CHAPTER 11

Identifying gifted students and their learning paths using data mining techniques

S. Bae¹, S.H. Ha² & S.C. Park³
¹Department of Industrial & Management Engineering, HANBAT National University, Korea.
²School of Business Administration, Kyungpook National University, Korea.
³Department of Industrial Engineering, Korea Advanced Institute of Science and Technology (KAIST), Korea.

Abstract

Today, educating the gifted and talented has become an important part of school education. School staff are increasingly aware of this. They develop special programs for the identification of gifted students and create a curriculum for them. In addition, existing gifted education pays too much attention to curriculum with methods such as a curriculum compacting, acceleration, and ability clustering. Currently, the identification of gifted students depends mainly on a simple identification test based on their age. But, the test results do not reveal all the potentially gifted students. In this chapter, we propose a neural network model for identification of gifted students and a framework for web mining in distance education to provide their learning path. With a specially designed questionnaire, we measure the implicit capabilities of giftedness and cluster the students with similar characteristics. The neural network and data mining techniques are applied to extract a type of giftedness, their characteristics, and their learning path.

1 Introduction

At present, formal education provided in schools is structured on the assumption of age-related development and intellectual homogeneity [1]. It further assumes that student’s intellectual, emotional and social development depends on age. It denies
the real existence of difference in intellectual, social and emotional development of students of similar chronological age. Also, gifted education has not been recognized as an important part of education.

In general, gifted education is only limited to a few gifted students because the frequency of its occurrence in the population is very low. Also, the most serious barrier to gifted education is the negative attitudes of teachers, parents of students who are not gifted, and the education policy for a homogenous education system. But, today, these negative situations are slowly changing. As mentioned before, all school staff have increased awareness and knowledge about gifted education and most parents of gifted children want to get a special school program for their children.

Thus, gifted education starts with the identification of potentially gifted students. Traditionally, to identify gifted students, a paper-based test was taken. The participants of the test would solve difficult and complex problems. But the ‘real’ gifted students might not pass the test due to the students’ immature intellectual ability. A big chance is lost.

To avoid this, we developed an easy and simple test for measuring the implicit capabilities of giftedness and built a model for distinguishing gifted students from other students using neural network and data mining techniques. In addition, we identified a giftedness feature and classified the type of giftedness. Then, based on these results, we suggest how to design educational programs considering the type of giftedness and how to provide the feature of giftedness to potentially gifted students.

2 Data mining in education

2.1 Gifted education: a short review

In general, gifted education is composed of three parts. The first part is awareness and knowledge, the next is the identification of gifted students, and the third part refers to implementation of gifted programs, a supportive environmental and differentiated curriculum [1].

Research to identify gifted students is divided into two categories: an explicit approach and an implicit approach [2]. In the explicit approach, researchers gather and analyze a large amount of data from a participant group. They define characteristics of giftedness under the assumption that the group has specific characteristics of giftedness. In addition, it measures a degree of giftedness with the number of questions a student answers correctly. On the other hand, the implicit approach identifies characteristics of giftedness which are not measured from traditional IQ tests by investigating and analyzing implicit things which the general population has. Contents for explaining characteristics of giftedness are easy and simple. However, it is difficult for the approach to be used for measuring characteristics of giftedness because of reliability.

In Queensland, Australia, almost all schools are using identification procedures which include the UNICORN model of identification published by Education Queensland. The model uses the ‘bubble-up’ method. The UNICORN model is composed of four stages [3]. However, it needs a teacher’s observation of the student
Identifying Gifted Students and Their Learning Paths

Michael Sayler developed the Gifted and Talented Checklist for Teachers and Parents. It is the recommended identification tool for Australian state schools [4].

In a gifted education program, five components are considered commonly: thinking, curriculum compacting, subject acceleration, ability clustering, and extension activities [1]. Thinking, which is used for curriculum planning, enables differentiation in content and strategies to match the different abilities and learning styles of individual students. Curriculum compacting is a process used to streamline the regular curriculum. Subject acceleration occurs when a student takes a single subject 1 or 2 years earlier. Ability clustering is used to build ability classes for specific subjects. Lastly, extension activities include various programs such as individual program, think fest, and so on.

