I. J. Education and Management Engineering, **2024, 4, 15-22**
Published Online on August 8, 2024 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijeme.2024.04.02

Optimization of Curriculum Content Using Data Mining Methods

Firudin T. Aghayev
Institute of Information Technology, Baku, Azerbaijan
Email: agayevinfo@gmail.com
ORCID iD: https://orcid.org/0000-0002-7660-1406

Gulara A. Mammadova*
Institute of Information Technology, Baku, Azerbaijan
Email: gyula.ikt@gmail.com
ORCID iD: https://orcid.org/0000-0002-6332-7129
*Corresponding Author

Rena T. Malikova
Institute of Information Technology, Baku, Azerbaijan
Email: melikovarena22@gmail.com
ORCID iD: https://orcid.org/0000-0001-6589-6379

Lala A. Zeynalova
Institute of Information Technology, Baku, Azerbaijan
Email: lalamaillala@bk.ru
ORCID iD: https://orcid.org/0000-0003-0882-455X

Received: 07 December, 2023; Revised: 19 February, 2024; Accepted: 18 March, 2024; Published: 08 August, 2024

Abstract: The purpose of this article is to search and extract the necessary content, identifying curriculum topics. Classification and clustering of text documents are challenging artificial intelligence tasks. Therefore, an important objective of this study is to propose and implement a tool for analyzing textual information.

The study used Data Mining methods to analyze text data and generate educational content. The work used methods for classifying text information, namely, support vector machines (SVM), Naive Bayes classifier, decision tree, K-nearest neighbor (kNN) classifier.

These methods were used in developing the curriculum for the specialty "Cybersecurity" for the Faculty of Information and Telecommunication Technologies. About 48 curricula in this specialty were analyzed, topics and sections in disciplines were identified, and the content of the academic program was improved. It is expected that the results obtained can be used by specialists, managers and teachers to improve educational activities.

Index Terms: Curriculum content, Data Mining methods, Text Mining, semantic similarities, SVM, Naive Bayes classifier, decision tree, kNN classifier.

1. Introduction

Due to the rapid development of digital technologies throughout the world, significant changes are taking place in the education system. In the modern world, education is considered as an essential component of economic growth and development of the state. Its role for people and society is constantly increasing. As our world becomes increasingly digital, the education sector is increasingly filled with websites, apps, social media and learning environments. This leads to the accumulation of large amounts of information in education, the use of which becomes more and more difficult every year.

The increase in information flows and their versatility complicates decision-making in the field of education. The criterion for making an effective decision is the analysis of information received from participants in the educational process at its various stages.

In connection with the growing use of information technologies in education, interest has arisen in new methods and approaches, in the automated identification of new, sometimes hidden, relationships in data. Under these conditions to increase the efficiency of analyzing large volumes of information, there is a need for new non-traditional methods of information processing.

In this case, there is a need to create a repository of educational training programs and extract the necessary information from them. This is important for optimizing educational activities.

When creating a curriculum, teachers often use data from information resources on the Internet. Lecture notes and materials for practical training can be found in the public domain on the websites of educational institutions and in search engines. At the same time, content generation is carried out manually. The teacher spends a huge amount of time on this process. In this regard, the use of tools and methods for intellectual processing of text information becomes relevant. Data Mining methods are used to analyze text data and generate educational content [1]. Text mining is a branch of artificial intelligence, the goal of which is "the computer-aided discovery of new, previously unknown information by automatically extracting information from various written sources". Using these tools reduce the time and effort required to create a curriculum.

Putting learning content into practice will help educators figure out what needs to be done to make the learning experience more engaging and effective. This will also help students see their mistakes, on the basis of which recommendations will be given for repeating the necessary material to eliminate deficiencies.

At the stage of developing a training program, it is necessary to create or select learning resources that ensure the achievement of planned learning outcomes, develop all elements of the course, and fill the course with both content and organizational components. Since a large number of digital educational resources have already been accumulated and their repositories have been organized, there are good conditions for creating recommender systems in this area. Difficulties at this stage may include a shortage of content, its obsolescence, and the need to adapt it to different audiences.

Semantic comparison of different training programs can also provide more accurate results by creating a concrete model of computer terminology. Existing systems do not yet have the ability to intellectualize the learning process, flexibly adapt educational content to the individual needs of students, and do not contain convenient intelligent tools for automating routine operations for structuring educational content. To solve this problem, it is necessary to select the necessary educational material from a large collection of educational materials and determine ways to structure it according to the relevant subject and topic (Fig. 1).

