World Academy of Science, Engineering and Technology
International Journal of Educational and Pedagogical Sciences
Vol:9, No:4, 2015

Educase – Intelligent System for Pedagogical Advising Using Case-Based Reasoning

Elionai Moura, José A. da Cunha, César Analide

***Abstract*—**This paper introduces a proposal scheme for an Intelligent System applied to Pedagogical Advising using Case-Based Reasoning, to find consolidated solutions before used for the new problems, making easier the task of advising students to the pedagogical staff. We do intend, through this work, introduce the motivation behind the choices for this system structure, justifying the development of an incremental and smart web system who learns bests solutions for new cases when it's used, showing technics and technology.

***Keywords*—**Case-based Reasoning, Pedagogical Advising, Educational Data-Mining (EDM), Machine Learning.

I. Introduction

ARTIFICIAL intelligence [1] is the study in animals, mankind and machines, of intelligent behavior, and trying find ways to transform this behaviors into any artifact by the engineering.

Into the many process of Artificial Intelligence we have the Machine Learning that can be sectioned into a supervised learning and unsupervised learning. Unsupervised Learning "covers those cases where we have not met the standard, or even if there is a pattern" [1]. Among the various techniques and algorithms used in this learning model, it includes clustering and genetic algorithms, among others.

In other hand, the supervised learning is used, e.g., for solving classification or regression cases, to recognize spam or in the fingerprint recognition. A technique widely used for this type of learning is an algorithm for generating a decision tree**.**

Nagy et al. [2] argue that the learning algorithm with a decision tree was proposed, in first time, by Ross Quinlan as an extension of the ID3 algorithm. They argue that it can handle continuous and discrete attributes and dataset with unknown values of attributes too.

As we can see in this brief introduction, there is a wide range of approaches to machine learning and AI. This work, notwithstanding the various subfields of AI and its algorithms, it focuses on a set of processes and technologies, which are part of the technique commonly known as Case-Based Reasoning (CBR) [3].

We combine this technique with a clustering algorithm called k-Nearest Neighbor (kNN) and the web platform and its technologies in building an expert system where should use the principles and concepts of Artificial Intelligence, to serve as decision support in the pedagogical counseling of students.

Case-based Reasoning (RBC) is a set of techniques for building an intelligent system, with the operations centered at a base of prior knowledge, what is composed of contextualized experiences, describing the problems and their solutions, called case-library. These solutions may be suggested (reused) to be applied to any new cases or problems that are presented, by the similarity relation (retrieve) of the presenting problem with existing cases and can even make use of these new cases (retain) to expand the knowledge base [3].

II. Contextualization

A. Pedagogical Counseling

We noted the need to develop a support tool at the context of educational counseling in interview with educators of Pedagogical Staff of the Federal Institute for Education, Science and Technology of Rio Grande do Norte, in Natal, Brazil.

Early 2012, the service even had a standardized a document for registration of cases, and subsequently created a pedagogical care sheet. Later, they developed a form using Google Docs platform, which generates a shared spreadsheet between members of the crew with all registered information. These people exposed how they work and handle records of cases through recent years, but not yet reach the central problematic of care that is the qualification of the information and the transformation of tacit knowledge of each team member into explicit knowledge of all staff.

The educator says in the interview:

"*The monitoring of students going on by checking the notes at the end of semester, written in the minute book of situations or team schedule (Until 2011). From 2012, we adopted a model of individual records that facilitated the monitoring of the progress of referrals and solutions and allowed to identify recurrences. This model was also important, which brought the possibility of sharing information among pedagogues board about attendances. There has already been a major breakthrough with the cards, but we began to realize that some service situations were quite common (problems requiring psychological counseling, e.g.) and how to check it? How it could work in our favor.*"

Elionai Moura is with the Federal Institute of Education, Science and Technology of Rio Grande do Norte, Natal, CEP 59015-300 Brasil (corresponding author to provide phone: +55 84 9129-2244; e-mail: elionaimc@outlook.com).

José A. Cunha, is Professor at Academic Board of Management and Information Technology at the Federal Institute of Education, Science and Technology of Rio Grande do Norte, Natal, CEP 59015-300 Brasil. He is now in PhD at the Department of Informatics, University of Minho, Braga, Portugal (e-mail: jose.cunha@ifrn.edu.br).