2.2 Web mining

Web mining has been used in three distinct ways, such as web content mining, web usage mining, and web structure mining [14]. Of them, web content mining is the process of information discovery from sources across the Web. In recent years it has prompted researchers to develop more intelligent tools for information retrieval, such as intelligent web agents, and to extend data mining techniques to provide a higher level of organization for semi-structured data available on the Web.

Web usage mining is the process of the automatic discovery of user browsing and access patterns from web servers. Organizations, which run distance education sites, collect large volumes of data, generated automatically by web servers and collected in server access logs. Other sources of user information include referrer logs that contain information about the referring pages for each page reference, and user registration or survey data. Analyzing such data can help organizations determine the thinking styles of learners, cross studying patterns across subjects, and effectiveness of a web site structure. It can also provide information on how to restructure a web site to create a more effective web site presence, and shed light on more effective management of collaborative study group communication and web server infrastructure.

Web structure mining discovers the link structures of the Web. They are discovered on the basis of the topology of the hyperlinks. These structures can be used to categorize web pages and are useful to generate information such as the similarity and relationship between different web sites. Web structure mining can be used to discover authority sites for specific subjects and to discover overview sites for the subjects that point to many authorities.

3 Identification of gifted students using neural network and data mining

As mentioned before, identification of gifted students plays an important role in gifted education. Existing methods of identifying gifted students only reveal the

and well-designed checklists which are hard to develop. Michael Sayler developed the Gifted and Talented Checklist for Teachers and Parents. It is the recommended identification tool for Australian state schools [4].

In a gifted education program, five components are considered commonly: thinking, curriculum compacting, subject acceleration, ability clustering, and extension activities [1]. Thinking, which is used for curriculum planning, enables differentiation in content and strategies to match the different abilities and learning styles of individual students. Curriculum compacting is a process used to streamline the regular curriculum. Subject acceleration occurs when a student takes a single subject 1 or 2 years earlier. Ability clustering is used to build ability classes for specific subjects. Lastly, extension activities include various programs such as individual program, think fest, and so on.
degree of giftedness. But, today, the degree of giftedness – such as IQ – is not enough. Other measures of giftedness such as EQ are proposed by many pedagogues and psychologists. In this section we propose a framework for identification of gifted students based on data mining techniques to classify their type of giftedness and each group’s characteristics, as depicted in Fig. 1.

The proposed questionnaire measures various implicit capabilities of giftedness such as scientific attitude, leadership, morality, creativity, challenge, and motivation of achievement. In the third step, we compare gifted students’ data with average students’ data and extract and refine features that distinguish gifted students from average students. The questionnaire and analysis results are stored in the database.

After that, to identify the types of giftedness, we divide the students into several groups with similar characteristics such as interested fields and their capabilities. Next, we identify each group’s features and patterns of each types of giftedness. When the type of giftedness is decided, we classify extracted features of each group into common and distinct features. With these results, we can define characteristics of each type of giftedness.

In the sixth step, we build a model to create a giftedness quotient of the type of giftedness. The giftedness quotient can be used as a measurement for evaluating similarity between students’ characteristics and students’ type of giftedness. In this manner, we could decide a student’s type of giftedness by comparing each student’s giftedness quotients. Finally, we can compile a pertinent learning guide by taking into consideration the giftedness features to develop capabilities of gifted students.

### 3.1 Design of questionnaire

For designing an easy and simple questionnaire for students, we investigated the implicit capabilities of giftedness. Implicit capabilities of giftedness are suggested...
by researchers such as Renzulli [5], Gardner [6], and Clark [7]. In the proposed questionnaire, we performed a pilot testing of the questionnaire to measure implicit capabilities of giftedness.

For verifying each question’s effectiveness, we randomly selected some students in the gifted students’ group and average students’ group and performed the identification test. With a one-way ANOVA test, we examined each question’s power of discrimination. Based on this result, the final questionnaire was refined.

