Evaluation System For International Student Application Using Nearest-neighbor Case Retrieval Technique

Khine Thuzar Nyunt, Daw Gilmour Hole
University of Computer Studies, Yangon, Myanmar
khinethuzarnyunt@gmail.com,dgmhucsy@gmail.com

Abstract

Important decisions are made based not on the information-rich data stored in databases but rather on a decision maker's intuition because the decision maker does not have the tools to extract the valuable knowledge embedded in the vast amount of data. Data mining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and educational and medical research.

Among the variety of knowledge based approaches to decision support, Case-Based Reasoning (CBR)is increasingly emerging as one of the most promising approaches for complex data rich domains such as education, health and business. The principal method used in the memory is case-based reasoning method which can provide solving new problem by adapting previous solution to similar problems. CBR's cyclical process is used to support enhancing a process's performance of an expert. This method retrieves the appropriate cases from a large set of cases. If the similar between a new case and the retrieved case are very high, the previous solution to that case is returned to users. This system is tested on international student data at AIT's school of engineering and technology. International student data who applied for AIT's school of engineering and technology and who had permission to enter the university are used as old cases. Unknown cases are matched with old cases. If unknown cases and old cases are the same, the system displays accept. Nearest-neighbor case retrieval technique (NNR) is used to find the similarity measure for the cases which are not found exactly in old cases. If the similarity measure is over 0.5, the system displays accept. If not, the system displays reject.

1. Introduction

Case-based reasoning is a problem solving paradigm that in many respects is fundamentally different from other major AI approaches. Instead of relying solely on general knowledge of a problem domain, or making associations along generalized relationships between problem descriptors and conclusions, CBR is able to utilize the specific knowledge of previously experiences, concrete problem situations (cases). A new problem is solved by finding a similar past case, and reusing it in the new problem situation.

J.L. Kolodner (Ed.)et.al, [5] presented a case can be defined as a conceptualized part of knowledge representing past experience in general. The importance of the represented knowledge in a case is because it supplies a wide range of contents in a form of the computer and/ or human readable formats. Case representation encloses a detailed problem description and a detailed solution description. The detailed problem description consists of a new problem and a solved problem description. When the new problem issue arises, the retrieval process identifies the case with the most similar problem description in the past cases. If there is any stored problem description, it represents the description of detailed solution of that case. If it is necessary, adaptation occurs and a new solution is created. Within a case representtation, most types of data can be stored in a case. However it may be difficult to represent large amount of inter -related data. Therefore the functionality and acquisition of information need to be clarified first to decide what should be represented in cases.

Case retrieval is a technique to decide similar case to a source case. The two most widely used techniques of case retrieval are: Nearest-neighbor retrieval and inductive retrieval. Nearest-neighbor retrieval (NNR) is a technique to measure how similar the target case is to a source case. The paper is organized as follows. Section 2 provides related work for the system. Section 3 describes case-based reasoning. Section 4 presents system flow of the proposed system section 5 shows case-based reasoning in evaluating of international student application. Section 6 supports analysis of data. The paper ends with conclusions.

2. Related Work

In today's transnational admission environment, evaluating applicant qualifications is becoming increasingly challenging. Nguyen Thi Ngoc Hien et.al, [4] presented an approach to deriving a case-based retrieval mechanism from the Bayesian network predication model in such a way that the similarity measure used by the case-based system is consistent with the predictive model. The case-based component retrieves the past student most similar to the applicant being evaluated. The Bayesian network model is evaluated using ten-fold cross validation. Using Bayesian networks to predict graduating cumulative Grade Point Average based on applicant background at the time of admission.

Hadkkinen I et.al [7] and Golding P et.al, [8] suggests using student outcome as a good basis to assess applicant's qualifications. Fortunately, AIT has a large database of information on past and current applicants.

Bekele R et.al, [9] used Bayesian networks to predict mathematics performance of high school students. Their model categorized students into three categories: below satisfactory, satisfactory and above satisfactory. The work reported in the present paper differs from their in the highly international nature of the applicant pool and the more fine grained prediction.

3. Case-Based Reasoning

Alamodt. et.al, [1] presented case-based reasoning is a problem solving paradigm that in many respects is fundamentally different from other major AI approaches. It is based on two observations of real world problems. The first is that similar problems tend to have similar solutions. The second is that the types of problems encountered tend to reoccur over time. When the two of this observation hold, it is worthwhile to remember and reuse prior cases.

