# IPB Prototype - Technical Report

Ana Isabel Pereira, M. Fátima Pacheco, Florbela Fernandes, Beatriz Flamia Azevedo, Felipe Gimenez, Gabriel Leite, Daniela Figueiredo, Sofia Romanenko, and Tânia Silva

**Abstract**

This report presents the OptLearn algorithm proposed by the Polytechnic Institute of Bragança for intelligent and customized question selection on the MathE platform. According to this proposal, the selection must be according to each user's needs. The developed proposal is based on two main methodologies: clustering and graph theory. The results are promising, as they were generally superior to those obtained with the random selection version.

## 1 Introduction

The development of learning capabilities and the learning process are strongly dependent on the active involvement of students in their education [1]. Active learning is a student-centered pedagogical technique that promotes undergraduate student learning more effectively than traditional or professor-centered approaches [9, 13]. The fundamental role of active learning is to encourage students to participate in their education by analyzing, debating, researching, and producing, either in groups or alone. In this way, the student is no longer just a listener and becomes an active participant in their learning ecosystem [8]. Based on this, university lecturers have been encouraged to adopt innovative methodologies and teaching tools to implement an interactive and appealing educational environment.

The MathE platform (mathe.pixel-online.org) is an online educational system that aims to help students who struggle to learn college mathematics as well as students who want to deepen their knowledge of a multitude of mathematical topics, at their own pace. On the platform, students worldwide have free access to resources such as videos, exercises, training tests, and pedagogical materials covering several areas of mathematics taught in higher education courses. Furthermore, MathE also has a Youtube channel and Social media (Facebook and Instagram). The MathE platform was created to provide students and teachers with a new perspective on mathematical teaching and learning dynamically and appealingly, relying on digital interactive technologies that enable customized study. The MathE platform has been online since 2019, and has been used by many students and professors worldwide. However, the necessity for some improvements on the platform has been identified by previous studies. In [5] the difficulties and potentialities of the platform were investigated, as well as the profile, preferences, opinions, and suggestions of the MathE users. In its turn, [3] had investigated the profiles of different groups of students exclusively in the Linear Algebra topic, which is the most used topic of the platform, [4]. The study suggested some clustering techniques to reorganize the resources available on the platform into different difficulty levels. Finally, [6] analyzed topics available on the platform that need to be restructured regarding question level and also analyzed the students' performance according to the countries they belong to. The information acquired from previous research helped the platform developers trace the path to provide intelligence for the MathE platform, based on optimization algorithms and machine learning, to make autonomous decisions tailored according to the needs of each user.

This technical report describes the proposal for the OptLearn algorithm customization developed by the Polytechnic Institute of Bragança (IPB - Portugal) researchers. The algorithm proposed aims to make the learning experience adaptive and personalized, considering the individual characteristics and needs of the student and their trajectory on the MathE platform. Thence, the methodology developed encompasses clustering methods to categorize the questions by difficulty levels and graph theory methods to customize the path according to the student's needs and abilities.

# 2 Methodology

The necessity to replace the conventional teaching approach has arisen with the advent of the digital revolution and the growing dependence on technology [1]. Gardner [11] affirms that individuals differ in their profiles of strengths and weaknesses across this intelligence, and educational and professional success depends on the ability to leverage one's strengths and compensate for weaknesses. Considering this, teaching effectively to every student's unique learning style is impossible. The methods applied to the algorithm development are described.

## 2.1 Clustering Techniques

In this work, the $k$-means clustering algorithm was used to classify the questions according to their difficulty level. Clustering is an unsupervised data partitioning method, being its main purpose to divide the elements of a dataset into groups (clusters) based on the similarities and dissimilarities of the elements, focusing on discovering underlying patterns in an unsupervised manner [17].

The k-means partitioning clustering algorithm is one of the most well-known clustering algorithms. It consists of trying to separate samples into groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares (WSS) [2]. As $k$-means is not an automatic clustering algorithm, it requires the definition of the initial parameter $k$, which represents the number of clusters division. Once this value is established, the $k$-means algorithm divides a set of $X$ samples $X_1, X_2, ..., X_m$ into $k$ disjoint clusters $C_k$, each described by the mean of the samples in the cluster, $\mu_i$, also denoted as cluster "centroids". In this way, the $k$-means algorithm aims to choose centroids that minimize the inertia, or within-cluster sum-of-squares criterion, presented in Equation 1 [2].

$$WSS = \min \sum_{i=1}^{k} \sum_{x \in C_i} ||X - \mu_i||^2, \tag{1}$$

From these centers, clustering is defined, grouping data points according to the center to which each point is assigned.