César Analide is PhD in Artificial Intelligence. He is a Head at Department of Informatics, University of Minho, Braga, Portugal (e-mail: analide@di.uminho.pt).

B. Similarity and the Aggregation Problem

A general question facing researchers in many different areas is how to organize observed data into manageable structures, i.e., how to develop taxonomy. In face of this problem, a set of some algorithms is proposed. For all these techniques, well known as Cluster Analysis, mentions to the common characteristic grouping of features or variables.

The concept of cluster is found in the most diverse situations in our day-to-day. For example, in the school environment, there are different groups of students separated by room, or in a restaurant, it can separate people into groups that share same tables. Cluster analysis is a tool for exploratory data analysis that takes different objects and separates them into groups so that the degree of association (similarity) between two objects is maximal if they belong to the same group and minimal otherwise case. Accordingly, we can use cluster analysis to find data structures without resorting to an explanation or interpretation.

An important classification example, using clustering, is that biologists do with mankind, including into groups (clusters) where man is part of primates (low level group), mammals, and animals (high level), for example. Man has greater similarity with chimpanzees (in primates) than with alligators (in animals).

The taxonomy applied in this work uses the descriptions of situations (attributes), in groups (classes) that can be built from this information. The result was the separation into 5 classes of attributes, separated by aspects of the real world and the influence in the academic life of students. Table I lists the classes and the respective weights.

TABLE I
DISCRIMINATION OF CLASSES AND THEIR WEIGHTS, ORDERED BY RELEVANCE

Classes	Weight
Cognitive or psycho educational difficulty	5
Psychological conflicts	4
Relationship/behavior problems in schools	3
Disciplinary problems	2
Difficulty in affective/family relationship	1

We also determined the weight class, 1-5, so that the demands that need more attention by the professional educator is contained in the higher weight classes, while demands considered less relevant belongs to classes that have lower weight. Table II describes all the demands that have been listed (total of 26), which will be calculated attributes in the search for similarities between the cases, relating these attributes with weights given according to pedagogical analysis, in the classification previously determined.

The case-library in this study has 53 cases. Every instance in the case-library is originated from real world experiences. To survey of them, we have the support of experts in tutoring, to undertake the development of a list of demands and solutions applied to major cases, which had knowledge during the 1st semester of the year 2014, at Federal Institute of Education, Science and Technology of Rio Grande do Norte - IFRN.

TABLE II
DISTRIBUTION OF DEMANDS IN THEIR RESPECTIVE CLASSES AND WEIGHTS AFTER ANALYSIS OF EDUCATIONAL SPECIALIST

Classes	Weight	Class
Family conflict	1	E
Parents in separation	1	E
Problem of relationship with mother	1	E
Relationship problem with her father	1	E
Relationship problem at home	1	E
The student or family socioeconomic order problem	1	E
Severe disciplinary order problem	2	D
Mild disciplinary order problem	2	D
Medium disciplinary order problem	2	D
In behavior problem	3	C
Relationship student x student problem	3	C
Teacher x student relationship problem	3	C
Bulling	4	B
Conflict about self-sexual orientation	4	B
Conflict due to emotional or relational situation	4	B
Disincentive for the choice of course	4	B
Disincentive for low income	4	B
Need for secular orientation	4	B
Abusive situation (moral, sexual etc.)	4	B
Situation of exclusion in classroom	4	B
Related to shyness situation	4	B
Constant delays	5	A
Imbalance or psychological problem	5	A
Learning difficulties in discipline	5	A
Many faults	5	A
Need for mentoring	5	A

In defining the case-library, each demand represents an attribute of the case, and, among the listed cases, some have more than one attribute, however, given the limits of our source of information, not all attributes is registered in the case-library.

Those attributes can be evaluated conditionally to be related to the case, and were expressed numerically with 0 (absent) or 1 (present), i.e., if the attribute exists it receives the value 1 (one), if not, the value 0 (zero).

