Measuring Similarity of Educational Items Using Data on Learners' Performance and Behavioral Parameters: Application of New Models SCNN-Cosine and Fuzzy-Kappa

Khadidja Harbouche, Chabane Khentout, Mahieddine Djoudi, Adel Alti

IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 28, No. 1, February, 2023, pp. 1-11
Journal homepage: http://iieta.org/journals/isi

Measuring Similarity of Educational Items Using Data on Learners' Performance and Behavioral Parameters: Application of New Models SCNN-Cosine and Fuzzy-Kappa

Khadidja Harbouche[1*], Chabane Khentout[1], Mahieddine Djoudi[2], Adel Alti[1,3]

[1] Computer Science Department and LRSD Laboratory, University Ferhat Abbes of Setif, Setif 19000, Algeria
[2] Computer Science Department and TECHNE Labs, University of Poitiers, Poitiers Cedex 9 86073, France
[3] Department of Management Information Systems & Production Management, College of Business and Economics, Qassim University, P.O. Box 6633, Buraydah 51452, Saudi Arabia

Corresponding Author Email: khadidja.harbouche@univ-setif.dz

https://doi.org/10.18280/isi.280101

Received: 27 December 2022
Accepted: 20 January 2023

Keywords:
Convolutional Neural Networks, cosine similarity, educational items, fuzzy logic, item-to-item similarity, kappa inter raters agreement, Siamese Neural Networks, latent trait

ABSTRACT

Currently, educational systems typically contain a wide range of items drawn from students' learning experiences, questions, problems, and other items. They are collected using modern computer-assisted learning platforms. These items can be grouped into Knowledge Components (KCs) using a similarity measure based on the learner's performance and the number of correct or incorrect answers. This approach leads to wrong assignment of items to the cluster with less interpretation of learners' skills. Consequently, looking for other performance parameters to enhance accuracy has become paramount. The subject's ability to respond to each item called latent trait includes hints, attempts and time response, which can ensure better accuracy and enhance clustering of different items into KCs. In this paper, we propose new similarity models based on combining Siamese Convolutional Neural Network (SCNN), cosine similarity and fuzzy logic systems for Kappa inter-rater agreement. These models aim to achieve efficient item grouping based on learners' performance and behavioral parameters. The proposed approach has been tested on Algebra and French Language datasets to experiment its performances. Experimental findings show the superiority of SCNN-Cosine in terms of clustering analysis measured by ARI, correlation, Calinski-Harabasz and Davies-Bouldin indices, less complexity and faster response, compared to Fuzzy-Kappa.

1. INTRODUCTION

Most course curricula follow a logical progression from relatively simple lessons to more difficult ones. Prerequisite skill frameworks are included in educational systems to help determine the order in which concepts should be taught to students to ensure their success [1]. Many skills have a strong causal link, requiring that the student have to master one before he may move on to another (e.g., a hierarchy of skills according to prerequisites). The prerequisite structure refers to the links between skills that put severe restrictions on the order in which skills can be obtained in an Intelligent Tutoring System (ITS) to sequence learning. It is often produced via a method that involves programmers, learning scientists, and domain specialists [2].

Item-to-skill mappings (also called Q-matrix) are desirable because they allow more interpretable diagnostic information. They also allow for discovering prerequisites among items based on their skills mapping [3, 4]. They are standard representations used to specify the relationships between individual test items and target skills.

Item-to-skills mapping includes two types of approaches: model-based and similarity-based. The model-based approach is used to reduce the dimensionality of the problem and infer the latent factors (e.g., Knowledge Component (KC)) that underlie the items. Since the construction of a Q-matrix from learners' answers is an active research topic; Learning Factor Analysis (LFA) [2] and matrix factorization have been proposed to improve expert-based Q-matrix. The similarity-based approach is based on the assumption that learners will tend to perform similarly on items that require the same skill. This must identify a similarity between pairs of items [5]. It is noted that our approach is similarity-based.

Although it is still difficult to estimate human capacity from stochastic responses, in statistics and related fields. A similarity measure, which is a real-valued function that quantifies the similarity between two objects, has been studied [6]. This method is widely used in cognitive science and psychology, for example, in language acquisition and development, item similarities are automatically detected. Item similarities are the first and necessary step in further analysis such as clustering of the items (e.g., assigning the elements of a cluster to a single KC), which is useful in several ways, with one particular application being learner modeling [7].

Learner models are created using large volumes of data drawn from students' experiences, questions, problems, and other items. They are collected using recent computer-assisted learning platforms. These models offer an estimate of skill level at a particular moment [2]. In tutoring systems, learner models are frequently used to personalize instructions or predict students' performance in the future. Imagine answering a series of multiple-choice questions. For example, consider a homework assignment, or a school entrance examination.

Selecting a response to each question is an interaction between your "ability" (knowledge or features) and the characteristics of the question, such as difficulty, as shown in Figure 1. The subject's ability to respond to each item is called a latent trait. The latent trait is defined as function of obtained a score and made error [8].

In reality, it is important to know that not all questions are created equal. However, some questions can include many technical aspects that are hard to understand while others may test cognitive concepts that are more difficult. Moreover, several experimental studies [9] have been conducted to demonstrate and prove that the probability of responding to an item can be expressed by a logistic function of person's latent trait level and one or more item-characterizing parameters, such as difficulty, discrimination or pseudo-chance parameters [10]. Another aspect is assessing each student's ability to respond questions while unknown, and some external effects may distort his perception, like fatigue and stress, the altered mental state of the subject, emotional state, the tendency to guess, etc. Under these circumstances, it is vital to include these behavioral parameters to ensure effective and efficient items similarities and quality clustering approach.

Most of the research on the item-similarity methods used in educational data mining for clustering items suffered from common limitations, which are: (1) the latent trait is assumed to be fixed during the learning activity that generated the responses; and (2) recent methods based on similarity measures are still very simple and only consider the number of correct responses and neglects many behavioral parameters, such as differences in question quality [7].

To address these two gaps, we aim to propose a latent trait model to measure the items' similarity in the context of educational field by taking into consideration the most important behavioral parameters that enhance components' skills using a new hybrid model. This model is built based on three different models: a fuzzy model, kappa inter-raters agreement, a Siamese Convolutional Neural Network (SCNN) and a cosine method. The main goal of this approach is to ensure high accuracy by considering relevant and discriminatory parameters in qualitative estimation of similarity between items. The main contributions of this work are:

(1) Introducing new behavioral factors such as hints, response time, and number of attempts into the latent trait model. This promotes discrimination between items and achieving good accuracy with quality items clustering.
(2) Developing a new hybrid approach based on a combination of two different — Siamese Convolutional Neural Network (SCNN) combined with the cosine similarity method and fuzzy analysis combined with Cohen's kappa. The proposed approaches consider both learner's performance and behavioral factors that may alter the learner's response, and,
(3) Grouping similar items and ensuring quality clustering using soft computing.
(4) Evaluating the proposed approaches on real-world data.

The remainder of the paper is organized as follows. Section 2 presents related works in the area of similarity-based methods to item-skills mapping. Section 3 details the proposed method based on a fuzzy analysis system, cosine method, and Siamese Neural Network. Section 4 presents the item-skill mapping. Section 5 discusses the experimental results obtained from the proposed approaches. Finally, Section 6 concludes the paper.