## 2.2 Graph Theory Approach

Graphs are mathematical structures that are often used to model problems involving relations between different objects. In its simplest form, a graph consists of a set of nodes (or vertices), which represent the objects of interest, and edges, that denote the relations between these objects.

Despite their apparent simplicity, graphs provide revealing visual representations of complex systems and enable us to analyze and understand underlying patterns and connections. For example, in physics, graphs can depict particle interactions and networks, providing a visual representation of fundamental forces and their effects; graph models can also provide insight into understanding the intricate connections in ecological food webs, analyzing brain networks in neuroscience, predicting the behavior of financial markets, and in many other types of problems across all areas of science [7, 10, 12, 14, 16, 18].

In a graph, the degree of a vertex represents the number of edges that are directly connected to it.

In order to improve the strategy for the progression of the students within the platform, an approach that considers the set of questions and keywords of a given subtopic as vertices and edges of a graph and takes into account the degree of each vertex as the indicator (among others) for the choice the following question is currently being conceived.

The idea behind this approach is that the degree of a vertex can indicate its centrality or importance within the network. Vertices with higher degrees often serve as crucial connectors, linking multiple parts of the graph. Although there is a strong belief that this approach holds promise, it has not been implemented yet.

# 3 Input and Output Data

The algorithm developed in this work aims to make the learning experience adaptive and personalized, taking into account the individual characteristics and needs of the student and their journey on the platform.

To develop the proposed prototype and personalize the student's experience as a user of the MathE platform, it is necessary first to define the difficulty level of the questions. This categorization considers both the student's opinion and the professor's opinion who develops the question since, in many cases, their opinions may differ regarding the question's difficulty level.

Thus, to evaluate the student's opinions, historical data from the MathE platform collected between 2019 and 2022 were used. Based on this data, it is possible to calculate the number of attempts and the number of correct answers for each question. This information is used to calculate the student's opinion denoted by $Stdscore_q$, for each question $q$.

On the other hand, the professor's opinion is obtained through the score assigned by the professor when adding a question to the platform. In this case, the scoring follows the Likert scale [15], divided into 5 levels. The questions with the lowest difficulty level are assigned a score of 1, while those with higher difficulty levels receive a score of 5. The remaining questions are distributed among the intermediate levels according to the difficulty level assigned by the professor, denoted by $Pscore_q$.

In this way, a score is calculated for each question by weighing the opinions of both students and the professor regarding the question, as presented in Equation (2).

$$score_q = \frac{\alpha \times Pscore_q + \beta \times Stdscore_q}{\alpha + \beta} \tag{2}$$

Note that $\alpha$ and $\beta$ represent weights for each of the scores, aiming to, when necessary, weigh the opinions of students and professors for each of the questions.

In cases where a question has had few responses, the professor's score prevails until a minimum number $\gamma$ of responses is reached to define the student's score, $Stdscore_q$. The calculated score, obtained by weighing the opinions of both students and the professor, is used for the division of clusters. Based on this concept, the questions were grouped into clusters representing different levels of difficulty. Five clusters were proposed to develop the prototype, where questions with lower difficulty are allocated in cluster 1, while the most challenging questions are in cluster 5.

Additionally, a set of keywords associated with each question was also used to establish the relationship between level-question and keywords.

As an example, consider Table 1 where a set of 40 questions have their corresponding keywords.