C. K-Nearest Neighbor Algorithm

There followed the development work this process, defining a calculation method of similarity or dissimilarity between the cases, in order to make the retrieval and indexing processes effectively functional. Using the Euclidean distance was determined to calculate the distance between the cases where the distance calculated for each attribute of the cases.

$$D(x, y) = \sqrt{\sum_{a=1}^{n} (x_a - y_a)^2} \quad (1)$$

Equation (1) describes the calculation for Euclidean dissimilarity measures where D is the distance, *x* and *y* are two vectors (cases) with *n* numeric attributes.

The Euclidian distance is a dissimilarity measure, which is defined as a line segment with the shortest distance between

two points, being chosen for this work to be applicable in multidimensional spaces with discrete or continuous values.

The vectors to be used in the calculations of this work are all the requirements of each case registered on the case-library, attributes correspond to the demands detailed in survey and classification as listed in [II].

In principle, the user provides a set of descriptions of the current problem. Then it applies a function to calculate the similarity measure chosen for all instances of knowledge contained in the case-library. Then sort the cases in order of similarity, returning one or more cases, up to the limit set by k, between the nearest.

A clear disadvantage of this approach is the high computational value of the algorithm, it is necessary to compare the data from each new case presented with all the other cases on the case-library, recalculating distances each time it is used.

The main features that lead up to use this algorithm lies in its ease of implementation and ability to adapt to different entries including with multiple dimensions.

The first step is to define the weight of each entry, as classified in Table II. Then it runs through the input vectors, i.e., the case-problem (case) and a case present in case-library (neighbor), each attribute are multiplying by the corresponding weight. It follows then the sum of the squares of the differences of all attributes and finally divides the range of demands and it is the square root calculation, returning its value, i.e., the distance between the cases. Repeat the calculation for each entry recorded in case-library.

III. Web Application

The system architecture can be defined as a description of the organization of the various system components and how them, and the system interact with each other and with the environment in which they live. We selected for the development of this work the architecture client-server, the common web environment.

This architectural model is especially understood as a layered model, which allows understand the role of each layer isolated on the system.

The proposed system is comprised of three layers: 1) user interface - is the layer that users use to interact with the system; 2) business layer - present on the server side logic corresponds to the application domain; 3) persistence layer - this layer application data are stored in a repository or database.

The most superficial layer of the application is the user interface. We developed the interface of this study using a set of documents and forms written using HTML5 and Javascript libraries that can be interpreted by a web browser.

Among the libraries used, there is AngularJS that focuses on manipulating interface via a technology named two-way data binding. This technology allows the interface to be changed simultaneous with changes in the logic of the system, without the need to reload the application to view the new information, and reflect the user's actions on the system logic at runtime [4].

That is, if we correlate the fields of a form to an object that can be manipulated with AngularJS, every interaction the user has to complete this, the object is modified simultaneously, allowing, for example, that the completed data can be checked even before the user decides to terminate the interaction.

Using this technology, the model becomes a SSOT - Single Source of Truth. SSOT refers to the practice of structuring information so that there is only one source of information that given any other way of access to data is given for reference only [4]. Fig. 1 illustrates the operation of two-way data biding.

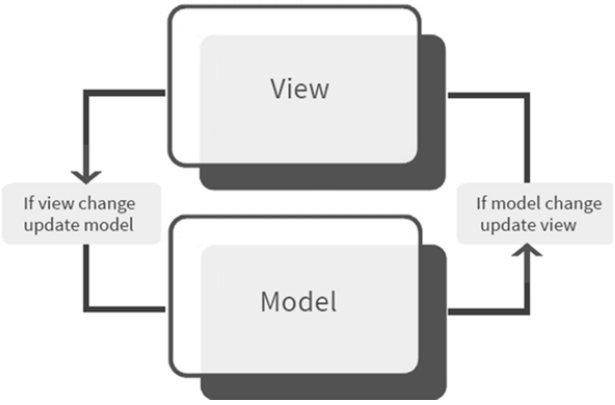

Fig. 1 Two-way data binding in Angular JS library

The second layer, allocated on web-server, is responsible for receiving request from the user interface, process this information and return the relative data response according to the result of processing the request. This middle tier runs in a server environment developed for this purpose by the technology used to process server-side JavaScript called NodeJS.

NodeJS is a processing platform technology that uses event-driven model with the non-blocking I/O, lightweight and efficient. Commonly used in applications where the real time communication and high volume of data and the intense transmission of it is necessary, working as distributed system [5].