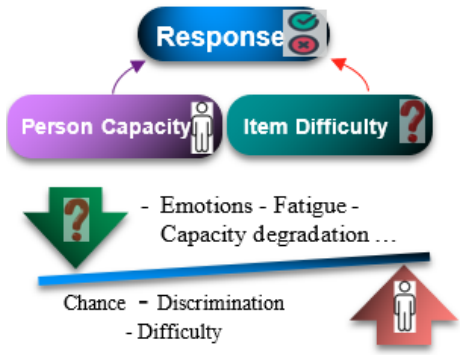


Figure 1. Synoptic of the triadic relation: Person Capacity-item difficulty- response

2. RELATED WORKS

As mentioned above, item-to-item similarity is a crucial phase to perform relevant educational item clustering. Therefore, many approaches and various similarity techniques were proposed in the literature. In this section, we give an overview of recent similarity methods and item-skills mapping approaches.

Rihák et al. [11] applied an automatic computation of item similarities based on learners' performance data using different measures such as Pearson, Yule and Jaccard with different settings and different levels. Through their experimental results, they showed that the Pearson correlation coefficient provides good performance results compared to the other experimented measures.

Dharaneeshwaran et al. [6] calculated the user-item similarity–based on Pearson's and Cosine correlation using three different techniques to recommend desired items. The proposed approach presented high accuracy compared to some existing similarity measures.

For studying performances of similarity measures, a new systematic approach has been proposed in the study [12]. The approach is based on the evaluation of similarity measures. These measures are evaluated in several introductory programming environments.

Nazaretsky et al. [13] proposed a new item-similarity measure named Kappa learning for clustering educational items by identifying similarity between them. It is based on the analysis of learner's response data. The proposed approach outperformed traditional techniques and improved the sequence of the knowledge skills but required a high computation time.

In the research, Mathisen [14] et al. developed a framework using an artificial neural network model to show the main differences between various types of similarity measures.

Overall, there have been little interest and few studies into the deployment of a wider range of methods and criteria in the field of education. At present, similarity measures across the educational data mining community are mostly immature and receive limited attention. Nevertheless, there are a few key research directions that we must follow to help improve the education field. This study attempts to solve two problems: the first is technical with an extrinsic scope to the latent trait and the second is cognitive with intrinsic item characteristics.

The similarity measure solves the problem of mapping items to skills and covers several aspects such as content classification, clustering, motif discovery, etc. The most widespread measures of similarity related to item-to-item similarity are either Euclidean distance or elastic distance measures (Pearson, Kappa, Yule, Jaccard, Sokal, Fisher, etc.). These measures were evaluated on real or simulated data. However, all these methods are distance-based, which calculates similarity by accumulating the distances between optimally matched pairs and neglects the intrinsic differences of items. In order to solve this drawback, we propose to combine distance-based techniques with Siamese Convolutional Neural Network (SCNN) and fuzzy analysis.

The similarity-based methods show more interest than distance-based methods in educational data mining. These methods ensure item clustering and implicitly assume that the latent trait is fixed during the learning activity that generated the responses. The learners' responses to items are considered to be highly correlated to the same knowledge components. Items are assumed to be exchangeable, and their properties are not part of the model for creating a latent trait estimate. Moreover, when assigning items to clusters, the number of correct answers as a single criterion is not sufficient. We must consider other behavioral characteristics to improve the accuracy rate when clustering items, such as response time, hints, and number of attempts to solve the item. This operation is part of the item mapping techniques since it consists of performing clustering on discriminatory features, which are powerful to any changes within continuous assessments.

3. ITEM-TO-ITEM SIMILARITY MODELS

To efficiently ensure item-item mapping in the educational field, it will be helpful to exploit both behavioral parameters and learner performance data. Each item is solved by a group of learners either $\{L_k\}$.

We keep the following parameters as indicators of the learner's response to item I_i: 1) - the item's response is either correct (C_{ki}= 0/1), or incorrect (I_{ki} = 0/1), 2) - the learner may or may not ask for hints (H_{ki} is the number of hints), 3) - the response time of an item is different from one learner to another and from one item to another (T_{ki} is the response time), and 4) - the learner can have several chances or attempts to answer the item (A_{ki} is the number of attempts). Thus, for each item I_i we can determine the matrix of features as shown in Table 1.

Table 1. Features matrix of the i[th] item

	C	I	H	A	T
L1	C1i	I1i	H1i	A1i	T1i
L2	C2i	I2i	H2i	A2i	T2i
…	…	…	…	…	…
Lk	Cki	Iki	Hki	Aki	Tki

Table 2. Item-to-item similarity matrix

	I_1	I_2	I_3	…	I_n
I_1	**1**				
I_2	sim(I1, I2)	**1**			
I_3	sim(I1, I3)	sim(I2,I3)	**1**		
	…	…	…	…	
I_n	sim(I1, In)	sim(I2, In)	sim(I3,In)	…	**1**

We propose to measure Item-to-Item similarity using three different models. After each model, we obtain an item-to-item similarity matrix as shown in Table 2. A similarity matrix is symmetric, which means that the similarity between item *i* and *j* is the same as the similarity between item *j* and item *i*.

3.1 The cosine method

Cosine similarity measures the similarity between two vectors of an inner product space [15]. This measure is evaluated by cosine for the angle theta (θ) between two vectors and determines whether they are pointing roughly in the same direction. We will use this metric to measure the similarity between two items I_i and I_j characterized by their features matrix as shown in Table 1.

1. We start by calculating the vector V_x for each item I_x: V_x= $(V_{x1}, V_{x2}, V_{x3}, V_{x4}, V_{x5})$ as given by Eq. (1):

$$V_{x1} = \sum_{k=1}^{N} C_{kx} \Big/ N \quad , \quad V_{x2} = \sum_{k=1}^{N} I_{kx} \Big/ N$$
$$V_{x3} = \sum_{k=1}^{N} H_{kx} \Big/ N \quad , \quad V_{x4} = \sum_{k=1}^{N} A_{kx} \Big/ N \qquad (1)$$
$$V_{x5} = \sum_{k=1}^{N} T_{kx} \Big/ N$$

where, N is the total number of learners who answered the item I_x.

2. We calculate the similarity cosine between each pair of items I_i and I_j (Eq. (2)) and keep the result in a similarity matrix as shown in Table 2:

$$sim(I_i, I_j) = \frac{I_i \bullet I_j}{\|I_i\| \bullet \|I_j\|} \qquad (2)$$

where, $\|I_x\|$ is the Euclidean norm of a vector I_x, which is the length of the vector, and it is defined in Eq. (3) as follows:

$$\| I_x \| = \sqrt{I_{x1}^2 + I_{x2}^2 + I_{x3}^2 + I_{x4}^2 + I_{x5}^2} \qquad (3)$$

This measure computes the cosine of the angle between vectors I_x and I_y. A cosine value of zero (0) means that the two vectors are at 90° to each other (i.e., orthogonal) and have no match. The closer the cosine value to one (1), the smaller the angle and the greater the match between vectors.

