Data Acquisition Model for an Intelligent System Integrating the Creation of Educational Programs and Professional Standards

DINARA KAIBASSOVA, ARAILYM ASHIMBEKOVA
Department of Computing Engineering,
Astana IT University,
Mangilik El avenue, 55/11, Astana,
KAZAKHSTAN

Abstract: - The Digital Kazakhstan program ensures that higher education meets the needs of modern industrialization by developing students' professional skills and competencies. These qualities guarantee that students will be able to quickly adapt to the professional environment and remain competitive in the national and global labor markets after graduation. The aim of this research is to improve the content of educational programs by developing models, methods, and algorithms for an intelligent system for designing educational programs that consider the interrelationships between the subjects studied and the competencies formed through natural language analysis and processing. To achieve the goal, several tasks were solved, including the creation of a data model of educational content (subject programs) and professional content (requirements of professional standards) to study their structural components. This paper presents a model of the professional competencies database and the concept of creating an educational program corresponding to the required skills. This model is based on the analysis of unstructured texts describing the content of educational courses. The work performed has led to two important results. First, the successful prototyping of an intelligent system that dynamically links course content to the required professional competencies. Second, the development of an algorithm that improves the accuracy of matching learning outcomes to industry standards. These achievements prove that educational programs can be designed to better meet the changing demands of the labor market, equipping graduates with the necessary skills to succeed.

Key-Words: - data analysis, natural language texts, frequency matrix, education content, education program, intelligent system, natural language, professional competencies**,** professional standards, matching learning.

Received: February 16, 2024. Revised: August 12, 2024. Accepted: September 7, 2024. Published: October 4, 2024.

1 Introduction

To achieve the required educational results, regulatory and advisory documents, such as competency and qualification frameworks, and educational and professional standards, are developed and implemented. In this regard, the task of creating modern methods of planning and organizing the educational process that considers the needs of employers and labor market trends become especially urgent, [1]. As employer requirements are constantly changing and evolving, the methods developed must support the development of professional skills through specialized training programs.

The 2020 rating of educational programs at higher education institutions in RK, conducted by the National Chamber of Entrepreneurs “Atameken”, demonstrated that there is a 48% demand for graduates and that employers in RK are satisfied with the quality and relevance of the educational programs. The criteria are united in three main blocks: career prospects of graduates, quality of educational programs, and achievements of students. "The rating methodology was adopted as a result of open discussion with the public of higher education institutions and experts. Accordingly, the methodology of the rating is constantly being improved. The Ministry and National Chamber of Entrepreneurs "Atameken" hold open discussions with the heads of universities and the rating is aimed at protecting the interests of students, who should and will understand what career prospects are open to them after graduation. The authors [2], [3] assert that the key stakeholders in developing an educational program aimed at developing professional competencies should be students, representatives of the labor market (employers), and teaching staff.

In the process of developing educational programs, analysis of the content of academic disciplines, their relationship with the requirements of educational and professional standards, as well as considering the demand of the labor market for specialties and areas (profiles) of education, becomes crucial. Consequently, curriculum design includes the process of creating a basis for professional competencies based on professional standards, which will be the basis for the process of creating and updating educational programs.

For studying this subject area, the main document is state standards in the field of information technology and professional standards [4] developed by the National Chamber of Entrepreneurs «Atameken» under the project "Development of Labor Skills and Job Incentives" within the framework of the Partnership Agreement between the Government of the Republic of Kazakhstan and the International Bank for Reconstruction and Development.

2 Subject Area

In the context of academic freedom in universities, the quality of preparing specialists directly depends on the relevance of educational programs. Moreover, the professional competencies of future specialists should be shaped in accordance with the labor market demands in specific regions of the country. When developing educational programs, one of the primary tasks for educational institutions is to identify a set of disciplines that align with the required competencies. However, there is a lack of software applications in the market that would allow for a comparative analysis of the content, objectives, and outcomes of different disciplines and courses, irrespective of their subject area, to update and develop educational programs in line with the latest requirements of educational and professional standards.