Table 1: Relation between question and its keywords

| Question ID | Keywords ID |
|---|---|
| 0 | 114, 115, 116, |
| 1 | 116, 121, 122 |
| 4 | 118, 119, 120 |
| 6 | 116, 122, 123 |
| 11 | 114, 116, 121, 122, 123 |
| 12 | 116, 121, 122, 123 |
| 14 | 115, 116, 117, 122, 123 |
| 15 | 119, 122, 123 |
| 2 | 114, 115, 116, 121 |
| 3 | 114, 115, 116, 121, 122 |
| 9 | 114, 115, 116, 117, 121 |
| 10 | 114, 115, 116 |
| 13 | 114, 115, 116, 123 |
| 24 | 116, 118, 119, 120, 121 |
| 28 | 114, 115, 116, 117, |
| 37 | 115, 116, 121, 122, 123 |
| 5 | 114, 115, 116 |
| 7 | 114, 116, 122, 123 |
| 8 | 116, 122, 123 |
| 16 | 114, 115, 116 |

| Question ID | Keywords ID |
|---|---|
| 23 | 121, 122, 123 |
| 25 | 114, 115, 116, 117 |
| 27 | 114, 115, 116, 117 |
| 29 | 117, 118, 119 |
| 17 | 114, 116, 122, 123 |
| 18 | 116, 117, 121, 122, 123 |
| 19 | 118, 119, 120 |
| 20 | 114, 115, 116, |
| 21 | 114, 115, 116, 124 |
| 22 | 114, 115, 116, |
| 26 | 115, 119, 122, 123, 125 |
| 30 | 116, 117, 118, 119, 120 |
| 31 | 114, 115, 116, 117 |
| 32 | 121, 122, 123, 124, 125 |
| 33 | 121, 122, 123 |
| 34 | 119, 122, 123 |
| 35 | 116, 121, 122, 123, 124 |
| 36 | 116, 121, 122, 123 |
| 38 | 115, 120, 121, 122 |
| 39 | 116, 118, 120, 121, 123 |

# 4 Proposed Algorithm

This section presents the proposed Algorithm developed by the IPB team, named OptLearn Prototype. This algorithm is divided into four phases. In Phase $I$ the score of the question is defined based on the professor's and students' opinions. After that, in phase $II$, the question's difficulty level is defined by a clustering algorithm, through the analysis of the score previously defined. In Phase $III$, the graphs theory approach establishes the relationship between questions and keywords. In turn, in phase $IV$, the questions that will be provided for the students are finally selected.

To better explain the algorithm concepts, let's consider a set of questions $Q = \{1, ..., nq\}$ and a set of keywords $KW \in \{kw_1, kw_2, ..., kw\}$ in which each question $q$ has the following parameters:

- number of output at MathE platform $Nout_q$,
- number of incorrect answers $Nerror_q$,
- professor evaluation parameter in terms of difficult level $Pscore_q$.

The following algorithm should be applied to each topic or subtopic.

**Algorithm 1** IPB Optlearn Algorithm

**Phase I: Score calculation**
Consider the input parameters: $Q = \{1, ..., nq\}$, $Pscore_q \in \{1, ..., SMax\}$.
$Nout_q \longleftarrow$ Number of question's output, therefore $Nout = \{Nout_{q_1}, ..., Nout_{nq}\}$;
$Nerror_q \longleftarrow$ Number of incorrect answers of the question, therefore $Nerror = \{Nerror_{q_1}, ..., Nerror_{nq}\}$;
$\alpha \longleftarrow$ weight of student opinion;
$\beta \longleftarrow$ weight of professors opinion;
$\gamma \longleftarrow$ minimal number of answers to be considered the student's opinion.
Evaluate $NPscore_q = \frac{Pscore_q - 1}{SMax - 1}$
Evaluate $Stdscore_q = \frac{Nerror_q}{Nout_q}$
**if** $Nout_q > \gamma$ **then**
$score_q = \frac{\alpha \times NPscore_q + \beta \times Stdscore_q}{\alpha + \beta}$
**else** $score_q$=$NPscore_q$
**end if**

**Phase II: Difficulty level definition by cluster**
Consider $score = \{score_q, ..., score_{nq}\}$ as input parameters;
$k \longleftarrow$ number of cluster partitioning;
Obtain the clustering $k - means(score, k)$ associating each question to a given cluster

**Phase III: Questions-keywords Graphs**
Consider $Q \in \{1, ..., nq\}$, $KW \in \{kw_1, kw_2, ..., kw_n\}$
Identify the list of relations between questions $q \in Q$ and keywords $kw \subset KW$.
Create a graph considering questions and keywords as nodes of the graph, connected through edges representing the connection. Considering $G = [G_{qw}]$ the adjacency matrix of the graph, $G_{qw} = 1$ if the question $q$ is related to keyword $w$ and $G_{qw} = 1$ otherwise.