3.2 The fuzzy analysis combined with the Cohen's kappa coefficient

The fuzzy logic-based distortion effect estimation provides values about the effects of environment on item response quality. As shown in Figure 2, the fuzzy system [16] consists of three main steps:

Step 1: The preprocessing

During this step:

1. For each item I_i, we calculate the average number of hints (AH_i), the average response time (AT_i) and the average number of attempts (AA_i) as given in Eq. (4), using the feature matrix presented in Table 1.

$$AH_i = \frac{\sum_{k=1}^{N} H_{ki}}{N} \ , \ AT_i = \frac{\sum_{k=1}^{N} T_{ki}}{N} \ , \ AA_i = \frac{\sum_{k=1}^{N} A_{ki}}{N} \qquad (4)$$

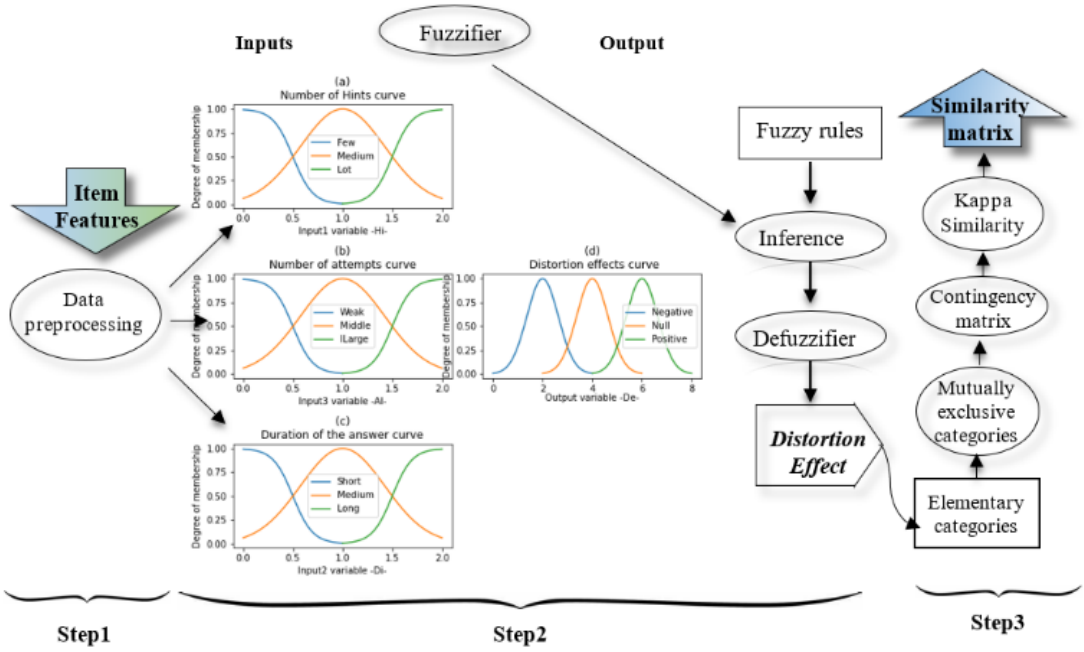


Figure 2. Synoptic of the fuzzy distortion analysis combined to Kappa

where, N is the total number of learners who answered the item I_i.

2. For each learner L_k, we calculate the new values H'_{ki}, T'_{ki} and A'_{ki} as given in Eq. (5).

$$H'_{ki} = \frac{H_{ki}}{AH_i} \quad , \quad T'_{ki} = \frac{T_{ki}}{AT_i} \quad , \quad A'_{ki} = \frac{A_{ki}}{AA_i} \tag{5}$$

3. Finally, we normalize the hints (H'_{ki}), the time response (T'_{ki}) and the attempts (A'_{ki}), by delimiting their values between zero (0) and two (2).

Step 2: The fuzzy distortion effect estimation

Fuzzy logic is an extension of Boolean logic, introduced in 1965 with the proposal of "fuzzy set theory" by Lotfi Zadeh. It gives very appreciable flexibility to reasoning while taking into account imprecisions and uncertainties [17-20].

One of the benefits of using fuzzy logic to formalize human reasoning is that the rules are expressed in natural language. Fuzzy logic is an appropriate way to draw up a distortion effect according to the item or person parameters: hints, number of attempts, and time response.

This step consists of four sub-steps:

Step 2.1 –The fuzzifier: The inputs to the fuzzy logic system to calculate the distortion effect estimation are the number of hints, the answer time and the number of attempts to give a solution.

- The first parameter of the proposed system is H'_{ki} (Hints), Figure 2.a shows the membership function used for this parameter. The membership functions are determined to be "Few" (F), "Medium" (M) and "Lot" (L).
- The second parameter is the duration of the answer or the answer time T'_{ki} that is determined, as shown in Figure 2.b, by the membership functions selected as "Short" (S), "Medium" (M) and "Long" (L).
- The third parameter is the number of attempts A'_{ki} and is determined as shown in Figure 2.c, by three membership functions: "Weak" (W), "Middle" (M) and "Large" (L).

We suggest to distort the response based on the results obtained from three input parameters. The output of the fuzzy logic system is the distortion effect D_{ki} that is shown with three membership functions in Figure 2 (d): "Negative" (N), "Null" (Nl), and "Positive" (P).

The input linguistic values are converted into fuzzy input values using the input membership function f(x) given in Eq. (6).

$$f(x) = \begin{cases} 1/e^{-(x-1)/0.2} & where\ \ x \in [0..2] \\ 1/(1+e^{-10(x-1.5)}) & where\ \ x \in [0..1] \\ 1/(1+e^{10(x-0.5)}) & where\ \ x \in [1..2] \end{cases} \tag{6}$$

The Gaussian membership function is used as membership function for the output, as shown in Figure 2 (d). This function is defined in the universe of discourse X = [0..8] given in Eq. (7):

$$Fc(x) = 1/e^{-(x-\mu)^2/0.8} \tag{7}$$

Step 2.2 –The inference: The fuzzy knowledge base is made up of linguistic variables and fuzzy rules. We use a set of IF-THEN rules. The proposed system consists of 27 rules. Table 3 illustrates some examples of fuzzy rules.

Based on these 27 fuzzy rules and input membership functions, fuzzy outputs are obtained.

Table 3. A sample of the fuzzy rules

Rule#	Rule statement
1.	**IF** NH is F **and** NA is W **and** TR is S **THEN** Effect is N
3.	**IF** NH is L **and** NA is W **and** TR is S **THEN** Effect is N
14.	**IF** NH is M **and** NA is M **and** TR is M **THEN** Effect is Nl
20.	**IF** NH is M **and** NA is W **and** TR is L **THEN** Effect is Nl
25.	**IF** NH is F **and** NA is L **and** TR is L **THEN** Effect is P
27.	**IF** NH is L **and** NA is L **and** TR is L **THEN** Effect is P

The obtained fuzzy outputs are converted to output values using output membership functions. According to 1st rule given in Table 3, if the number of hints is few, the number of

attempts is weak, and the response time is short then the effect is negative. The choice of the fuzzy operators allows us to determine the inference engine, which is generated, in our case using the Min/Max Zadeh operators.

Let *S* be the set of ordered pairs (fuzzy variables, crisp variables). This set is defined as follows:

S = {(NH(F), X1), (NH(M), X2), (NH(L), X3), (NA(W), X'1), (NA(M), X'2), (NA(L), X'3), (TR(S), X"1), (TR(M), X"2), TR (L), X"3), (Effects(N), Y1), (Effects(Nl), Y2), (Effects (P), Y3)}.

We apply the fuzzy rules, quoted above, with the MIN/MAX operators and obtain the following rules:

Y_1 = MAX(MIN(X''$_1$, MAX(X'$_1$, MIN(X'$_2$, MAX(X_1,X_2)) , MIN (X'$_3$, X_1))), MIN(X''$_2$, MAX(MIN(X'$_1$, 1-X_3), MIN (X'$_2$, X_1))), MIN(X"$_3$, X'$_1$, X_1))

Y_2= MAX(MIN(X"$_1$,X'$_2$,X_3) , MIN(X"$_1$,X'$_3$,X_2) ,MIN(X"$_2$,X'$_1$,X_3) ,MIN(X"$_2$, X'$_2$,1-X_1) ,MIN(X"$_2$,X'$_3$,X_1) , MIN(X"$_3$,X'$_1$,X_2) ,MIN(X"$_3$,X'$_2$,X_1))

Y_3= MAX(MIN(X"$_3$,X'$_3$), MIN(X"$_2$, X'$_3$, 1-X_1) , MIN(X"$_1$,X'$_3$,X_3) ,MIN(X"$_3$,X'$_2$,1-X_1) ,MIN(X"$_3$,X'$_1$,X_3))

Step 2.3 – Defuzzifier: in this step, the Mean of Maxima (MOM) method will be applied [20, 21]. The defuzzified value is taken as the element with the highest membership values. If there is more than one element having maximum membership values, we can get the mean value of the maxima as follows:

$$Y = MAX(Y1, Y2, Y3)$$

We calculate the inputs (abscissas) named *A* corresponding to the outputs *Y*, as given by Eq. (8):

$$Y = e^{(A-\mu)^2/0.8} \quad (8)$$

where, $\mu \in \{2, 6, 10\}$.