In this regard, the possibility of integrating components of data analysis into a wide range of applied software becomes relevant. Designing an educational program is a multi-step process, as it reflects the dynamics of student learning and should allow for timely adjustments in the teaching process, [5].

The research aims to enhance the quality of educational program content by developing models, methods, and algorithms for an intelligent system that shapes educational programs, considering the interrelation between studied disciplines and the formation of competencies, based on the analysis and processing of natural language texts.

To achieve the stated goal, several tasks need to be addressed, and one of them is as follows:

Develop data models for educational content (syllabus thematic plans) and professional content (requirements of professional standards) to study structural elements.

In essence, the research seeks to create an intelligent system capable of analyzing and processing textual information in natural language to optimize and improve the design of educational programs. By considering the interaction between different disciplines and aligning them with the requirements of professional standards, the system aims to ensure that the educational content remains relevant and meets the demands of the labor market.

The object of the research is professional standards and discipline curricula (syllabus). In other words, the main subjects of the study are normative documents defining the requirements for professional training of specialists, as well as educational programs that include information about the content and organization of education for individual disciplines.

The subject of the research is natural language text models designed for text data search, classification, and clustering. The research aims to develop and apply methods of natural language text analysis and processing to structure and systematize the information contained in professional standards and discipline curricula.

The practical significance of this research lies in the fact that the obtained results and software can be used to optimize and shape educational programs. By utilizing the developed methods and algorithms for text analysis, researchers and educational institutions can better understand the interrelations between disciplines and competencies, adapt the content of educational programs to changing labor market demands, and improve the quality of specialist training. For the successful completion of the task of researching a method for selecting textual documents necessary for organizing education based on the identified competencies of the educational program, complex intelligent algorithms need to be developed. The first algorithm aims to extract terms characterizing the given domain of scientific knowledge, while the second algorithm aims to determine the similarity of these terms with the competency base of the subject area developed in this work, [6].

Natural language texts have an informal structure, consisting of sentences based on the rules of natural language and using the full diversity of its vocabulary, [7]. Unlike fixed constructions in programming languages, the absence of such fixed structures in natural language makes data extraction highly relevant and often serves as the first step in text processing.

Data extraction is the process of finding, collecting, and storing information in various formats. Specific characteristics of this process include:

- processing individual texts or collections of texts;
- extracting data relevant to specific problems, questions, or topics.

To address this task effectively, the use of sophisticated intelligent algorithms is essential in order to extract relevant data and information for organizing education and shaping educational programs based on the identified competencies.

3 Materials and Methods

To address the research task of selecting textual documents necessary for organizing learning based on the identified competencies of the educational program, sophisticated intelligent algorithms need to be developed. These algorithms will facilitate efficient information extraction and processing from the text documents.

Let's describe the model for creating an educational program that meets the specified competencies. The problem statement can be formulated as follows:

Let's assume we have a collection of syllabuses $D=(d_1,d_2,...,d_n)$ and a dictionary of terms from the competency database $S=(s_1,s_2,...,s_n)$, which interacts with the user's query. Consider a document from the collection of syllabuses $d_i \in D$ and represent it as a vector in the space. Then this vector will have the form as in (1):

$$d_i=(tf_{i1},tf_{i2},...,tf_{in}) \quad (1)$$

where tf_{ij}- is the number of occurrences of the word from the query s_j in the document d_i. In the vector model, the document is considered as an unordered set of terms.

The goal is to obtain a matrix:
$X=\{x_{ij}\}$,

where $x_{ij}=tf(s_j,d_i)$ - represents the frequency of terms occurring in the document.

Let's consider the sequence of actions for organizing the search s_j in the document d_i:

- select a text document from the document collection;
- remove stop words from the text;
- considering the word morphology, calculate the frequency of occurrence of each term.

Rank the terms in descending order of their occurrence frequency.

Form the term-document matrix.