**Phase IV: Question selection**
For a given student, do:
**while** Number of questions selected for the current test is not reach **do**
**if** the student has a cluster association, $c$ **then**
Identify the questions correctly answered in the cluster, defined as $QC$
Identify the keywords associated with the questions $QC$
Calculate the graph edge's weight of that current cluster
Select the edge with the maximum value
**else**
Select randomly a question belonging to cluster 1
**end if**
**if** $q$ has a corrected answered **then**
$c = \min(c + 1, SMax)$
**else**
$c = c - 1$
**end if**
**end while**

Some extra situations needs to be satisfied in Algorithm 1, named:

- If in Phase IV the set $QC$ is empty the question is randomly selected in the cluster $c$;
- In a given test, the questions can not be repeated;

If in Phase IV the set $QC$ is empty or the student.

# 5 Results and Discussion

A set of 40 questions related to the subtopic Matrices and Determinants was used to test the proposed algorithm. These questions were distributed into 5 clusters, $k = 5$. To ensure the algorithm's functionality, some adjustments were made in the distribution of questions among the clusters so that each cluster contains 8 questions. It is important to highlight that these adjustments were only necessary due to the small number of available questions to test the approach.

## 5.1 Questions Categorization Results

Table 1 provides an organized representation of the questions and their respective keywords at different difficulty levels. In this way, the column *Cluster* presents the level assigned for each question, which is also the cluster where the question is located. The column *Questions Assigned* identifies the questions assigned for each cluster.

Table 2: Question Categorization by level (clusters)

| Cluster | Questions Assigned |
|---|---|
| Cluster 1 | ID: 0, 1, 4, 6, 11, 12, 14, and 15 |
| Cluster 2 | ID: 2, 3, 9, 10, 13, 24, 28, and 37 |
| Cluster 3 | ID: 5, 7, 8, 16, 23, 25, 27, and 29 |
| Cluster 4 | ID: 17, 18, 19, 20, 21, 22, 26, and 30 |
| Cluster 5 | ID: 31, 32, 33, 34, 35, 36, 38, and 39 |

## 5.2 Graphs Approach Results

Within each cluster, the questions are interconnected with their respective keywords through a graph structure, where questions and keywords are represented as vertices, and the edges represent the possible paths that each student can follow. In addition, for each edge, a weight is assigned to indicate how suitable each question is for the student. Furthermore, the student's progression between clusters depends on the type of responses provided. The student advances to a higher-level cluster for each correct answer, while for each incorrect answer, the student returns to a lower-level cluster. Therefore, the student is considered successful if all questions in the last cluster are answered correctly.

To better describe the results, let's consider cluster 1 (level 1) as an example. The questions belonging to this cluster are interconnected through keywords. These connections form a graph in which both the questions and the keywords are represented as vertices and are connected by edges. Figure 1 presents the graph for this level:

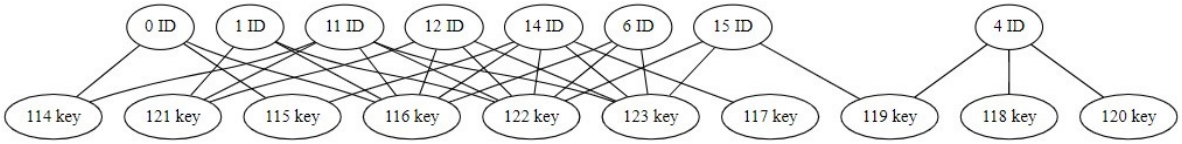


Figure 1: First level graph

In this context, when a student accesses the platform for the first time, the prototype randomly selects a question for the student. This question has connections with other questions through the keywords associated with it. So, let's suppose that the selected question is question ID 15, and this question is related to keywords $KW_{119}$, $KW_{122}$, and $KW_{123}$, establishing connections with other questions beyond the initially chosen one.

If the student answers the selected question correctly, they will advance to the next level of difficulty. At this point, the selection of the next question will no longer be random but based on the keywords with which the student has already interacted and succeeded. The system will utilize the established

connections between the questions and keywords to identify the relevant questions for the student's level of knowledge based on their previous correct answers.

The prototype adapts the selection of questions based on the student's performance and the relationships between the questions and keywords, providing a more personalized and targeted learning experience.