3.2 Clustering and classification of gifted students

The next step is clustering and classification. We divide the survey data into several clusters using the \( k \)-means clustering algorithm. The \( k \)-means clustering algorithm is a method for partitioning \( n \) objects into \( k \) clusters by measuring cluster similarity.

After that, we identify the characteristics of each cluster. In this step, we use a classification tool – C4.5 [8]. In general, classification tools are used to identify characteristics of each cluster and to build a model to predict clusters where unclassified data are classified. Decision-tree-based classification is obtained to a directed graph showing the possible sequences of questions, answers and classification. We can identify features and patterns for distinguishing a student’s giftedness type from the others.

To realize the method, we must examine the most suitable number of type of giftedness as evaluating the \( k \) value which has a high intra-cluster similarity value and a low inter-cluster similarity value in the \( k \)-means algorithm. Also, with C4.5, we can select the features that distinguish one type of giftedness from another.

3.3 Creating a giftedness quotient using neural networks

A giftedness quotient can be a measurement for similarity between students’ characteristics and their type of giftedness. That is, if a student has a high giftedness quotient, we can assume that the probability of the student belonging to the specific type of giftedness is high.

Creating a giftedness quotient using a neural network gives us two advantages. The first is that we can evaluate students’ type of giftedness and distinguish excellent students. The second advantage is that we can measure a significant degree of features in the type of giftedness.

As shown as Fig. 2, we use a back-propagation neural network for creating giftedness quotient. A neural network is a set of connected input and output nodes where each connection has a weight associated with it. We build a neural network model which consists of \( n \) inputs for \( n \) questions, one output value between 1 and 100 and one hidden layer which has \( m \) hidden nodes. The hidden layer has \( n \times m \) weight vectors which connect between each input and hidden node. The appropriate number of hidden nodes is known to double or triple the number of input nodes. In our model, a sigmoid function is used as the threshold activation function and the output node’s activation function uses a linear function. The following is the procedure for building the neural network model for creating the giftedness quotient.
We choose a specific type of giftedness in interested fields such as science, liberal art, and so on. Let us assume that we choose Type X in the science field.

1. We randomly select a part of the survey data and assign a target output value of 100 if the type of giftedness of the cluster that the selected data belongs to is Type X, otherwise 1. As depicted in Fig. 2, target output value for cluster C1 is 100 because the giftedness type of cluster C1 is Type X and that of the others are 1.

2. With a back-propagation algorithm, we reduce a difference between an output value and its target output value until the difference is less than a termination criteria.

3. When training is terminated, we can obtain the neural network model for creating the giftedness quotient. If the giftedness quotient of a student is close to 100, we can classify the student to Type C in a science field.

Based on our model, we can measure a significant degree of the features because we can calculate the variation of the output value when one question is increased or decreased in the input node.

### 3.4 Applications

To implement our neural network model, we developed an online identification system to distinguish gifted students and to identify their type of giftedness. In Fig. 3,
the proposed system performs two types of test; a general and a specific test of identification of gifted.

When a participant logs in the system, a general test of identification begins. Based on his or her answers, we compute a participant’s giftedness quotient and view participant’s position. If participant’s giftedness quotient is more than a specific threshold value, we decide that a participant belongs to the gifted student group.

After the general test, the system gives various questionnaires retrieved in the database for identifying a participant’s type of giftedness and computes a giftedness quotient for each giftedness type. Then, it shows characteristics of giftedness type and a learning guide. We applied our identification system to the two giftedness fields, a science and a liberal art field in Korea.

### 3.4.1 General and specific test: identifying of giftedness and their type

To extract the implicit capabilities of giftedness, we organized an advisory committee which consists of scientists, professors, teachers and parents of students. We chose seven implicit capabilities such as scientific attitude, leadership, motivation of achievement, morality, creativity, challenge and general ability. Based on that, we composed 77 questions for the measuring implicit capabilities of giftedness.
Then, we conducted the survey on 130 students in high school, 280 students in the science high school (science field) and 183 students in the foreign language high school (liberal art field). We identified features of ordinary type, features and patterns of all giftedness types related in each giftedness field. We obtained four giftedness types for each field and selected 34 features which explained giftedness types using clustering and classification techniques. We used $k$-means as the clustering algorithm and C4.5 as the classification tool.