Let *A* be a fuzzy set with membership function *Y*(*x*) defined over $x \in X$, where *X* is the universe of discourse.

The defuzzified value, let named x^*, of a fuzzy set is defined in Eq. (9):

$$x^* = \left(\sum_{xi \in M} xi\right) / |M| \quad (9)$$

where,

M = {xi| Y (xi) is equal to the height of the fuzzy set A}, and $|M|$ is the cardinality of the set *M*.

Step 2.4 –The output: The result of the last step of the fuzzy system is to interpret the effect of the three parameters hints, attempts and response time on the quality of the response. There are three effects: negative, null and positive. Therefore, for each item I_i, solved by a learner L_k, we keep the following primitive categories: Correct response (C), Incorrect response (I), Positive effect (P), Negative Effect (N) and Null effect (Z).

Step 3: Cohen's kappa item-to-item similarity calculation

The evaluation of Inter-Rater Reliability (IRR) is often necessary for research designs where data are collected through ratings provided by trained or untrained coders (raters) [21]. The assessment of IRR provides a way of quantifying the degree of agreement between two or more coders who make independent ratings about the features of a set of subjects (*people, things, or events that are rated in a study*) [22]. The data must meet the following assumptions to calculate Cohen's Kappa similarity metric.

1. Two categorical outcome variables are ordinal or nominal.
2. Two outcome variables must have the same categories
3. Each subject is rated twice by two independent raters or methods.
4. The same two evaluators are used for all participants.

We wish to consider items as raters and learners as subjects. The primitive categories obtained from the previous step are combined and transformed into six *mutually exclusive categories:* CP, CN, CZ, IP, IN and IZ, where,

CP: The answer to the item is correct with a positive effect,
CN: The answer to the item is correct with a negative effect,
CZ The answer to the item is correct with a null effect,
IP: The answer to the item is incorrect with a positive effect,
IN: The answer to the item is incorrect with a negative effect,
IZ: The answer to the item is incorrect with a null effect.

For each item and each category, we calculate the total number of learners. The computed values are defined in a contingency matrix as shown in Table 4 where,

– *i* is a category for one observer (from 1 to 6).
– *j* is a category for the other observer (from 1 to 6).

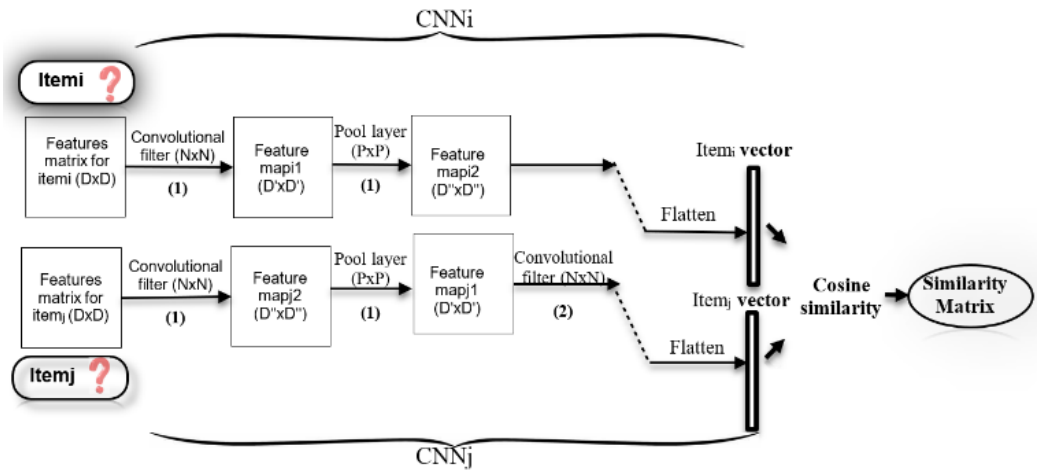


Figure 3. Synoptic of the Siamese Convolutional Neural Network model

- N is the total number of learners who answered all items.
- n_{ij} is the number of learners who answered both items $Item_i$ and $Item_j$ with the categories j and i respectively.
- n_{i+} is the margin per row as given by Eq. (10).

$$n_{i+} = \sum_{j=1}^{6} n_{ij} \quad (10)$$

- n_{+i} is the margin per column as given by Eq. (11)

$$n_{+i} = \sum_{j=1}^{6} n_{ji} \quad (11)$$

- P_{i+} is the marginal proportion per row as given by Eq. (12)

$$P_{i+} = \frac{n_{i+}}{N} = \frac{\sum_{j=1}^{6} n_{ij}}{N} \quad (12)$$

- P_{+i} is the marginal proportion per column calculated using Eq. (13)

$$P_{+i} = \frac{n_{+i}}{N} = \frac{\sum_{j=1}^{6} n_{ji}}{N} \quad (13)$$

Table 4. Contingency matrix

items / items	CP	CN	CZ	IP	IN	IZ	Total
CP	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	n_{16}	$\mathbf{n_{1+}} = \sum \mathbf{n_{1j}}$
CN	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}	n_{26}	$\mathbf{n_{2+}} = \sum \mathbf{n_{2j}}$
CZ	n_{31}	n_{32}	n_{33}	n_{34}	n_{35}	n_{36}	$\mathbf{n_{3+}} = \sum \mathbf{n_{3j}}$
IP	n_{41}	n_{42}	n_{43}	n_{44}	n_{45}	n_{46}	$\mathbf{n_{4+}} = \sum \mathbf{n_{4j}}$
IN	n_{51}	n_{52}	n_{53}	n_{54}	n_{55}	n_{56}	$\mathbf{n_{5+}} = \sum \mathbf{n_{5j}}$
IZ	n_{61}	n_{62}	n_{63}	n_{64}	n_{65}	n_{66}	$\mathbf{n_{6+}} = \sum \mathbf{n_{6j}}$
Total	$\mathbf{n_{+1}} = \sum \mathbf{n_{j1}}$	$\mathbf{n_{+2}} = \sum \mathbf{n_{j2}}$	$\mathbf{n_{+3}} = \sum \mathbf{n_{j3}}$	$\mathbf{n_{+4}} = \sum \mathbf{n_{j4}}$	$\mathbf{n_{+5}} = \sum \mathbf{n_{j5}}$	$\mathbf{n_{+6}} = \sum \mathbf{n_{j6}}$	***Total learners*** = N

Cohen's kappa (abbreviated as Kappa and denoted K) is a measure of inter-rater agreement [23] calculated with Eq. (14):

$$K = \frac{Po - Pe}{1 - Pe} \quad (14)$$

where,

- P_o is the overall accuracy of the model or the concordance proportion (observed percentage) calculated using Eq. (15)

$$P_o = \frac{\sum_{i=1}^{6} n_{ii}}{N} = \sum_{i=1}^{6} P_{ii} \quad (15)$$

- P_e is the measure of the agreement between the predictions model and the actual class values calculated using Eq. (16).

$$P_e = \sum_{i=1}^{6} P_{i+} * P_{+i} \quad (16)$$

It should be noted that if the Kappa is negative, then we speak in random agreement, which has the same quality as a null agreement.