Thus, we obtain the term-document matrix, where the rows correspond to documents from the collection, and the columns correspond to terms, [8]. As a result, there might be zero rows corresponding to syllabi that do not match any competency, as they are not considered for further analysis, and such rows are removed, [4], [7]. There might also be zero columns, representing uncovered competencies, in which case it is necessary to add syllabi covering those competencies.

For vectorizing the corpus, various methods are used, such as frequency-based methods, direct encoding methods, term frequency-inverse document frequency methods, and distributed representation methods. These methods allow transforming documents from the transform corpus into the vector space model to make predictions later.

The frequency vector is the simplest method of vectorization, which involves filling the vector with frequencies of word occurrences in the document [9]. The encoding of each document is represented as a set of its constituent lexemes, and the value for each word position in the vector is the counter of the corresponding word. This method is suitable for models based on the Bayesian approach, [10]. The drawback of this method is that as the result of vectorization, lexemes that occur very frequently may appear to be more "significant" than those occurring much less frequently, [11].

In turn, professional competencies describe a set of essential typical characteristics of a particular specialty, defining the specific direction (profile) of an educational program. They manifest in a specialist's ability to address a set of professional tasks within their chosen field of activity based on specific knowledge, skills, and abilities. The list of professional competencies is structured in accordance with the main types of professional activities that a graduate should be prepared for, such as scientific research, project management, production and technological skills, and organizational and managerial competencies.

The main elements of a professional standard are as follows:

- Direction of Professional Activity: this describes the field or area of work where the standard applies.

- Job Title: The official designation of the profession for which the standard is defined.
- Qualification Level: This specifies the requirements regarding education and work experience within the National Qualifications Framework.
- Labor Function: A set of interconnected work actions that contribute to achieving a specific task or objective.
- Requirements for Skills and Knowledge: The necessary skills and knowledge needed to perform a particular labor function effectively.

In this context, the present work focuses on the development of a database of professional competencies, which will be described in more detail below.

a. Professional Competencies Database Model

The problem statement is to extract data from the database to construct a query, which consists of the following procedures:

Processing the query, including:

- selecting from directions;
- selecting professions;
- selecting functional tasks;
- selecting competencies.

Viewing, creating, moving, editing, and adding competencies, functional tasks, professions, and directions.

The database structure consists of 5 main tables (Figure 1), interconnected by parent-child relationships.

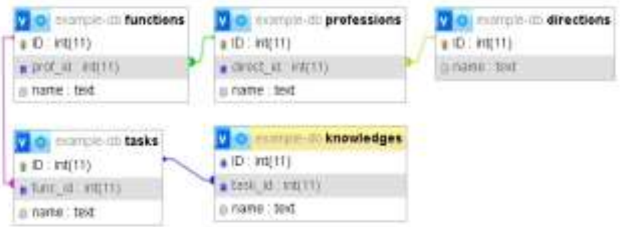

Fig. 1: Database model «Professional competencies»

The «Directions» table is intended for storing information about educational program directions: direction identifier (id) and direction name (name).

The «Professions» table contains information about professions related to the corresponding directions: profession identifier (id), a field (direct_id) serving as a foreign key, and the name of the profession (name). It's important to note that a foreign key is a column or a combination of columns whose values correspond to the primary key in another table. In this case, the values of the «direct_id» column (foreign key) correspond to the values of the «id» column (primary key) from the "Directions" table.

The other tables «Functions», «Tasks» and «Knowledges» are created using the same scheme. Building the database based on a relational DBMS allows for utilizing built-in full-text search, regulating relationships between competencies, and implementing certain linguistic algorithms using stored procedures and functions.