We discovered eight giftedness types which have similar patterns (features and their values) and selected features identifying them. As shown in Fig. 4, we could present the pattern of giftedness types with seven feature sets. The ratio of each type means the number of gifted students, who belonged to each giftedness type, to the total number of gifted students in the given field. For example, in the science field, the giftedness Type C is the most common type of gifted students.

We defined eight giftedness types which have different values of feature sets. For example, Type D has high f2 (leadership) and f5 (creativity) values and low f1 (scientific attitude) and f7 (challenge) values. The others are zero.

With this result, we could provide a learning guide that could maximize their creativity and leadership. In addition, we could stimulate their scientific attitude and challenge and encourage their motivation of achievement, morality and general ability.

<table>
<thead>
<tr>
<th>Fields</th>
<th>Giftedness Type</th>
<th>Giftedness Pattern</th>
<th>Ratio of each Type to Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>Type A</td>
<td><img src="image" alt="Pattern" /></td>
<td>34.8</td>
</tr>
<tr>
<td>Science</td>
<td>Type B</td>
<td><img src="image" alt="Pattern" /></td>
<td>22.13</td>
</tr>
<tr>
<td>Science</td>
<td>Type C</td>
<td><img src="image" alt="Pattern" /></td>
<td>38.93</td>
</tr>
<tr>
<td>Science</td>
<td>Type D</td>
<td><img src="image" alt="Pattern" /></td>
<td>3.69</td>
</tr>
<tr>
<td>Liberal Art</td>
<td>Type E</td>
<td><img src="image" alt="Pattern" /></td>
<td>42.86</td>
</tr>
<tr>
<td>Liberal Art</td>
<td>Type F</td>
<td><img src="image" alt="Pattern" /></td>
<td>18.18</td>
</tr>
<tr>
<td>Liberal Art</td>
<td>Type G</td>
<td><img src="image" alt="Pattern" /></td>
<td>12.99</td>
</tr>
<tr>
<td>Liberal Art</td>
<td>Type H</td>
<td><img src="image" alt="Pattern" /></td>
<td>21.43</td>
</tr>
</tbody>
</table>

Figure 4: Characteristics and patterns of giftedness type (f1: scientific attitude, f2: leadership, f3: motivation of achievement, f4: morality, f5: creativity, f6: challenge, f7: general ability).
3.4.2 Evaluating the results of the identification test

To evaluate the results of the identification test, we compared average students and gifted students in the science field distinguished by our model. As shown in Table 1, the test category is composed of four types of capabilities – memorization, cognition, logic, and evaluation.

We perform a significance test to examine the difference between the gifted students group and the average students group using one-way ANOVA analysis. If the significance test is performed with $\alpha = 0.05$ and the $p$-value is less than 0.001, then the null hypothesis is not accepted. That is, we could say that there exists a significant difference between two groups.

Table 1 shows the results of the ANOVA analysis for the four types of capabilities. Because the $p$-values of the four capabilities are less than 0.001, we could confirm that the difference between the gifted students group and the average students group does exist. That is, it shows that gifted students distinguished by our model have more capabilities than average students in the fields of memorization, cognition, logic, and evaluation.

4 Web mining for extracting learning path

We have developed a framework for web usage mining in distance education, which is presented in Fig. 5. The framework starts with using web server logs and user login information, which is stored, maintained within a distance education system. Because the log cannot solely carry on the complete information for web mining and, therefore, include the difficulty in identification of unique learners as well as learner sessions or transactions, it is necessary to develop a framework to integrate any web information with back-end operational data, such as user login information [9]. Everyone who enrolls for distance education has to register in the system and his/her demographic information is stored in a user demographics database.