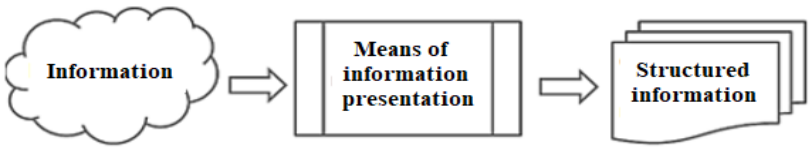

Fig. 1. Structuring information by means of its presentation

In the learning process, text is one of the main ways of transmitting and interacting with educational information. Thus, learning materials mining is essential to utilize this rich source of information.

This process includes searching for the necessary information, automatically constructing a repository of educational materials, annotations, frequency distribution of key terms, pattern recognition tasks, including analysis of connections and associations, visualization and predictive analytics. In electronic education, Text Mining methods are applied to written resources: electronic textbooks, lectures and other materials posted on the websites of educational institutions and on the Internet in general.

The purpose of this article is to search and extract the necessary content, identifying curriculum topics.

2. Related Works

The development of modern intelligent learning systems includes the following main stages [2]:

- determination of the purpose of the system and methods of use;
- determination of the structural components of educational materials;
- selection of educational material and its structuring.

To accomplish this task, it is necessary to determine ways to select educational material from the entire set and ways to structure it in a specific discipline and relevant topic.

The structuring of the content of educational material must meet certain requirements. Thus, work [3] presents the following principles for structuring content:

- modularity of content;
- integration (establishing a relationship between the structural components of educational content);
- taking into account the student's subjective characteristics and the dynamics of the learning process;

- informativeness of training (providing the necessary information for the student to develop the necessary professional competencies);
- content distribution (network placement of information based on specified requirements).

At the same time, the main requirement of the curriculum is a focus on developing professional competencies in the student.

Any curriculum consists of a basic (core) and an advanced (individualized) component. The basic component includes topics that are required to be studied. They form the student's basic skills in a given subject area. The individualized (extended) component includes topics that expand the student's understanding of a given topic and serve to develop the necessary competencies in his future professional activity.

In research [4], the authors proposed a method for selecting training modules to build the structure of an individualized curriculum depending on the student's input set of skills and abilities and the output (planned learning outcomes).

In works [5-7], to form an individual educational trajectory when teaching IT specialties, they used an intelligent model of the educational process in the form of a labeled Petri net.

The study [8] describes in detail the main methods of text mining and text sentiment analysis: summarization, information extraction, categorization, visualization, clustering, topic tracking, question answering. Methods of text mining, as well as its sentiment, are used in [9], a method for semantic comparison of the content of educational programs using text similarity methods is presented in publication [10]. Thus, an analysis of sources shows that modern learning management systems are aimed at supporting intelligent methods for developing and maintaining electronic training courses.

3. Methodology and Research Methods

There are quite a large number of varieties of the listed problems, as well as methods for solving them.

Unstructured text analysis techniques lie at the intersection of several fields: data mining, natural language processing, information retrieval, data mining, and knowledge management [11].

Currently, many applied problems are described in the literature, the solution of which is possible using the analysis of text documents. These include classic problems of data mining: classification, clustering and tasks typical only for text documents: creating automatic text annotations, extracting key concepts from text, etc. [12-14].

The purpose of classification in text data mining is to determine for each document one or more predefined categories to which it belongs, in other words, to discover the probability of each document belonging to a certain category. Formally, the task of classifying text documents is to assign them to a predetermined category or group of categories based on the content of the document [15].

We describe the set of educational materials obtained from various sources in the following form:

$$D = \{d_1, \dots, d_i, \dots, d_n\} \tag{1}$$

Let us denote the categories of documents:

$$C = \{c_r\} \text{ , where r = 1, …, m.} \tag{2}$$

Thus, for each category there should be many attributes:

$$F(C) = (c_r) \text{ , where } F(c_r) =< f_1, \dots, f_l, \dots, f_t > . \tag{3}$$

Such a set of features is often called a dictionary, because it consists of lexemes that include words and/or phrases that characterize the category.

Similarly, like categories, each document also has characteristics by which it can be assigned with some degree of probability to one or more categories:

$$F(d_i) =< f_1^i, \dots, f_j^i, \dots, f_k^i > \tag{4}$$

The set of characteristics of all documents must coincide with the set of characteristics of categories, i.e.:

$$F(C) = F(D) = F(d_i) \tag{5}$$

In the classification problem, it is necessary to construct a procedure based on these data, which consists of finding the most probable category from the set *C* for the document d_i under study.