3.1. Case-based Reasoning Cycle

The processes involved in CBR can be represented by a schematic cycle. To utilize specific knowledge of previously experienced problem situations to solve any problem. It has described CBR typically as a cyclical process comprising the four REs:

- Retrieve the most similar case comparing the cases to the library of past cases by using retrieval algorithm.
- Reuse the case(s) to attempt to solve the problem. Generally the solution of the retrieval case is transferred to the new case directly as its solution case.
- Revise the proposed solution if necessary. The case solution generated by reuse process is necessary when the solution proves incorrect.
- Retain the new solution as a part of a new case by indexing the case for future retrieval.

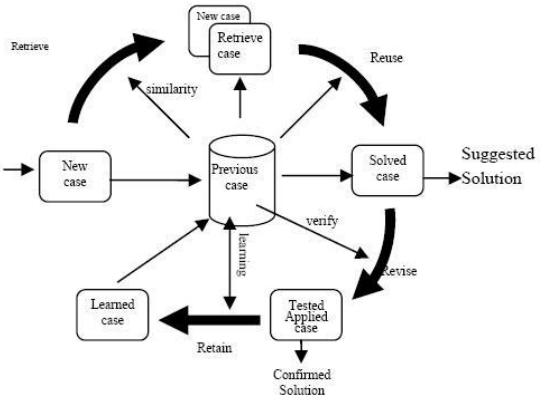

Figure1. Case-based reasoning cyclical process

4. Overview of the Proposed System

A new problem is solved by retrieving one or more previously experienced cases by using nearest neighbor retrieval algorithm, reusing the case in one way or another, revising the solution based on reusing a previous case, and retaining the new experience by incorporating it into the existing knowledge-base (case-base) The four processes each involve a number of more
specific steps, which will be described in the task model. The overview of this system as follows Figure2:

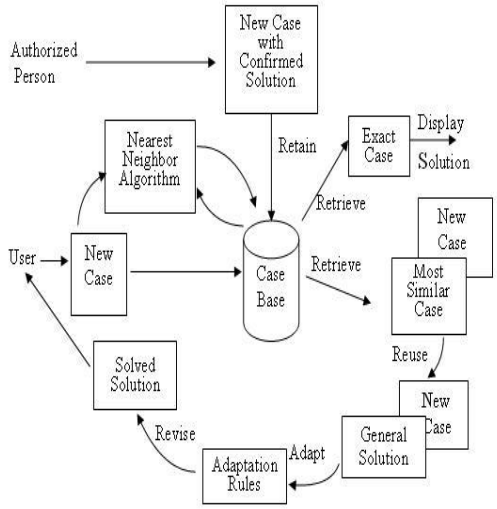

Figure 2. Overview of the Proposed System

4.1 System Flow of the Proposed System

In the start of the system, the user enters the case then the system accepts this case and matching the old cases. If new case is found in old cases, it will display the acceptance of applicant. Otherwise, the system will revise the solution using nearest neighbor algorithm and check this solution whether to retain or not. If retain state, this system save this case in case-base and display the solution. If not, it displays the solution. Then, this system goes to accept state and ready for another process. The system flow is as shown in figure 3.

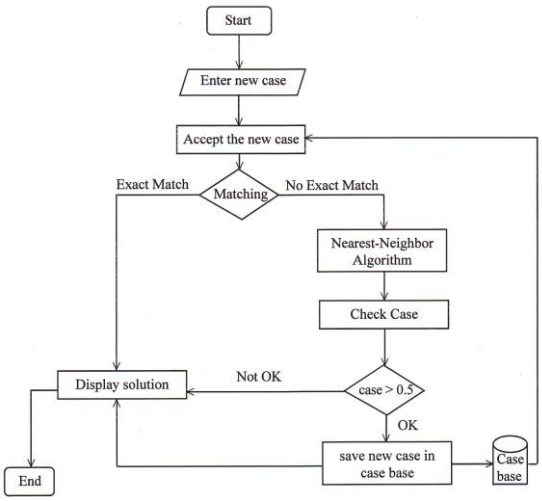

Figure3. System Flow Diagram of the Proposed System

5. Case-based Reasoning in evaluating of international student application

Case Base Reasoning, a method of AI, has been used in this evaluating international application system. As in all CBR based system, the system attempts to produce a solution to new problem by making use of 4R's: Retrieve, Reuse, Revise and Retain. The international student application that the system used is evaluating of Master admission at AIT's school of engineering and technology. The system uses the attributes of each student as the features of the case to store in the casebase.