Before extracting the access histories of learners, on which the mining algorithms can be run, a number of data pre-processing issues, such as data cleaning and transaction identification, have to be addressed.

The major pre-processing task is data cleaning. Techniques to clean a server log, to eliminate irrelevant items, are of importance for this type of web log analysis. Elimination of irrelevant items can be reasonably accomplished by checking the suffix of a file name in the URL (uniform resource locator). We can remove all log entries with filename suffixes such as gif, jpeg, GIF, JPEG, which indicate graphic files. Identifying individual learners and their sessions can be done relatively easily because the system keeps login histories of each learner. For more details about the problem of transaction identification, refer to [10] and [11].

It is clear that the extracted access histories of each individual learner represent the physical layout of web sites, with web pages and hypertext links between pages, which are just given by distance educators (push-type knowledge structures). Once user access histories have been identified, there are several kinds of access pattern mining, such as path analysis, discovery of association rules and sequential patterns, and clustering and classification.
### Table 1: Results of ANOVA analysis for four types of capabilities.

<table>
<thead>
<tr>
<th></th>
<th>Memorization</th>
<th>Cognition</th>
<th>Evaluation</th>
<th>Logic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean square</td>
<td>F</td>
<td>p-value</td>
<td>Mean square</td>
</tr>
<tr>
<td>Between groups</td>
<td>0.734</td>
<td>52</td>
<td>&lt;0.001</td>
<td>1.371</td>
</tr>
<tr>
<td>Within groups</td>
<td>0.014</td>
<td>0.011</td>
<td></td>
<td>0.018</td>
</tr>
</tbody>
</table>
4.1 Customized education

There are many different types of path analyses for customizing education. The most obvious is a graph representation [12], since a graph represents some relation defined on web pages; a graph with web pages as nodes and hypertext links between pages as directed edges, a graph with edges representing similarity between pages, or creating edges that give the number of users who go from one page to another.

Our work relating to path analysis involves determining most frequent traversal patterns from the physical layout of a web site.

The path analysis is performed from two points of view: aggregate and individual paths. The first part, aggregate paths, includes the process of clustering registered learners. A registration at the onset of learning can capture important personal information (gender, age, ZIP code, and education level) that can be enhanced and mined later on [13].

A web site database, which is especially created for user registration, can be segmented by one of the clustering techniques such as a self-organizing map, also known as a Kohonen feature map, to discover learners with similar characteristics. The resulting learner segments with access histories can be used to determine most frequently visited paths of learners in each segment in a web site (aggregate paths).

Figure 5: Web mining for distance education.

We perform web page traversal path analysis for customized education and web page associations for virtual knowledge structures, which could be formed by learners themselves as they navigate web pages.
Examples of aggregate paths that can be discovered through path analysis are listed in Table 2.

The first aggregate path indicates 52% of learners in a segment with high educational level start at /Information/DataMining/Classification and proceed to a more advanced topic, such as /Information/ArtificialIntelligence/NeuralNetwork. The second aggregate path indicates that 65% of learners with low education, at first, study on a data structure of tree in a subject, ‘Information’. But they go back to the data structures of stack and queue and list for acquiring background knowledge to better understand the tree structure.

The second part of path analysis is about discovering individual paths. It amounts to determining a set of frequently visited web pages accessed by a learner on his/her visits to the server during a certain period of time.

Discovering such aggregate and individual paths for learners engaged in distance education can help in the development of effective customized education. In addition, aggregate paths discovered from World Wide Web access logs can give an indication of how best to organize the educator organization’s web space. The aggregate paths make suggestions on learning sequences to learners who belong to the same segment and share similar characteristics (e.g. education level). Later on, they can facilitate the development and execution of future web space design, such as dynamically changing a particular page (contents, hypertext links) for learners belonging in different segments.

Table 2: Aggregate paths of learners.