Most text classification methods are one way or another based on the assumption that documents belonging to the same category contain the same features (words or phrases), and the presence or absence of such features in a document indicates its belonging or non-belonging to a particular topic.

The decision to classify document di as category c_r is made based on the intersection of $F(d_i)$ and $F(c_r)$.

The degree of information content of the *j*-th text material can be determined from the relationship:

$$S_j = \frac{n_j^k}{W_j}, \tag{6}$$

where n_j^k - number of keywords used in *j*-th text material, W_j - volume of text material in characters.

The resulting numerical value S_j determines the degree of concentration of the educational material with keywords. A large number of keywords in a small amount of information determines the degree of concentration of the material on the subject of the search.

Most algorithms for clustering text information require that the data be represented as a vector space model [16]. This is the most widely used model for information retrieval.

It is used to reflect semantic similarity as spatial proximity. In this model, each document is represented in a multidimensional space, in which each dimension corresponds to a word in a set of documents.

This model represents documents as a matrix of words and documents:

$$\mathrm{M} = |\mathrm{F}| \times |\mathrm{D}| \tag{7}$$

$$F = \{f_1, \ldots, f_k, \ldots, f_z\} \tag{8}$$

$$D = \{d_1, \ldots, d_i, \ldots, d_n\}, \tag{9}$$

where *F* - a set of features that is constructed by excluding rare words and words with high frequency (stop words), d_i — vector in *z*-dimensional space R_z.

Each feature f_k in document d_i is associated with its weight $\omega_{k,i}$, which indicates the importance of this feature for this document. To calculate the weight, different approaches can be used, for example, the TF-IDF (Term Frequency Inverse Document Frequency) algorithm [17]. The idea of this approach is to guarantee that the weight of a feature will be in the range from 0 to 1. Moreover, the more often a word appears in the text, the higher its weight, and vice versa: the lower the frequency, the lower the weight. The formula by which weight is calculated is as follows:

$$\omega_{k,i} = \frac{(1+\log(N_{i,k}))\log(D/N_k)}{\sqrt{\sum_{S \neq K}(\log(N_{i,s})+1)^2}} \tag{10}$$

Where $N_{i,k}$ – number of occurrences of a feature f_k in the document d_i;

N_k- number of occurrences of a feature f_k in all documents of the set *D*;

D – number of documents (the power of set D).

It should be noted that the denominator contains the amount for all documents except the one under consideration. This way, the weight of the feature is normalized across all documents. This model is often called the "bag-of-words model".

In addition to the TF-IDF method, the TLTF (Term Length Term Frequency) approach is often used for weighting terms. Clusters in this model are represented similarly to documents in the form of vectors:

$$C = \{c_1, \ldots, c_j, \ldots, c_m\}, \tag{11}$$

where c_j — vector in *z*-dimensional space R_z. The vector c_j is often the center of the cluster (centroid).

At the same time, the purpose of clustering is to group documents (represented by vectors) into clusters in accordance with their proximity to centers.

4. Research Results and Discussions

In our research, hundreds of web pages of higher educational institutions of both the Republic of Azerbaijan and leading universities in the world were analyzed. 48 training programs in the specialty "Cybersecurity" were identified and selected for analysis. The syllabus texts were retrieved from university websites and uploaded into the system.

Our task was to extract the names of topics and sections of curricula from a variety of curriculums and classify them according to their similarity (Fig. 2). To do this, we used intelligent methods for classifying text materials. Key words and terms were identified from each educational material, then the terms were normalized (unnecessary ones were excluded, all words were reduced to a single form). A dictionary of curriculum terms was compiled. Topics and sections similar in the dictionary fell into one group.

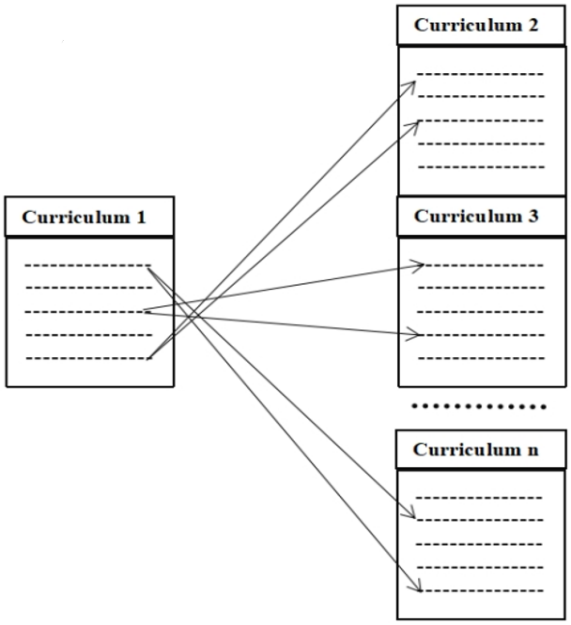

Fig.2. Comparison of the names of topics in disciplines of various educational curriculumss in the same specialty.