The case attributes (along with its values) allow to evaluate the similarity between cases in order to retrieve appropriate information. In this system, each case contains 15 attributes which can be used safely for a good evaluating of our chosen attributes. Table 1 shows the index-features of cases which can be used for case retrieval.

Table 1. Index-feature table

Features	Data Type
Institution Rank (H)	Boolean
Institution Rank (M)	Boolean
Institution Rank (L)	Boolean
GPA (H)	Boolean
GPA (M)	Boolean
GPA (L)	Boolean
GNI (H)	Boolean
GNI (M)	Boolean
GNI (L)	Boolean
English Score (H)	Boolean
English Score (M)	Boolean
English Score (L)	Boolean
Professional Qualification	Boolean
Pre-major	Boolean
Field of Study	Boolean

Table 2 shows the unindex-feature of cases which can be used for keeping a record of the student.

Table 2. Unindex-feature table

Features	Data Type
Name	String
Father's Name	String
Passport No	String
Address	String
Sex	String
Marital Status	String
Email	String

5.1. Weight of Features

The weight assigned to each feature of the case tells how much attention to pay to matches and mismatches in the field when computing the distance measures of a case. One way to assign importance values is to have a human expert assign them as the case library is being built. The system uses this way to assign the importance values to different features in the case base.

Table 3. Weight of Feature Table

Features	Weight Point
Institution Rank (H)	5
Institution Rank (M)	4
Institution Rank (L)	3
GPA (H)	5
GPA (M)	4
GPA (L)	3
GNI (H)	4
GNI (M)	3
GNI (L)	2
English Score (H)	3
English Score (M)	2
English Score (L)	1
Professional Qualification	3
Pre-major	3
Field of Study	1

Explanation of weight of feature table

Institution Rank (H) = 1 to 50
Institution Rank (M) = 51 to 100
Institution Rank (L) = > 100
GPA (H) = A to B +
GPA (M) = B to C +
GPA (L) = C to D
GNI (H) = UMC (Upper Middle Income)
GNI (M) = LMC (Low Middle Income)
GNI (L) = LIC (Low Income)

English Score

(H) TOFEL Scores in
Paper-based 101 to 120
Computer-based 261 to 300
Internet-based 621 to 677
IELTS Scores in 7.5 to 9.0
(M) TOFEL Scores in
Paper-based 81 to 100
Computer-based 211 to 260
Internet-based 561 to 620
IELTS Scores in 6.1 to 7.4
(L) TOFEL Scores in
Paper-based 61 to 80
Computer-based 173 to 210
Internet-based 500 to 560
IELTS Scores in 4.5 to 6.0

5.2. Retrieval using Nearest-neighbor Technique

The final goal of a CBR system is to find the case which has the maximum similarity to the input case. The features (attributes) of the input case are assigned as indices characterizing the case. These indices are used to retrieve a similar past case(s) from case memory. The system uses the nearest-neighbor algorithm that finds the closest matches of the cases already stored in the database to the new case. The key thing in nearest-neighbor algorithm is the calculation of an attribute's comparison value for a feature between the previously stored cases and the input case.

A typical algorithm for calculating nearest neighbor matching is:

$$similiary\ \ (T,S) = \frac{\sum_{i=1}^{n} f(T_i, S_i) * W_i}{\sum_{i=1}^{n} W_i}$$

Where,
T is the target case.
S is the source case.
n is the number of attributes in each case.
i is an individual attribute from 1 to n.
f is a similarity function for attribute i in cases T and S.
w is the importance weighing of attribute i.
So, the weight is introduced in the case retrieval and the similarity between cases is considered to be the weighted summation of the similarity between attributes (features). Although each case contains 15 attributes, showing an example of how to calculate the similarity between the new case and the old cases by using nearest neighbor technique.

