(2)$
Fluency vs Flexibility	$Q_6(2) \Rightarrow Q_{12}(2); Q_7(1) \Rightarrow Q_{12}(2); Q_7(3) \Rightarrow Q_{10}(3)$
	$Q_7(4) \Rightarrow Q_{10}(3); Q_8(3) \Rightarrow Q_{12}(2); Q_8(5) \Rightarrow Q_{10}(3)$
	$Q_5(4), Q_7(3) \Rightarrow Q_{10}(3); Q_5(5), Q_8(3) \Rightarrow Q_{12}(2)$
	$Q_6(2), Q_7(1) \Rightarrow Q_{12}(2); Q_6(2), Q_7(2) \Rightarrow Q_{11}(2)$
	$Q_6(2), Q_7(2) \Rightarrow Q_{12}(2); Q_6(2), Q_8(3) \Rightarrow Q_{12}(2)$
	$Q_6(4), Q_7(3) \Rightarrow Q_{10}(3); Q_6(4), Q_8(4) \Rightarrow Q_{10}(3)$
	$Q_7(1), Q_8(3) \Rightarrow Q_{12}(2)$

display the relationship regarding $Q_5(8)$. In addition, there exist no rules in the relationship between initiation and flexibility. Therefore, the strategy using DBSCAN contributes more meaningful information than the one using non-localized features.

For obtaining precise association rules, we filter the association rules based on the second localized feature – correlations. To increase the association rules, we prune the association rules in which the correlation of the questions with the association rules are equal to or greater than 0.4. Therefore, the association rules in DBSCAN can be pruned and the results are shown in **Table 12**. These association rules provide the prediction for the next phase.

The degrees of imagination are listed in **Figure 4** for comparison with two different strategies, namely *non-localized features*, *DBSCAN*, and *FCI*. The degrees of the imagination in non-localized features, *DBSCAN*, and *FCI*

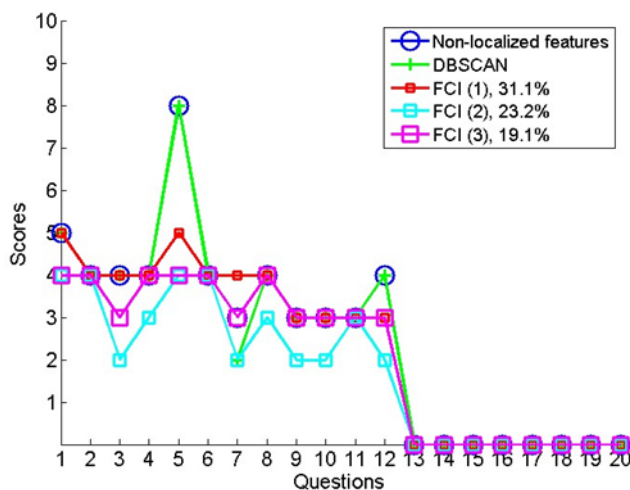


Figure 4. Comparisons with three strategies for predicting students' imagination

can be expressed in terms of the global frequent scores without using any localized feature; frequent scores using DBSCAN clustering; and scores predicted by combining the concept of association rules, linear regression, and frequent scores.

The trends of imagination that originate from the scores of $Q_1 - Q_{20}$ in two strategies – non-localized features and DBSCAN – display similar evolutions of components. The students under these two strategies had moderate initiation and fluency, low flexibility, and zero originality. As for the FCI approach, there exist three predictions denoted by square marks of different sizes in the experiments. The red one (FCI (1)) used the fourth priority to Q_1 for the prediction; and the purple and cyan ones (FCI (3) and FCI (2)) used the first priority of initiation to Q_1 for prediction. All three FCI predictions show decreasing trends in terms of thought processes.

The three predictions made using FCI are different from the ones made using the two strategies. In FCI (1), one of the predictions has higher weight (31.1%) than the other two predictions and most of the scores obtained with FCI (1) are similar to the ones with the two strategies. The differences between FCI (1) and the two strategies are the scores on Q_5 , Q_7 , and Q_{12} , and the scores are five, four, and three points, respectively. The scores of Q_5 and Q_7 were obtained by using second priority, and Q_{12} was inferred with linear regression using the variables Q_9 , Q_{10} , and Q_{11} . These three scores imply (1) a slight increasing trend when students started to answer the first questions in fluency and (2) steady thought processes in fluency and flexibility.

In addition to FCI (2), the prediction has medium weight among the three predictions and shows the worst case of localized features in terms of imagination because the scores of initiation, fluency, and flexibility are unsteady. This prediction is considerably different from those made using the two strategies. We inferred that these students may not have concentrated when answering the questions. Although they were unsteady, the scores of initiation, fluency, and flexibility were within the upper bound of the lower levels. As for FCI (3), the prediction is very similar to the one in FCI (1) and has the lowest weight in FCI predictions. Moreover, like FCI (1), it has slight differences with the prediction made using the other two strategies. We inferred these students are similar to the ones predicted in FCI (1). The only difference is they might have a slight weakness on some questions.

CONCLUSIONS AND FUTURE WORKS

In the study, we proposed two strategies using FPM and ARE for analyzing imagination data and provided degrees and thought processes for supervisors. We first used DBSCAN to objectively distinguish imagination input data into several groups to localize the most similar input data. By discussing the differences between the frequent scores and association rules by using non-localized features and the ones using DBSCAN, the degrees and thought processes of imagination can be revealed easily. Besides clustering, we performed a second imagination analysis to provide predictions for supervisors. By means of integration from association rules and the frequent scores generated after DBSCAN, we filtered the association rules with correlations to decrease the number of association rules. According to association rules, frequent scores, and linear regression, supervisors can obtain the most representative degrees and processes of imagination.