The interaction between the database and the visualization module is organized using a retrieval command, which, in turn, performs the function of transforming the input text into a matrix, where the values represent the occurrences of a given key (word) in the text.

b. The structural elements of an educational program

According to the Working Educational Program (Syllabus), include the following components:

- Information about Instructors: Details about the teaching staff involved in delivering the course.
- Description of the Studied Discipline: An overview of the subject or course being taught.
- Objectives and Goals of the Discipline: Clearly stated aims and objectives of the course.
- Brief Content, Topics, and Duration of Study: An outline of the course content, the topics covered, and the duration of each segment.
- Assignments and Requirements: Guidelines for self-study assignments and the expectations set by the instructor.
- Evaluation Criteria and Schedule: The criteria for assessment, timelines for task completion, and submission deadlines for assignments related to the discipline.
- Student Assessment Criteria: The criteria used for evaluating students' knowledge and performance.
- List of Main and Additional Literature and Information Resources: A compilation of primary and supplementary reading materials and other resources relevant to the course.

Additionally, the working educational program for a discipline may include variations of examinations, course projects, and assignments tailored for students in distance learning programs.

From the perspective of educational program development, the following crucial concepts can be highlighted:

1. Educational Goals and Objectives of the Course: The competencies that the educational program aims to develop and the expected learning outcomes for the students after completing the

discipline. The course's place within the educational program is determined, specifying the entry requirements (prerequisites) for students and the subsequent preparation (post requisites) for the educational program.

2. Course Structure: This includes a breakdown of the sections, key topics, objectives, and content of lectures and practical sessions, as well as assignments for independent student work.

Much of the information in educational programs and discipline syllabuses is presented in large blocks of loosely structured text. Some sections are formatted as tables (e.g., workload of the course, thematic plan of the discipline, list of practical and laboratory work) or lists (e.g., objectives and goals of mastering the discipline, etc.).

This shortcoming can be elucidated in greater detail by examining several pivotal factors.

1. The prevalence of frequently occurring words. Words that occur frequently, such as prepositions, conjunctions, and general terms, are assigned high values in the frequency vector. This results in an overestimation of their significance. Consequently, words that are significant but infrequent may be eclipsed.

2. The inability to distinguish between synonyms and polysemantic words. The Frequency Vector method is unable to distinguish between synonyms and polysemantic words. In the event that a word possesses a similar denotation yet disparate forms, the same word will be represented by distinct vector elements. Furthermore, the same word in disparate contexts will have an identical representation.

4 Results and Discussion

The primary findings of the content analysis of educational programs (subject programs) are presented in the form of frequency distributions of keywords and phrases, as well as identified connections between disciplines and competencies.

1. Analysis Methodology: The analysis of curriculum content was conducted in several stages, including:

- the texts comprising the curriculum were then collected and prepared for analysis;
- the texts were then parsed, and keywords and phrases were highlighted;
- the frequency distribution of keywords and phrases was determined;
- the identification of connections between disciplines and competencies.

2. Analysis Results: The results of the analysis are presented in the form of frequency distributions of keywords and phrases, as well as identified connections between disciplines and competencies.

The frequency distribution of keywords and phrases demonstrates which terms are most frequently encountered in curriculum texts. This enables the identification of the principal topics and areas of focus within each discipline that are of greatest importance and relevance.

Connections Between Disciplines and Competencies: the identification of connections between disciplines and competencies enables the determination of which disciplines contribute the most to the formation of specific competencies. This facilitates the optimization of educational program structures and ensures their alignment with professional standards.

5 Conclusions

In summary, the work involved analyzing the development of educational programs in accordance with current educational and professional standards. It also addressed the challenges of aligning the professional domain with the educational program. Additionally, the study explored the structural elements of educational programs and described models for organizing and representing knowledge in intelligent systems, specifically for educational and professional content. Finally, a formal and structured model was developed to represent the elements of educational and professional content and their relationships.

This study makes a significant contribution to the existing research in this field.

1. An innovative solution for designing educational programs has been developed. This solution comprises the creation of an intelligent system that dynamically links course content with professional competencies. This allows educational programs to be more accurately adapted to the requirements of the labor market.

2. The algorithm developed for this project increases the accuracy of educational programs' alignment with professional standards, thereby enhancing the quality of student training and their competitiveness in the labor market. Furthermore, the introduction of natural language text analysis algorithms allows educational institutions to more effectively update and adapt their programs to meet changing requirements.