The steps the tool takes to find the most frequently occurring topics in different curricula are:

- From the curriculum repository, a specific curriculum is selected one by one and compared with other programs. The occurrence of each topic in other curricula is calculated.
- Topics from all curricula of the same subject are combined and duplicate topics are removed.
- Topics are sorted in descending order of their frequency. This allows you to determine how many educational programs included this topic for training.

We extracted the names of disciplines, topics and sections of the curriculum from web pages, then converted the information into text format. Many websites abroad presented their programs in English, and the curricula of Azerbaijani universities were presented in the national language. All information has been translated into English. Comparison, analysis and classification of key terms was carried out over words of the English language. . Below is a fragment of the curriculum in the "Cybersecurity" direction at Stanford University (Fig.3.).

155. Computer and Network Security
https://courseware.stanford.edu/pg/courses/lectures/349991
- Control hijacking attacks and defenses
- Tools for robust code
- Exploitation techniques and fuzzing
- Dealing with legacy code
- Operating system security
- Cryptography overview
- Basic web security model
- Web application security
- Session management and user authentication
- HTTPS: goals and pitfalls
- Network protocols and vulnerabilities
- Network defenses
- Denial of service attacks
- Malware
- Mobile platform security models
- Mobile threats and malware
- The Trusted Computing Architecture

255. Introduction to Cryptography:
http://crypto.stanford.edu/~dabo/cs255/
- History and overview of cryptography
- Identification protocols
- Basic symmetric-key encryption
- One time pad and stream ciphers
- Block ciphers
- Block cipher abstractions: PRPs and PRFs
- Attacks on block ciphers
- Message integrity: definition and applications
- Collision resistant hashing

Fig. 3. Fragment of the curriculum of Stanford University (USA) in the specialty "Cybersecurity"

When solving the problem of classifying the text of educational programs, the following steps were performed:

1. At the preliminary stage of information processing, the text is normalized, that is, the entire text of the curriculum is translated into lower case, each word in the text is presented in the nominative singular case, prohibited characters and stop words are removed;
2. Tokenization (conversion of text into a list of lexemes) and lemmatization (conversion to normal dictionary form) was carried out;
3. Stop words (words used with high frequency) have been removed from the list of tokens;
4. A dictionary of tokens has been compiled;
5. We divided the set of program texts into two subsets (training and testing): at the first stage, we optimized the classification parameters and made an initial comparison of various classification algorithms; at the second stage, we carried out a final check of the model's quality;
6. For each classification algorithm, the parameters were optimized to find the best loss function value. Based on the loss function, the results are compared with each other.

As part of the experimental study, various classification algorithms, as well as their ensembles, were compared. Stages 1-4 make it possible to reduce the dimension of the feature space of the input data, which reduces the complexity of the classification algorithm and is a means of combating overfitting.

It should be noted that not only the terms, but also the names of the topics may differ in the curriculum. For example, in one curriculum the title of the topic is indicated as "The Role of Information Security Standards", in another - "Comparative Analysis of Information Security Standards". Or in one - "Security classes of computer systems", in the other - "Classification of methods for ensuring the security of computer networks".

Based on the example above, there is a need to search for similar topics as many topics are presented differently or have different names for the same topic. Therefore, it is necessary to make a semantic comparison. At this stage, the topics of each curriculum were compared with other topics of other curriculums and the overall amount of semantic similarity for each topic was determined. The degree of semantic similarity of the names of topics and sections of the academic discipline was determined using formulas 7-10 given above.

Therefore, when comparing topics and sections of the academic discipline of different universities, we also determined the degree of semantic similarity. The semantic approach improves results and accuracy. Semantically similar terms should be considered the same terms. For example, the curriculum terms presented in Table 1 are semantically similar, so they should be considered the same terms.