3.3 The Siamese Neural Network (SNN) in conjunction with cosine measure

3.3.1 The Siamese Neural Network (SNN)

Siamese Neural Networks (SNNs) are an important model widely used in different fields, including face recognition, signature verification, prescription pill identification, etc. Siamese neural network consists of two identical subnetworks (sister networks). The outputs of the two subnetworks are combined, and then the final output similarity score is returned [24, 25]. We propose a standard Siamese Convolutional Neural Network (SCNNs) model to learn the similarity between two items, as shown in Figure 3.

Preparing and performing processing tasks on features matrix is important before presenting them to the SNN. The processing phase is described as follows:

1. Keep only learners who answered both items I and J.
2. Transform the two-feature matrix I and J into $D \times D$ square matrix by duplicating the columns and duplicating or reducing the rows (reshaping features matrix). Each row represents a learner and each column represents the item response features as shown in Table 5.

Table 5. The square feature matrix of i[th] item

	C	I	H	A	T	C	…	T
L1	C1i	I1i	H1i	A1i	T1i	C1i	…	T1i
L2	C2i	I2i	H2i	A2i	T2i	C2i	…	T2i
…	…	…	…	…	…	…	…	…
LD	CDi	IDi	HDi	ADi	TDi	CDi	…	TDi

The model consists of a sequence of three convolutional layers. Each layer uses a single channel with filters of a fixed size $N \times N$. The number of convolutional filters is fixed to optimize performance. Output feature maps are followed by a pooling layer of size $P \times P$.

1. Input two item features matrix (i, j) of dimension $D \times D$ to the Siamese model.
2. Create two sub-networks with the same architecture and same parameters. These two sub-networks mirror each other, which means that if the weights in one sub-network are updated, the weights in the other sub-network will also be updated. A Convolutional Neural Network (CNN) architecture [26] is applied for each sub-network. This architecture consists of a convolution layer, a dimensionality reduction layer or a pooling layer and the flatten. These layers are applied as follows:
   - First, a convolutional filter of size $N \times N$ is applied on the item matrix to extract the features.
   - Then, an average pooling of size $P \times P$ is applied to progressively reduce the spatial size of the feature map which decreases the number of parameters and computational time. The pooled feature map gets a small size.
   - Finally, the pooled feature map matrix is transformed into a one-dimensional vector which will be used with the cosine measure

3.3.2 The cosine similarity measure

After flattening the last features maps of $Item_i$ and $Item_j$, we obtain two vectors, I_i and I_j where $I_i = (I_{i1}, I_{i2}, I_{i3}, I_{i4})$ and $I_j = (I_{j1}, I_{j2}, I_{j3}, I_{j4})$. The cosine similarity measure between these outputs is computed using Eq. (2), which measures how similar two input items are. The obtained-result is stored in a similarity matrix as shown in Table 2.

3.4 Comparison of different models

Table 6 shows formal and implicit comparisons of three used models in terms of data inputs/outputs, type of data processing and data size. Table 6 also presents some weaknesses and strengths of each model. As shown in Table 6 we can observe that the SCNN model method is better than other existing models (Cosine similarity and Kappa) in terms of information loss and similarity degree. Note that SCNN model requires large data set with low information loss unlike Kappa. However, the SCNN model is good when extracting deep features but is heavy in terms of processing and consumes more data compared to cosine similarity.

Table 6. Comparison of different models

Criteria \ Model	Cosine similarity	Fuzzy-Kappa	SCNN-Cosine
Input data	Features matrices of the items		
Data processing	Transforming item feature matrix into a vector	Measure the distortion effect + transforming data into categories mutually exclusive	Transforming item feature matrix into a square one
Output data	Item-to-Item similarity matrix		
Data size	Large	Less	Very large
Similarity	The angle between two vectors	The agreement between two items	The angle between two flattened feature maps
Advantages	Easy to implement	Measure the degree of the effect	Overcomes information loss
Disadvantages	Less accurate	Very complicated	Complicated

4. ITEMS CLUSTERING

Item-to-skill mapping is an important phase in the educational field, where a system plays a pivotal role in assigning appropriate items to clusters. It is measured through the Item-to-item similarity measure across pairs of items. A higher similarity value means high similarity between two items, and a lower similarity value means the absolute difference between two items. To this end, a similarity matrix between pairs of items is computed which could be clustered using unsupervised techniques. Furthermore, clustering algorithms can be achieved in several and various unsupervised methods such as centroid-based clustering (K-means), density-based clustering, distribution-based clustering and hierarchical clustering [27].

K-means is the most effective technique since it only computes centroid-based clustering to place individuals in the population 'closest' to them, unlike supervised techniques where the number of partitions is required to ensure individuals (i.e., items) clustering [28].

5. EXPERIMENTS AND RESULTS

5.1 Experimental data

For this work, we used data sets from real educational systems information about the realistic performance of techniques, but the evaluation would be complicated by the fact that we do not know the "ground truth" (the correct similarity or clusters of items).

All tests are applied on two standard datasets: Algebra [29], which is a log of students' step-by-step performance (correctness and timing) during problem solving, and French Language [30], which corresponds to French course e-learning.

After selecting features and eliminating irrelevant information and irrelevant items, we obtain subsets of data on which we will test the proposed models. Table 7 illustrates a panoramic view of the characteristics of both datasets using some statistical data, before and after data preprocessing.

Table 7. Data statistics

Characteristics \ Data		Algebra	French-Language
School year		2005-2006	2007-2010
Number of skills		515	7
Before data pre_ processing	Data shape Rows×Columns	809694 × 18	359804 × 45
	Number of learners	574	2064
	Number of problems	1084	1880 (steps)
After data pre_ processing	Data shape Rows Columns	52316 × 7	2132 × 7
	Number of learners	550	353
	Number of problems	1037	407

After data cleaning, we built the different similarity matrices using the three proposed models. Each one of them is handled differently. In the SCNN model, we kept only the learners who answered both items proposed to be compared in a similarity matrix. Then, we performed a reshape on these into shape (28, 28) and applied a 3×3×64 convolutional filter and a 2×2 average pooling on them.

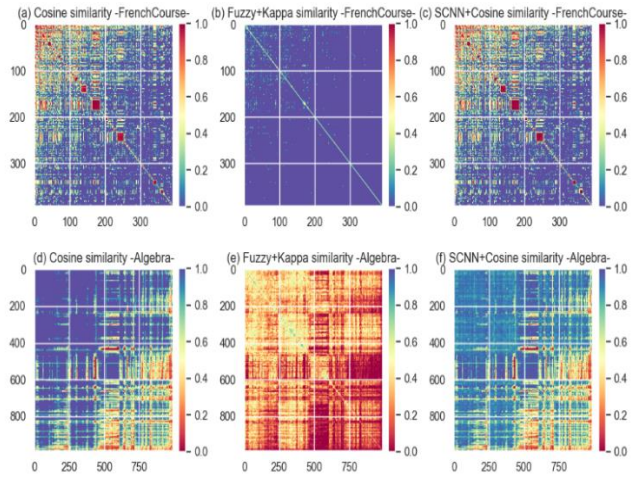


Figure 4. Visualization of the similarity matrices as an image

5.2 Results and discussions

5.2.1 Evaluating similarity matrix

Figure 4 shows similarity matrixes produced by each category of similarity measure. These can be visually inspected to gain a clear idea of how each matrix captures the components' structure. A ‘good’ matrix will have a strong ‘block’ structure, with each component clearly distinguishable, particularly on the diagonal.