3. The practical significance and potential applications of this research are twofold. Firstly, the

results obtained and the developed software can be used to optimize educational programs at universities and other educational institutions. Secondly, the study demonstrates how modern methods of data analysis and natural language processing can be integrated into educational processes to improve their quality and meet the requirements of the labor market.

By implementing this method and algorithm, it becomes possible to efficiently extract relevant data from large volumes of loosely structured texts and create valuable resources in the form of a competency database and a corpus of documents enriched with competencies. These resources can significantly aid in curriculum development, competency-based education, and research in the field of education and professional development.

Declaration of Generative AI and AI-assisted technologies in the writing process.

During the preparation of this work the authors used ChatGPT (Generative Pre-trained Transformer) in order to improve the readability and language of their manuscript. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References:

[1] V. Yavorskiy, D. Kaibassova, Y. Klyuyeva (2022, May 9-12) «Issues of developing measures to analyze storage medium for educational achievements of students», *in Proc. ENERGYCON 2022 - 2022 IEEE 7th International Energy Conference,* Riga, Latvia.

[2] B. Dauletbakov, Zh. Ivanova. Modernization of educational programs in the Republic of Kazakhstan // Scientific and methodological electronic journal «Concept». – 2017,T. 9, pp. 57–61, [Online]. http://e-koncept.ru/2017/870010.htm (Accessed Date: July 8, 2024).

[3] R. R. Khairutdinov, R. S. Safin, E. A. Korchagin, F. G. Mukhametzyanova, A. V. Fakhrutdinova, S. R. Nikishina. The Content of Educational Programs in Technical Universities: Quality of Applying the Modern Professional Standards. *International Journal of Instruction*, 2018, pp.357-370. https://doi.org/10.29333/iji.2019.12124a.

[4] National Qualifications Framework: approved. Republican tripartite commission on social partnership and regulation of social and labor relations. March 16, 2016. https://enic-kazakhstan.edu.kz/uploads/additional_files_items/23/file_en/national-qualifications-framework.pdf?cache=1555068553 (Accessed Date: May 31, 2024).

[5] A. Bakanova, N. Letov et al. The use of ontologies in the development of a mobile e-learning application in the process of staff adaptation, *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*. Vol. 8, Issue 2S10, pp. 780-789, 2019. https://doi.org/10.35940/ijrte.B1144.0982S1019.

[6] Unknown (1975) Minsky's frame system theory, in: Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing - TINLAP '75, Association for Computational Liquistics, Stroudsburg, PA, USA, pp. 104–116, https://doi.org/10.3115/980190.980222.

[7] Zhao, L., Alhoshan, W., Ferrari, A., Letsholo, K. J., Ajagbe, M., Chioasca, E.-V., & Batista-Navarro, R. T. (2021). Natural Language Processing for Requirements Engineering: A Systematic Mapping Study. *ACM Computing Surveys,* 54(3), 1–41. Article 55. https://doi.org/10.1145/3444689.

[8] G. Salton, C. Buckley Term-weighting approaches in automatic text retrieval, *Information Processing and Management*, 1988, Vol. 24(5), pp.513-523. https://doi.org/10.1016/0306-4573(88)90021-0.

[9] J. Teevan. Improving information retrieval with textual analysis: Bayesian models and beyond: MA thesis. – Massachusetts, 2001. – pages 1-121, [Online] http://hdl.handle.net/1721.1/86759 (Accessed Date: July 5, 2024).

[10] Sahar Tahvili, Leo Hatvani. Chapter Three - Transformation, vectorization, and optimization, *Academic Press*, 2022, p. 35-84. https://doi.org/10.1016/B978-0-32-391913-5.00014-2.

[11] G. Salton, A. Wong, C. Yang. A vector space model for automatic indexing, *Communications, ACM*, 1975, Vol. 18(11), pp.613-620. https://doi.org/10.1145/361219.361220.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed to the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP19677319).

Conflict of Interest

The authors have no conflicts of interest to declare.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)