Table 1. Semantically similar terms from the "Hardware and Software Security" curriculum

Term 1	**Term 2**
Security ID	Ensuring identical security
Traffic filtering	Firewall
Securing computer networks	Network and Web Security
Multi-level protection	Hierarchical level of protection
Software protection	Protection model
Trust Computing	Robust computing algorithms

When conducting experimental studies, 4 classification algorithms were used: support vector machine (SVM), naive Bayes classifier, k-nearest neighbors method, decision tree method [18-21].

75% of the analyzed information was taken as the training set, and the rest was used for testing (the data splitting factor was chosen based on an analysis of the performance of the various data pieces used).

Tables 2 and 3 show the accuracy and running time of various educational program classification algorithms.

Table 2. Accuracy of various educational program classification algorithms

Classification algorithms	
SVM	72.534
Na ïve Bayes	70.347
k-nearest neighbors	64.163
Decision tree	62.569

Table 3. The amount of time it takes a computer to train and test using various classification algorithms

Classification algorithms	**Training time (minutes)**	**Testing time (minutes)**
SVM	0.06137	0.00045
Na ïve Bayes	0.03175	0.00765
k-nearest neighbors	0.00361	4.09431
Decision tree	0.94129	0.02521

The tables show that when using different classification algorithms, different results were obtained. The highest accuracy result was obtained when using the SVM algorithm – 72.5. The time spent on the computing process is small (from 0.06 to several minutes).

All work on calculating accuracy indicators for various classification algorithms was performed using the Rapid Miner application.

In the future, we also intend to carry out work to optimize the curriculum, taking into account the time required to complete each topic and section of the academic discipline, as well as the entire curriculum as a whole.

5. Conclusion

In this study, we applied Text Mining methods to automate the construction of an educational program in the specialty "Cybersecurity". In the process of constructing the text of the program, the following algorithms were used: SVM, Naïve Bayes, k-nearest neighbors and Decision Tree. The best results were obtained using the support vector machine (SVM) algorithm, which showed an accuracy of more than 70 percent. Also, when comparing topics and sections of the academic discipline of different universities, the degree of semantic similarity was determined.

As a result of the work carried out, the structure of the academic discipline was improved, some new sections and topics were added to it. In the future, we intend to conduct research in other areas of knowledge and disciplines, as well as continue collecting information and apply other algorithms and software to optimize the educational process.

The results of the study can be used by university teaching staff to improve the process of developing and adjusting the educational program.

References

[1] Romero, C., Ventura, S. Data Mining in Education. WIREs Data Min. Know. Disc. 2013, 3, 12–27 https://doi.org/doi: 10.1002/widm.1075.

[2] Trembach V.M. The main stages of creating intelligent teaching systems // Software Products & Systems. № 3, 2012, pp. 147-151.

[3] Xramsova, E.O. Bochkarev P.V. Intelligent training systems // Theory. Practice. Innovation. 2017, №12, pp. 56–62.

[4] Marcos L., Pages C., Martinez J.J., Gutierrez J.A. Competency-based Learning Object Sequencing using Particle Swarms. 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007), IEEE, October 29-31, 2007. Patras, Greece, 2007, pp. 77–92.

[5] Shukhman, A.E. Work in progress: Approach to modeling and optimizing the content of IT education programs / A.E. Shukhman, I.D. Belonovskaya // Global Engineering Education Conference (EDUCON). 2015. pp. 865–867. DOI: 10.1109/ EDUCON.2015.70960741.

[6] Shukhman, A.E. Individual learning path modeling on the basis of generalized competencies system / A.E. Shukhman, M.V. Motyleva, I.D. Belonovskaya // Proceedings of the IEEE Global Engineering Education Conference (EDUCON), 13-15 March 2013. Berlin, 2013, pp. 1023–1026. DOI: 10.1109/EduCon.2013.6530233.

[7] Prilepina A.V. Methodology for developing educational programs for training specialists for the information technology industry / A.V. Prilepina, E.F. Morkovina, A.E. Shuxman // Vestnik of Orenburg State University. 2016, №1(189), pp. 41–46.

[8] Sathya R. A. Survey On: A Comparative Study of Techniques in Text Mining / R. Sathya // International Journal of Electrical Electronics & Computer Science Engineering Special Issue. 2018, pp. 45–48.

[9] Santos C.L., Rita P., Guerreiro J. Improving international attractiveness of higher education institutions based on text mining and sentiment analysis. International Journal of Educational Management, 2018, vol. 32, no. 3, pp. 431–447.