5.2.2 Evaluating clustering quality

The items clustering using the k-means model is the last step of our approach, which should ensure clustering quality. The clustering quality is an important factor in data mining, where the behavioral parameters play a pivotal role in determining the correct number of clusters. The clustering quality could be measured by many metrics, in the experimentation we use the Elbow method and Silhouette measure. The Elbow method is a decision rule that helps to find optimal number of clusters. The Silhouette is an efficient metric used for validation while clustering.

The elbow method is a graphical representation that optimizes the number of clusters *K* in a K-means clustering method. It works by finding WCSS (Within-Cluster Sum of Square). In the proposed approach, the elbow method is mainly employed to validate the optimal number of clusters. Figure 5 illustrated the elbow graph of the WCSS algorithm. It is noticed that among 7 clusters obtained using the proposed similarity methods and applied on two datasets: French Language and Algebra, the edge falls between three and five clusters. Therefore, the optimal number of clusters is four.

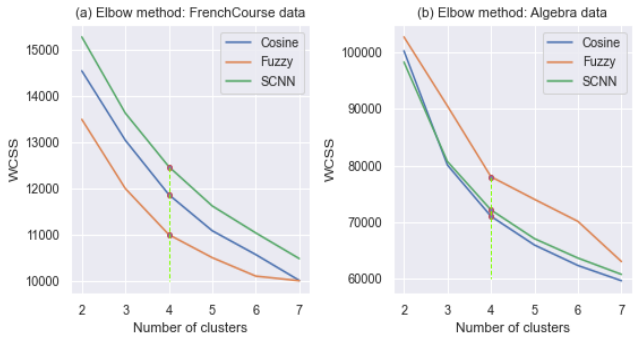


Figure 5. WCSS Elbow graph

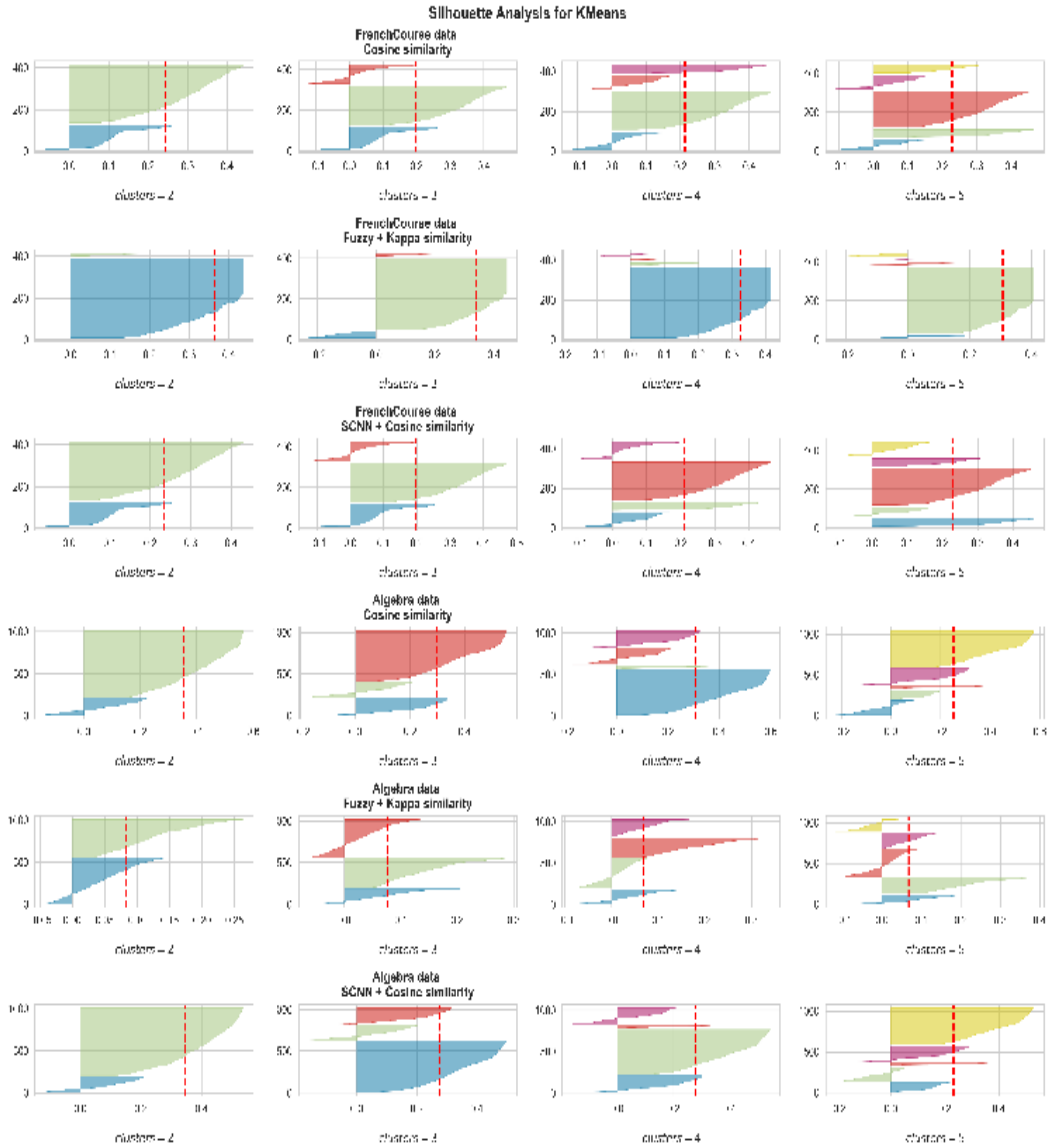
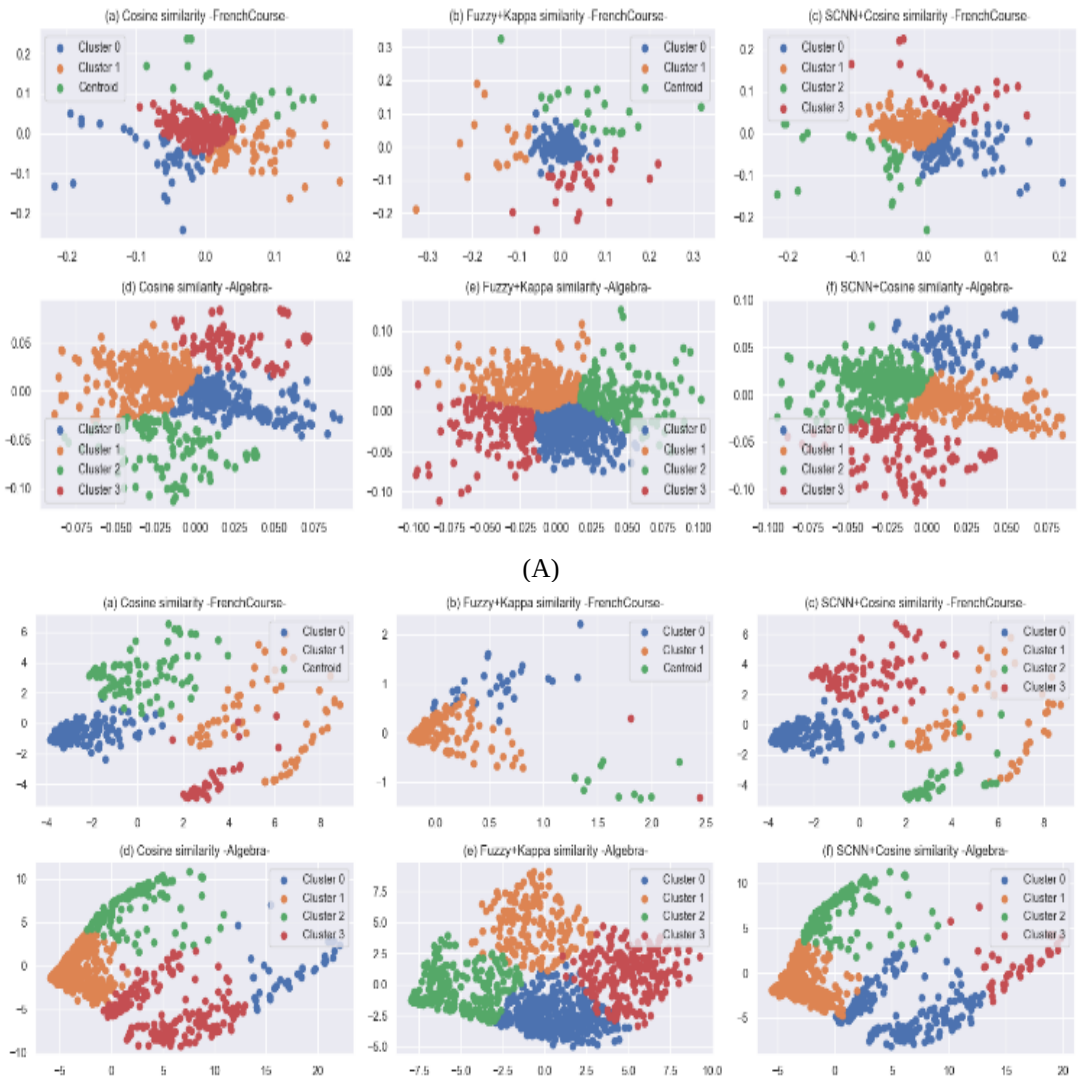