On the other hand, consider that the student does not succeed in answering the question while at difficulty level 2, even though they have previously succeeded in questions to reach that level. In this case, the system will redirect them to the lower level. The selection of the next question will be based on the keywords in which the student has previously succeeded, ensuring that there is no repetition of questions, but choosing questions related to the keywords in which they have already demonstrated competence. Figure 2 shows the graph of level 2.

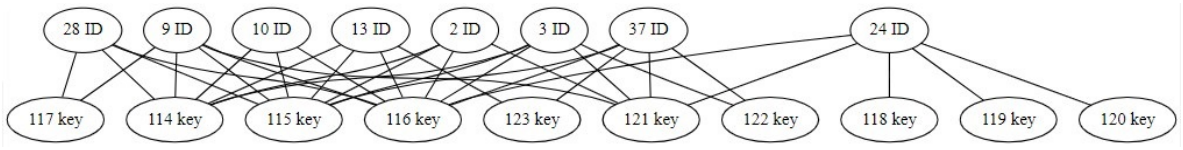


Figure 2: Second level graph

The same logic is applied to all levels, except for specific details described in Algorithm 1. Thus, Figures 3, 4, and 5 represent the graphs of the third, fourth, and fifth levels of difficulty, respectively.

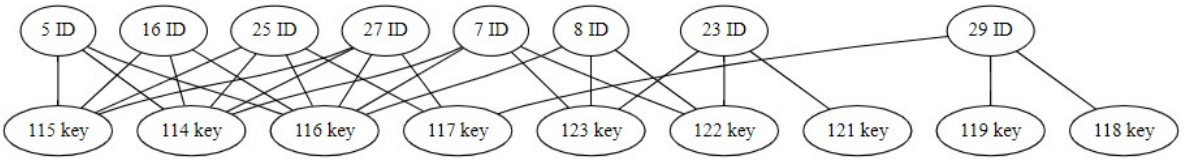


Figure 3: Third level graph

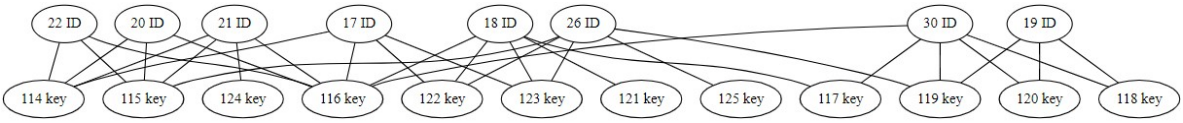


Figure 4: Fourth level graph

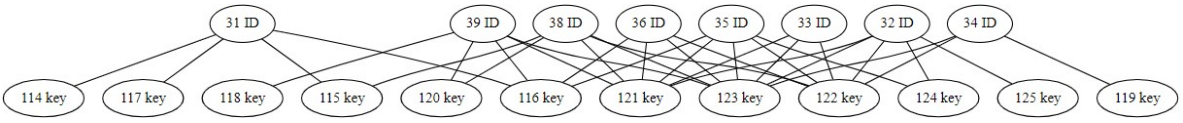


Figure 5: Fifth level graph

It is important to emphasize that the graph structure allows for a comprehensive analysis of the connections between the questions and keywords at each level of difficulty. This enables the identification of patterns and more complex relationships, contributing to the accurate selection of the next questions based on the student's performance and demonstrated skills.

Finally, the use of graphs as a representation of the connections between questions and keywords in the adaptive learning system provides a more efficient and personalized approach. This graphical representation allows for clear visualization of the relationships between system elements, facilitating the selection of the next questions based on the student's performance history. With this approach, it is possible to create a continuous flow of learning, directed and tailored to the individual needs of each student, maximizing the efficiency and effectiveness of learning.

More information about utilizing or reproducing the code can be found at the READ ME documentation.

# 6 Conclusion

This report presented the IPB approach for the OptLearn algorithm. Basically, the approach developed combines clustering and graph techniques to personalize the student learning path. In this way, it ensures the student progress occurs progressively considering a diversity of contents and mainly the system personalization according to the user's needs.