Table 4. Examples for comparison of cases

	Feature	Weight	New Case	Old Case	Loc. Simil	Wei x Sim
1	Institution Rank (H)	5	Yes	Yes	1	5
2	Institution Rank (M)	4	Yes	Yes	1	4
3	Institution Rank (L)	3	Yes	Yes	1	3
4	GPA (H)	5	Yes	Yes	1	5
5	GPA (M)	4	Yes	Yes	1	4
6	GPA (L)	3	Yes	Yes	1	3
7	GNI (H)	4	No	Yes	0	0
8	GNI (M)	3	Yes	Yes	1	3
9	GNI (L)	2	Yes	Yes	1	2
10	English Score (H)	3	No	Yes	0	0
11	English Score (M)	2	Yes	Yes	1	2
12	English Score (L)	1	Yes	Yes	1	1
13	Professional Qualification	3	Yes	Yes	1	3
14	Pre-major	3	No	Yes	0	0
15	Field of Study	1	Yes	Yes	1	1
	Sum	46				36

T is new Case.
S is old Case.
Nearest Neighbor Formula = sum of (weight* similarity)/ sum of weight
(T,S) = 36/46=0.78

5.3. Reusing the Solution

When the system successfully retrieves the case after careful comparison and matching, the next step is to "reuse" them and produce a result. If the problem completely resembles an existing case, and exact match has been found, the result of the existing case will be used as the final diagnosis without any change. If an exact match is not found, the system reuses the solution of the nearest match as the general solution for the new case.

5.4. Revision of the General Solution

The solution which was presented by utilizing the retrieved cases has to be evaluated according to its level of success. If the solution was fully successful then there remains no need for revision. If however, the solution failed to achieve its required goal, then it is revised. If the solution that was presented is repaired, then the errors which has occurred are given a suitable explanation and stored in the case library.

5.5. Saving New Case

This system is designed to successfully retain novel problems. However, saving new case can only be made by the person who has the authority to maintain the database (casebase). Any case that has the confirmed solution can be saved and that can be successfully used for future retrieval and reuse by the system

6. Analysis of data

Our proposed system supports the evaluating of international students cases concerning master students. The system has trained with 500 cases and tested with set of 200 cases for above stated master student. Firstly, a data set is used to train the system. The system is presented with a set of inputs that have known outputs. By comparing output of the system with the known outputs, we can examine the accuracy of the system. The accuracy is computed by

$$sensitivity = \frac{t - pos}{pos}$$

$$specificity = \frac{t - neg}{neg}$$

$$precision = \frac{t - pos}{(t - pos + f - pos)}$$

$$accuracy = sensitivity \frac{pos}{(pos + neg)} + specificity \frac{neg}{(pos + neg)}$$

Where
t-pos is the number of true positives.
pos is the number of positive.
t-neg is the number of true negatives.
neg is the number of negative.
f-pos is the number of false positives.

The accuracy of the system for evaluating of each student is shown in Table 5.

Table 5. The accuracy for each case

	Master Degree	Accuracy
1	Institution (high)	98.24 %
2	Institution (middle)	97.62 %
3	Institution (low)	95 %

7. Conclusion

This system is focused on developing architecture for the simulation of decision support system by using Nearest-neighbor case retrieval method. The student can enter their background educational information. Our approach using case-based reasoning for predicting applicant performance has the potential to meet this need .The technique can be applied at any institution that has a good database of student and applicant information.

REFERENCES

[1] Alamodt. & Plaza, *Case-based Reasoning: Foundational Issues, Methodological Variations, and System Approaches,* AI Communications, Vol.7 Nr.1, March 1994, pp 39-59.

[2] Chien-Chang Hsu, Li-We Pan, and Cheng-Seen Ho, "Using Induction Trees to Do Case Adaptation in Case-based Reasoning".

[3] Filipic, B. and Junkar, M., *Using Inductive Machine Learning to Support Decision Making in Machine Learning Process,* Computer in Industry,43, 2000, pp.31-41.

[4] Nguyen Thi Ngoc Hien and Peter Haddawy
Computer Science and Information Management Program, Asian Institute of Technology .

A Decision Support System for Evaluating International Student Application, Session F2A.

[5] J.L.Kolodner (Ed.), Case-Based Reasoning, Morgan Kaufmann Publishers: California, 1993.

[6] Rom: Satria Wahono and Behrouz Homayoun Far, "Reasoning with Cases in the CBR System: A Case Study for Applying OOExpert System", *IECI Japan Workshop*, 2000.

[7] Hadkkinen I., "Do Univeristy entrance exams predict academic achievement?", Working Paper Series, Department of Economics, Uppsala University, 2004.

[8] Golding P., Donaldson O., "Predicting academic performance", Proc. 36th ASEE/IEEE Frontiers in Education Conference, 2006, 21-26.

[9] Bekele R., Menzel W., "A Bayesian approach to predict performance of a student (BAPPS): A Case with Ethiopian Student", Proc. IASTED International Conference on Artificial Intelligence and Applications, 2005.