Figure 6. Silhouette analysis for K-means clustering



(A)

(B)

Figure 7. Visualization of the K-means clustering

Table 8. Silhouette scores

Similarity / Clusters	French Language data			Algebra data		
	Cosine	Fuzzy-Kappa	SCNN- Cosine	Cosine	Fuzzy-Kappa	SCNN-Cosine
2	**0.244**	**0.364**	**0.235**	**0.354**	**0.084**	**0.346**
3	0.197	0.338	0.196	0.298	0.076	0.275
4	**0.213**	0.325	**0.211**	**0.305**	0.069	**0.280**

The Silhouette score is a very useful method to find the number of optimal clusters K and help to validate and consolidate the results obtained by the elbow method [31]. Table 8 shows the Silhouette scores. It is noticed that in almost cases, the maximum value for K is 2 regardless of the method applied or the data used. However, that is not sufficient to select the optimal value of K. There should not be wide fluctuations in the size of the clusters as can be seen in Figure 6. Therefore, the Silhouette plot approach gives us $K = 4$ as the optimal value for different methods.

Figure 7 shows two clustering results: (A) is obtained when using eigenvectors and (B) is obtained after applying PCA on data to a lower dimensional subspace. Both PCA and eigenvectors are obviously applied on the different similarity matrices and then K-means is applied in the subspace. The clustering quality to a certain degree subjective and difficult to quantify as shown in Figure 7. However, clustering quality can be measured and used to compare similarity methods.

5.2.3 Results analysis

We consider here several types of analysis.

1. Clustering quality: We compare the quality of clustering by evaluating the metric Adjusted Rand Index (ARI) [32]. The ARI metric evaluates the agreement of two clustering with a chance-corrected agreement. Because we do not know the ground truth behind the data, we compared the similarity measures to each other. Table 9 shows a comparison between three proposed techniques in terms of ARI on two data sets: French Language and Algebra. As shown in Table 9 we can observe that Cosine and SCNN-Cosine measures are highly correlated (> 0.97) across both datasets and have almost the same ARI values. Larger differences (but only up to 0.1) can be observed between Cosine and Fuzzy-Kappa or Fuzzy-Kappa and SCNN-Cosine.

Table 9. Data statistics

Measures \ Data	French Language	Algebra I
Cosine vs. Fuzzy-Kappa	0.118	0.15
Cosine vs. SCNN-Cosine	0.99	0.97
Fuzzy-Kappa vs. SCNN-Cosine	0.116	0.14

2. Correlation measures [33]: To evaluate how similar two measures are, we take all similarity values for all pairs of items and compute the correlation coefficient in terms of three methods: Pearson, Kendall rank and Spearman. In the agreement matrix given by Table 10, the correlation values between Fuzzy-Kappa method and the other methods are small (around 0.18 or goes towards 0) which means there is a weak correlation between Fuzzy-Kappa and the other methods. On the other side, we notice a strong relationship between cosine and SCNN-cosine with values going toward 1.

Table 10. Agreement matrix

Correlation/ Data / Measures	French Course data			Algebra data		
	P	**K**	**S**	**P**	**K**	**S**
Cosine vs. Fuzzy-Kappa	0.18	0.18	0.18	0.15	0.14	0.18
Cosine vs. SCNN-Cosine	**0.996**	**0.95**	**0.92**	**0.95**	**0.83**	**0.91**
Fuzzy-Kappa vs. SCNN-Cosine	0.17	0.13	0.18	0.13	0.013	0.01
P = Pearson	K = Kendall			S = Spearman		

3. Clusters validity indices: To evaluate the cluster validity based on the average between and within clusters, we evaluate two indices: the Calinski-Harabasz Index (CHI) [34] and the Davies-Bouldin Index (DBI) [35]. CHI is a performance based on MSE (Mean Squared Error) that measures the average intra- and inter-cluster. DBI handles each cluster individually and seeks to measure its similarity to the closest cluster. As we can see in Table 11, the cosine similarity and the SCNN-cosine method give a high CH. This means these two similarities yield a better clustering since observations in each cluster are closer together (denser), while clusters themselves are further away (well separated).

Table 11. Clusters validity indices CHI and DBI

Correlation/ Indices / Measures	French Language data		Algebra data	
	CHI	**DBI**	**CHI**	**DBI**
Cosine	74.82	1.94	291.86	1.45
Fuzzy-Kappa	18.90	2.56	76.30	3.1
SCNN-Cosine	72.62	1.97	263.11	1.51

The Davies-Bouldin index shows that the better the clusters are separated, the better the clustering performance. Therefore, an optimal partition is one that minimizes the similarity between the clusters. As shown in Table 11, Cosine and SCNN-Cosine give better results than Fuzzy-Kappa.

Finally, the addition of the response time criterion to the similarity measure can change the meaning of similarity. However, the data size can increase the response times as well as computational complexity. Fuzzy-Kappa requires much time to calculate the contingency matrix and the distortion effect compared to SCNN-Cosine. In addition, it is more complex than SCNN-Cosine while Cosine method has the lowest response time and complexity.

The main drawback of the Fuzzy-Kappa method is its complexity. To overcome this disadvantage or shortcoming, we can reduce the membership functions of the inputs as well as those of the output.

For more precise results, we can assign weights to the different intrinsic characteristics related to the item (response time, number of hints and attempts as well as the value of the response).

6. CONCLUSION

This paper has studied an important issue which consists in automating computation of items similarities based on learners' performance data. These similarities are then used in further analysis of item-to-skills mapping. The item similarity approach is fairly straightforward, easy to realize, and it can be easily combined with other methods and applied on any sources of information. For these reasons, two similarity approaches (Cosine and Kappa agreement) as a baseline in proposals for more complex methods (SCNN and fuzzy analysis) have been combined. The proposed approach is based on several intrinsic characteristics related to the item, such as the response value, the latent trait, the response time and the complexity expressed in terms of hints and attempts and then selecting appropriate similarity methods and evaluating them on well-known educational datasets. The choice of a similarity measure is the most difficult step that should be guided by good clustering quality, correlation, and computational complexity that may be included significant behavioral information of learners. The SCNN-Cosine approach led to a good clustering quality, and computational complexity and times response, compared to Fuzzy-Kappa. This is demonstrated through experimental results regarding some metrics such as ARI, CHI and DBI to measure the quality of clustering. However, the approach suffers from a low data quality and data availability and should be focused on nearest works. In future research work, we will integrate other similarity measures such as-Pearson, Yule, Ochiai, Sokal, and Jaccard; as a basis for more complex methods. We can also use the IRT (Item Response Theory) to determine the ground of truth.