# References


[1] Alhawiti, N.M.: The influence of active learning on the development of learner capabilities in the college of applied medical sciences: Mixed-methods study. Advances in medical education and practice **14**, 87–99 (2023). https://doi.org/doi.org/10.2147/AMEP.S392875

[2] Arthur, D., Vassilvitskii, S.: K-means++: The advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. p. 1027–1035. SODA '07, Society for Industrial and Applied Mathematics, USA (2007)

[3] Azevedo, B.F., Rocha, A.M.A.C., Fernandes, F.P., Pacheco, M.F., Pereira, A.I.: Evaluating student behaviour on the mathe platform - clustering algorithms approaches. In: (In press) Book of 16th Learning and Intelligent Optimization Conference - LION 2022. pp. 319–333. Milos - Greece (2022)

[4] Azevedo, B.F., Amoura, Y., Rocha, A.M.A.C., Fernandes, F.P., Pacheco, M.F., Pereira, A.I.: Analyzing the mathe platform through clustering algorithms. In: Gervasi, O., Murgante, B., Misra, S., Rocha, A.M.A.C., Garau, C. (eds.) Computational Science and Its Applications – ICCSA 2022 Workshops. pp. 201–218. Springer International Publishing, Cham (2022)

[5] Azevedo, B.F., Pereira, A.I., Fernandes, F.P., Pacheco, M.F.: Mathematics learning and assessment using mathe platform: A case study. Education and Information Technologies (2021). https://doi.org/10.1007/s10639-021-10669-y

[6] Azevedo, B.F., Romanenko, S.F., de Fatima Pacheco, M., Fernandes, F.P., Pereira, A.I.: Data analysis techniques applied to the mathe database. In: Pereira, A.I., Košir, A., Fernandes, F.P., Pacheco, M.F., Teixeira, J.P., Lopes, R.P. (eds.) Optimization, Learning Algorithms and Applications. pp. 623–639. Springer International Publishing, Cham (2022)

[7] Bollobás, B.: Modern Graph Theory. Graduate Texts in Mathematics 184, Springer-Verlag New York (1998)

[8] Deslauriers, L., McCarty, L.S., Miller, K., Callaghan, K., Kestin, G.: Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. Proceedings of the National Academy of Sciences **116**(39), 19251–19257 (2019). https://doi.org/10.1073/pnas.1821936116

[9] Dunkle, K.M., Yantz, J.L.: Intentional design and implementation of a "flipped" upper division geology course: Improving student learning outcomes, persistence, and attitudes. Journal of Geoscience Education **69**(1), 55–70 (2021). https://doi.org/10.1080/10899995.2020.1787808

[10] Farahani, F.V., Karwowski, W., Lighthall, N.R.: Application of graph theory for identifying connectivity patterns in human brain networks: a systematic review. frontiers in Neuroscience **13**, 585 (2019)

[11] Gardner, H.: Frames of mind: The theory of multiple intelligences. No. 3, New York: Basic Books (2011)

[12] Harary, F.: Graph Theory. Addison-Wesley Publishing Company (1969)

[13] Indorf, J.L., Benabentos, R., Daubenmire, P., Murasko, D., Hazari, Z., Potvin, G., Kramer, L., Marsteller, P., Thompson, K.V., Cassone, V.M., Stanford, J.S.: Distinct factors predict use of active learning techniques by pre-tenure and tenured stem faculty. Journal of Geoscience Education **69**(4), 357–372 (2021). https://doi.org/10.1080/10899995.2021.1927461

[14] Liu, Y., Safavi, T., Dighe, A., Koutra, D.: Graph summarization methods and applications: A survey. ACM computing surveys (CSUR) **51**(3), 1–34 (2018)

[15] Norman, G.: Likert Scales, Levels of Measurement and the "Laws" of Statistics. Advances in Health Sciences Education **15**(5), 625–632 (2010). https://doi.org/10.1007/s10459-010-9222-y

[16] Priyadarsini, P.: A survey on some applications of graph theory in cryptography. Journal of Discrete Mathematical Sciences and Cryptography **18**(3), 209–217 (2015)

[17] Rehman, A.U., Belhaouari, S.B.: Divide well to merge better: A novel clustering algorithm. Pattern Recognition **122**, 108305 (2022)

[18] Riaz, F., Ali, Khidir, M.: Applications of graph theory in computer science. In: 2011 Third International Conference on Computational Intelligence, Communication Systems and Networks. pp. 142–145 (2011). https://doi.org/10.1109/CICSyN.2011.40