[10] A Text Mining Methodology to Discover Syllabi Similarities among Higher Education Institutions / G. Orellana et al.// 2018 International Conference on Information Systems and Computer Science (INCISCOS). – IEEE, 2018. Pp. 261–268.

[11] Ronen Feldman. The Text Mining Handbook. Cambridge University Press. 2006. 421 p.

[12] G. King, P. Lam, and M. Roberts, "Computer-assisted keyword and document set discovery from unstructured text," Copy at http://j. mp/1qdVqhx Download Citation BibTex Tagged XML Download Paper, vol. 456, 2014.

[13] M. Ergün. Using the techniques of data mining and text mining in educational research. Electronic journal of education sciences. 2017. Volume 6, pp. 180-189.

[14] Yogapreethi N., Maheswari S. A review on text mining in data mining. International Journal on Soft Computing (IJSC) Vol.7, No. 2/3, August 2016, pp. 1-8.

[15] Hartmann J., Huppertz J., Schamp C., Heitmann M. Comparing automated text classification methods. International Journal of Research in Marketing, Volume 36, Issue 1, 2019, pp. 20-38.

[16] James E. Dobson. Vector hermeneutics: On the interpretation of vector space models of text. Digital Scholarship in the Humanities, Vol. 37. No. 1, 2022, pp. 81-93.

[17] Automatic Keyword Extraction from Individual Documents [Electronic resource]. – Access mode: https://www.researchgate.net/publication/227988510 _Automatic_Keyword_Extraction_-d-from_Individual_Documents.

[18] Lipo Wang. Support Vector Machines:Theory and Applications. Springer, 2005

[19] Chatterjee, S., George Jose, P., & Datta, D. (2019). Text classification using SVM enhanced by multithreading and CUDA. International Journal of Modern Education & Computer Science, 11(1), 11–23. https://doi.org/10.5815/ijmecs.2019.01.02

[20] Liu, P., Zhao, H., Teng, J., Yang, Y., Liu, Y., & Zhu, Z. (2019).Parallel Naive Bayes algorithm for large-scale Chinese text classification based on spark. Journal of Central South University, 26, 1–12. https://doi.org/10.1007/s11771-019-3978-x

[21] Xu, B., Guo, X., Ye, Y., & Cheng, J. (2012). An improved random forest classifier for text categorization. Journal of Computing, 7(12), 2913–2920. https://doi.org/10.4304/jcp.7.12.2913-2920

Authors' Profiles

Firudin T. Aghayev graduated from "Automation and computing devices" faculty of Azerbaijan Polytechnic University. In 1997, he defended a thesis on specialty of "Distance aerospace researches" on subject of "Establishment of high speed radon-type changes in modeling process of pattern restoration methods in accordance with projections". In 2016, he was awarded with the title of associate professor by the Higher Attestation Commission under the President of the Republic of Azerbaijan. He is the author of 41 articles and 3 books. Currently he is the curator at Training Innovtaion Center of Institute of Information Technology. He is the head of department of Institute of Information Technology of ANAS.

Mammadova A. Gulara graduated from mechanical-mathematical faculty of Azerbaijan State University. She is involved in a pedagogical activity at the Training Innovation Center of the Institute of Information Technology. She is dealing with formation of information culture in the educational environment, development of technological bases of electronic teaching resources, increasing teachers' ICT training issues in the direction of problems of informatization of educational process. She is the author of 20 scientific works and one thesis. She works as senior research fellow at the Institute of Information Technology.

Rena T. Malikova graduated from the Faculty of Automatics and Computer Science, Azerbaijan Polytechnic Institute named after Ch.Ildirim. From 1986 until the appointment she worked for the ACS Department of Azerbaijan Academy of Sciences. She is engaged in training activities in the Training Innovation Center of the Institute. Research interests: data mining; social network analysis. She works as senior research fellow at the Institute.

Lala A. Zeynalova graduated from the Faculty of Automatics and Computer Science, Azerbaijan Polytechnic Institute named after Ch.Ildirim. From 1986 until the appointment she worked for the ACS Department of Azerbaijan Academy of Sciences. She is engaged in training activities in the Training Innovation Center of the Institute. Research interests: data mining; social network analysis. She works as senior research fellow at the Institute.

How to cite this paper: Firudin T. Aghayev, Gulara A. Mammadova, Rena T. Malikova, Lala A. Zeynalova, "Optimization of Curriculum Content Using Data Mining Methods", International Journal of Education and Management Engineering (IJEME), Vol.14, No.4, pp. 15-22, 2024. DOI:10.5815/ijeme.2024.04.02