REFERENCES

[1] Chen, Y., González-Brenes, J.P., Tian, J. (2016). Joint Discovery of Skill Prerequisite Graphs and Student Models. International Educational Data Mining Society.

[2] Cen, H., Koedinger, K., Junker, B. (2006). Learning factors analysis–a general method for cognitive model evaluation and improvement. In: Ikeda, M., Ashley, K.D., Chan, TW. (eds). Intelligent Tutoring Systems. ITS 2006. Lecture Notes in Computer Science, vol. 4053. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11774303_17

[3] Desmarais, M. C., Meshkinfam, P., Gagnon, M. (2006). Learned student models with item to item knowledge structures. User Modeling and User-Adapted Interaction, 16: 403-434. https://doi.org/10.1007/s11257-006-9016-3

[4] Vuong, A., Nixon, T., Towle, B. (2011). A method for finding prerequisites within a curriculum. In EDM, pp. 211-216.

[5] Balint, T.A., Teodorescu, R.E., Colvin, K., Choi, Y.J., Pritchard, D.E. (2015). Identifying characteristics of pairs of questions that students answer similarly. Physics Education Research Conference, 2015: 55-58

[6] Dharaneeshwaran, N., Nithya, S., Srinivasan, A., Senthilkumar, M. (2017). Calculating the user-item similarity using Pearson's and cosine correlation. In 2017 International conference on trends in electronics and informatics (ICEI), pp. 1000-1004. https://doi.org/10.1109/ICOEI.2017.8300858

[7] Nazaretsky, T., Hershkovitz, S., Alexandron, G. (2018). Kappa learning: A new method for measuring similarity between educational items using performance data. arXiv preprint arXiv:1812.08390. https://doi.org/10.48550/arXiv.1812.08390

[8] Bichi, A.A., Talib, R. (2018). Item response theory: An introduction to latent trait models to test and item development. International Journal of Evaluation and Research in Education, 7(2): 142-151.

[9] Ghali, R. (2010). Impact des émotions sur les performances. Thèses et mémoires électroniques de l'Université de Montréal.

[10] Wu, M., Davis, R.L., Domingue, B.W., Piech, C., Goodman, N. (2020). Variational item response theory: Fast, accurate, and expressive. arXiv preprint arXiv:2002.00276. https://doi.org/10.48550/arXiv.2002.00276

[11] Rihák, J., Pelánek, R. (2017). Measuring Similarity of Educational Items Using Data on Learners' Performance. International Educational Data Mining Society.

[12] Pelánek, R., Effenberger, T., Vaněk, M., Sassmann, V., Gmiterko, D. (2018). Measuring item similarity in introductory programming. In Proceedings of the Fifth Annual ACM Conference on Learning at Scale, pp. 1-4. https://doi.org/10.1145/3231644.3231676

[13] Nazaretsky, T., Hershkovitz, S., Alexandron, G. (2019). Kappa Learning: A New Item-Similarity Method for Clustering Educational Items from Response Data. International Educational Data Mining Society.

[14] Mathisen, B.M., Aamodt, A., Bach, K., Langseth, H. (2020). Learning similarity measures from data. Progress in Artificial Intelligence, 9(2): 129-143. https://doi.org/10.1007/s13748-019-00201-2

[15] Han, J., Kamber, M., Pei, J. (2012). Data Mining: Concepts and Techniques. Morgan Kaufmann Third edition. https://doi.org/10.1016/C2009-0-61819-5

[16] Andrea, G. (2020). Cosine Similarity Matrix using broadcasting in Python. Published in Towards Data Science.

[17] Godjevac, J. (1999). Idées nettes sur la logique floue. Presses Polytechniques et Universitaires Romandes. Lausanne, Switzerland.

[18] Gacôgne, L. (2001). Logique floue et applications. In: Institut d'Informatique d'Entreprise. Evry, France.

[19] Nodelman, U., Allen, C., Perry, J. (2008). Fuzzy Logic. Stanford Encyclopedia of Philosophy. Bryant University.

[20] Khentout, C., Harbouche, K., Djoudi, M. (2021). Learner to learner fuzzy profiles similarity using a hybrid interaction analysis grid. Ingénierie des Systèmes d'Information, 26(4): 375-386. https://doi.org/10.18280/isi.260405

[21] Landis, J.R., Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1): 159-174. https://doi.org/10.2307/2529310

[22] Hallgren, K.A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. The Quantitative Methods for Psychology, 8(1): 23-34. https://doi.org/10.20982:tqmp.08.1.p023

[23] Shweta, R., Bajpai, C., Chaturvedi, H.K. (2015). Evaluation of inter-rater agreement and inter-rater reliability for observation data: An overview of concepts and methods. Journal of the Indian Academy of Applied Psychology, 41(3): 20-27.

[24] Koch, G., Zernel, R., Salakhutdinov, R. (2015). Siamese Neural Networks for one-shot image recognition. ICML Deep Learning Workshop, 2(1).

[25] Gresse, A., Dufour, R., Labatut, V., Rouvier, M., Bonastre, J.F. (2018). Mesure de similarité fondée sur des réseaux de neurones siamois pour le doublage de voix. In XXXIIèmes Journées d'Études sur la Parole (JEP). https://doi.org/10.21437/jep.2018-2

[26] O'Shea, K., Nash, R. (2015). An introduction to Convolutional Neural Networks. arXiv:1511.08458.cs Cornell University, https://doi.org/1048550/arxiv.1511.08458

[27] Fahim, A.M., Salem, A.M., Torkey, F.A., Ramadan, M. (2006). An efficient enhanced K-means clustering algorithm. Journal of Zhejiang University Science A, 7(10): 1626-1633. https://doi.org/10.163/jzus.2006.A1626

[28] Laxmi, L., Govindasuvamy, P., Lakshmanaprabu, S.K., Ranya, D. (2018). Document clustering based on text mining K-means algorithm using Euclidean distance similarity. Journal of Advanced Research in Dynamical and Control Systems, 10(2).

[29] Algebra, I. (2010). KDD Cup 2010 Educational Datamining Challenge. Hosted by PSLC DataShop. https://pslcdatashop.web.cmu.edu/KDDCup/.

[30] Jones, C. (2007). French Language. OLI Elementary French. https://pslcdatashop.web.cmu.edu/.

[31] Wang, F., Franco-Penya, H.H., Kelleher, J., Pugh, J., Ross, R. (2017). An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity. 13th International Conference on Machine Learning and Data Mining MLDM 2017, New York, USA, pp. 291-305. https://doi.org/10.1007/978-3-319-62416-7_21

[32] Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, 66(336): 846-850.

[33] Achtert, E., Böhm, C., David, J., Kröger, P., Zimek, A. (2008). Global correlation clustering based on the hough transform. Statistical Analysis and Data Mining: The ASA Data Science Journal, 1(3): 111-127. https://doi.org/10.1002/sam.10012

[34] Kozak, M. (2012). A dendrite method for cluster analysis by Caliński and Harabasz: A classical work that is far too often incorrectly cited. Communications in Statistics-Theory and Methods, 41(12): 2279-2280. https://doi.org/10.1080/03610926.2011.560741

[35] Davies, D.L., Bouldin, D.W. (1979). A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, (2): 224-227. https://doi.org/10.1109/TPAMI.1979.476