Saying the model input and output is non-blocking means that the server when receiving a request will process it and at some point he can answer that request, however, for both, will not fail to receive and respond to other requests that may reaching the same.

One of the features that allows NodeJS be non-blocking is that it is single-threaded, i.e., the application will have only one instance of each process, which, however, possible to create clusters, as in the case of Mongoose library, written in Javascript to facilitate the work of developers who want to connect to the DBMS application with MongoDB.

The third and final layer is the layer of data persistence. A DBMS Manager (System Database) that allows us to store cases optimally and keep the same structure of documents used in the upper layers of the application, doing so with the data suffer the minimal possible change when used and only the necessary changes were made.

As an important factor to CBR, the case-library to be persisted in a database that maintains its structure in the form of collection of documents, and still allows native use of classification algorithms, Map-Reduce etc., in helping to get results so native tends to be highly efficient.

We used to develop the project object of this paper, a NoSQL DBMS, document-oriented, open source and free to use, named MongoDB [6].

The term NoSQL are widely used mostly to describe structures of databases that go beyond relational structure, which can include concepts of tuples, graphs, key-value pairs, tree, etc.

IV. Conclusion And Future Works

The main result, the product of this work was the development of an intelligent decision support system, which has as its target audience of users, pedagogical advisors staff. We developed the system with a set of technologies described in this document, and it is available as open source software under the MIT license.

As a means of provision, and seeking the possibility of future contributions from the community, we chose to create a public code repository using a web platform for developers, the community named github.com. This platform also allows the publication has a code versioning tool, which proved very useful during the development.

Why not be, however, the core of this work, the generation of reports and graphs, but the construction of an intelligent system applied to pedagogical counseling, these tools should be built in later activities, including being able to be exploited by other jobs that have a more statistical approach or focused on the development of this type of system.

The development of Intelligent Systems using CBR has shown that, despite the effort to build a large enough case-library, and the definition of algorithms and techniques for indexing and retrieval more efficient each day, a recurring problem that remains to be solved is the applicability of solutions based on how similar is the case with each other.

This is due to the similarity factor, referring to the knowledge, so as, to bring a set of experiences that may be used in solving the present problem. Since, in most cases, the number of cases is insufficient or inadequate to reflect the solution of the problems, an approach that has a high chance of being more efficient, or to bring appropriate solutions would be the use of algorithms that propose the solutions more useful, instead of suggesting solutions based on the most similar cases.

The main possibilities for expanding this work, in future developments, we point out as essential, to concentrate efforts in seeking a mechanism of inference that allows, the retrieval of cases, the most useful problem situation presented, or even the development of a mechanism for evaluating most recurrent cases or solutions that are most reused.

References

[1] B. R. Whitby, *A.I. A Beginner's Guide*, Oxford: One World Publications, 2003.

[2] H. M. Nagy, W. M. Aly, O. F. Hegazy, *An Educational Data-Mining System for Advising Higher Education Students.* World Academy of Science, Engineering and Technology: International Journal of Computer, Information Science and Engineering Vol.7, No. 10, 2013.

[3] I. D. Watson, *Case-based reasoning is a methodology not a technology*. Salford: University of Salford, 1999.

[4] Google Inc., *AngularJS API Docs*. Google Inc. Accessed in 14/8/2014. Disponible: https://docs.angularjs.org/api

[5] T. J. Fontaine *et al*., *NodeJS Manual & Documentation*. Joyent Inc. Accessed in 10/6/2014. Disponible: https://nodejs.org/documentation/

[6] K. P. Ryan, D. Merriman, E. Horowitz and others, *The MongoDB v2.6 Manual.* MongoDB Inc. Accessed in 14/08/2014. Disponible: http://docs.mongodb.org/v2.6/

[7] I. D. Watson, *Applying Case-Based Reasoning*. San Francisco: Morgan Kauffman Publishers, 1997.

[8] S. Tiwary, *Professional noSQL: a hands-on guid to leveraging noSQL Databases.* John Wiley & Sons. 2010.

[9] D. Crockford, *RFC 4627*. The Internet Society. Accessed in 23/9/2014. Disponible: http://www.ietf.org/rfc/rfc4627.txt