<table>
<thead>
<tr>
<th>Count</th>
<th>Aggregate paths</th>
<th>Segment</th>
</tr>
</thead>
<tbody>
<tr>
<td>52%</td>
<td>/Information/DataMining/Classification -&gt; /Information/ArtificialIntelligence/NeuralNetwork</td>
<td>High educational level learners</td>
</tr>
<tr>
<td>65%</td>
<td>/Information/DataStructure/Tree -&gt; /Information/DataStructure/StackQueue -&gt; /Information/DataStructure/List</td>
<td>Low educational level learners</td>
</tr>
</tbody>
</table>

Table 3: Association rules for discovering the virtual knowledge structure.

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Association rules</th>
<th>Segment</th>
</tr>
</thead>
<tbody>
<tr>
<td>45%</td>
<td>(/Information/DataStructure, /Mathematics/Algorithm)</td>
<td>High educational level learners</td>
</tr>
<tr>
<td>53%</td>
<td>(/Mathematics/Derivative, /Help/Mathematics)</td>
<td>Low educational level learners</td>
</tr>
<tr>
<td>34%</td>
<td>(/Bulletin, /Chatting)</td>
<td>All learners</td>
</tr>
</tbody>
</table>
4.2 Virtual knowledge structure

Discovering association rules is to find all associations and correlations among web pages where the presence of one set of web pages in a transaction implies (with a certain degree of confidence and support) the presence of other pages. In doing so, we discover the correlations among references to various web pages available on the server by a given learner or learners in a specific segment. We can find correlations such as those shown in Table 3.

The first rule in the Table 3 shows that 45% of learners who belong to a high educational level segment and access ‘/Information/DataStructure’ also access ‘/Mathematics/Algorithm’. The second rule states that 53% of learners in low education who access ‘/Mathematics/Derivative’ make reference to ‘/Help/Mathematics’. The last rule indicates that 34% of all learners go around ‘/Bulletin’ and ‘/Chatting’ regardless of their educational levels.

The discovery of association rules in web server access logs allows web-based distance educators to identify virtual knowledge structures against push-type knowledge structures and helps in reorganizing web space based on these structures. We call it ‘virtual knowledge structure’ because it includes differences between the physical topology of web spaces and user access patterns.

4.3 Application of web mining to a web-based education system

The distance education system for KAIST, which deals with four principles (information, mathematics, physics, and chemistry), was developed by the KAIST educational center for science education. The system mainly aims at middle school and high school students. It took about a year to develop the system, and it has been fully operational from 1998 (Fig. 6).
Figure 7 reveals several aggregate paths for high-level learners, as they frequently visit some web pages in order. The prototype system takes two parameters as an argument, such as an individual path file and confidence factor, with which sequential patterns discovered have to comply.

Figure 8 depicts the result of mining association rules among web pages for identifying a virtual knowledge structure. It also shows associations, correlations among web pages for the high-level learners.
5 Conclusions

In this paper, we have proposed a neural network model for giftedness identification and a learning path extraction in the distance learning environment using data mining techniques. We measured the implicit capabilities of giftedness with a specially designed questionnaire and classified the students with their types of giftedness. Data mining techniques such as clustering and classification were applied to extract the type of giftedness and their characteristics. The neural network was used to evaluate the similarity between characteristics of students and types of giftedness. The gifted quotient was gained from the trained neural network.

To evaluate our model’s effectiveness, we applied our model to the science and the liberal art field in Korea. The evaluation test results showed that there are various types of gifted students and they have different characteristics. If we can exactly identify students’ giftedness type, we can develop pertinent learning guides for maximizing their strong capabilities and encouraging their weak capabilities.

Discovering aggregate and individual paths for learners engaged in distance education could help in the development of effective customized education, and give an indication of how to best organize the educator organization’s web space. The discovery of association rules could make it possible for web-based distance educators to identify virtual knowledge structures. In order to reveal the possibilities of application of web mining to distance education, we developed a prototype system for web mining, and showed the results of using web mining for educational purposes.

In the future, we will refine our identification model by using various data mining techniques and develop an intelligent learning-guide system actually for potentially gifted students.

References

206 DATA MINING IN E-LEARNING