By analyzing and observing the trends from representative degrees, supervisors can conclude what the students need to improve the imagination. Based on the conclusion, supervisors can redesign the course. We use [Figure 4](#) as an example. The trend of FCI (1) shows that students have moderate levels in initiation and fluency, low levels in

flexibility, and zero levels in originality. According to those levels, supervisors can elegantly organize imaginative courses to strengthen students' initiation, fluency, flexibility and originality with the weights, 20%, 20%, 25% and 35%, respectively if the total percentage is equal to 100%.

In the future, we would like to explore the improvements on analyzing students' imagination and enhancing their imagination. For the improvement on analysis, we will consider correlation analysis. In the current study, we adopted linear correlation for pruning association rules. However, linear correlation sometimes cannot be satisfied with the current situation. To accurate the results, we will focus on the analysis of nonlinear analysis. For enhancing the students' imagination, some courses of the college of engineering can be redesigned with the feedbacks of students' imagination analysis. This work thus focuses on the course design and imagination evaluation.

REFERENCES

- Aher, S. B., & Lobo, L. (2011). Data mining in educational system using weka. In *International Conference on Emerging Technology Trends (ICETT)*, 3, 20–25.
- Cant, J. (2012). *IERG – A brief guide to imaginative education*. Retrieved from <http://ierg.ca/about-us/a-brief-guide-to-imaginative-education/>
- Eisner, E. (2001). *The educational imagination: On the design and evaluation of school programs* (3rd edition). (pp. 1–389). Pearson
- Ge, J., & Xia, Y. (2016). Distributed sequential pattern mining in large scale uncertain databases. *PacificAsia Conference on Knowledge Discovery and Data Mining*, Springer, pp 17–29. https://doi.org/10.1007/978-3-319-31750-2_2
- Hahsler, M., & Karpienko, R. (2017). Visualizing association rules in hierarchical groups. *Journal of Business Economics*, 87(3), 317–335. <https://doi.org/10.1007/s11573-016-0822-8>
- Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent pattern tree approach. *Data mining and knowledge discovery*, 8(1), 53–87. <https://doi.org/10.1023/B:DAMI.0000005258.31418.83>
- Ho, H. C., Wang, C. C., & Cheng, Y. Y. (2013). Analysis of the scientific imagination process. *Thinking Skills and Creativity*, 10, 68–78. <https://doi.org/10.1016/j.tsc.2013.04.003>
- Huang, C. Y., Kao, Y. S., Lu, H. H., & Wu, M. J. (2017). Curriculum development for enhancing the imagination in the technology commercialization process. *Eurasia Journal of Mathematics, Science and Technology Education*, 13(9), 6249–6283. <https://doi.org/10.12973/eurasia.2017.01062a>
- Inokuchi, A., Washio, T., & Motoda, H. (2000). An apriori-based algorithm for mining frequent substructures from graph data. *Principles of Data Mining and Knowledge Discovery*, 13–23. https://doi.org/10.1007/3-540-45372-5_2
- Khan, K., Rehman, S. U., Aziz, K., Fong, S., Sarasvady, S., & Vishwa, A. (2014). Dbscan: Past, present and future. *Applications of Digital Information and Web Technologies (ICADIWT), 2014 Fifth International Conference, IEEE*, pp. 232–238. <https://doi.org/10.1109/ICADIWT.2014.6814687>
- Kurumalla, S., & Rao, P. S. (2016a). K-nearest neighbor based dbscan clustering algorithm for image segmentation. *Journal of Theoretical and Applied Information Technology*, 92(2), 395.
- Lin, H. H., & Tsau, S. Y. (2012). The development of an imaginative thinking scale. *Imagination, Cognition and Personality*, 32(3), 207–238. <https://doi.org/10.2190/IC.32.3.b>
- Lindqvist, G. (2003). Vygotsky's theory of creativity. *Creativity Research Journal*, 15(2-3), 245–251. <https://doi.org/10.1080/10400419.2003.9651416>
- Safar, A., & Alkhezzi, F. (2016). Students' perspectives of the impact of online streaming media on teaching and learning at the college of education at kuwait university. *Eurasia Journal of Mathematics, Science & Technology Education*, 12(12), 2975–2989. <https://doi.org/10.12973/eurasia.2016.02317a>
- Salleb-Aouissi, A., Vrain, C., & Nortet, C. (2007). QuantMiner: A Genetic Algorithm for Mining Quantitative Association Rules. *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1035–1040.
- Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics* 42(1), 85–106. <https://doi.org/10.3102/1076998616666808>
- Vygotsky, L. S. (2004). Imagination and creativity in childhood (english translation). *Journal of Russian and east European psychology*, 42(1), 7–97. <https://doi.org/10.1080/10610405.2004.11059210>

Wang, H., Chu, H., Huang, J., & Kang, S. (2010). A road to imagination: The ideal training model and its application in college teaching. In *Proceedings of the TERA Conference on Education-Imagination of future education*, Kaohsiung, Taiwan.

Williams, F. E. (1980). *Creativity assessment packet*. Buffalo, NY: D.O.K.

<http://www.ejmste.com>