Data-Driven Approach to Curriculum Analysis

Iuri Nasu
IT in Business Department
HSE University
Studencheskaya street, 38, Perm, Perm Krai, 614070
yunas2002@iCloud.com

Mikhail Drobinin
IT in Business Department
HSE University
Studencheskaya street, 38, Perm, Perm Krai, 614070
msdrobinin@yandex.ru

Mark Efanov
IT in Business Department
HSE University
Studencheskaya street, 38, Perm, Perm Krai, 614070
msefanovv@gmail.com

Viacheslav V. Lanin
IT in Business Department
HSE University
Studencheskaya street, 38, Perm, Perm Krai, 614070
vlanin@hse.ru

Abstract**—The choice of an educational program is momentous in young people's lives. Given the shortage of time after exams, applicants usually do not have time to analyze possible educational tracks. Furthermore, it requires a thorough study of learning plans. This research addresses the problem proposing the algorithm to data-driven curriculum analysis based on natural language processing of course names or competences listed in learning plans. Moreover, the intelligent software system architecture is described. The method is tested on the curricula scraped from university websites. In order to store the content a data warehouse has been developed. At this time, there are few studies on this topic. The existing ones are either on the early stages of development or scarce on implementation details. They are briefly discussed in this paper.**

Keywords—Competences, Curricula, Data Analysis, NLP, Data Warehouse, Higher Education.

I. INTRODUCTION

Adolescence is a pivotal point in life. It is during this period that people shape their talents [1]. Therefore, after finishing secondary school, young people need a pathway to follow their calling. For the majority of teenage Russians, a typical way to do so is to obtain a university education.

Due to time pressure and stress, school-leavers can make an incorrect choice regarding an educational program due to a lack of information. The situation is common—it is estimated that nearly half of school-leavers are said to experience difficulties in choosing an educational program [2].

As a consequence, students are often dissatisfied with higher education [3]. This could be avoided if university entrants were more familiar with undergraduate programs, especially their curricula.

Data-driven decision making has already proven its usefulness in the business world, but is not widespread in the education sector. An intellectual system designed to decompose and visualize different parts of learning plans could alleviate the situation and bring about a change.

The curricula of HSE University and ITMO University were scrapped for the research. To store the retrieved data, a data warehouse was implemented. After that, the algorithm for the detection of similar courses was invented and tested.

II. PROBLEM STATEMENT

The issue addressed in this paper is the lack of digestible information about university programs. Curricula are not easily digestible for those unfamiliar with academic management because they contain a large amount of heterogeneous information. The main hypothesis is that the situation might be soothed by a business intelligence software system devised to provide consultancy to young people choosing an academic program.

The objective of the study is the construction of a research prototype of an intelligent system for inquiry into university curricula analysis. To achieve it Natural Language Processing and Data Warehousing methods are to be implemented. Course names and competence definitions listed in learning plans might be used to find similar courses of different programs.

The functional purpose of the system is to allow the user to compare curricula of educational programs by searching for the most similar courses, highlighting the competencies developed during the training, calculating and comparing the number of classroom hours in different programs.

III. RELATED WORKS

Currently, Russia applies a competency-based approach to higher education [4]. Modern Russian educational standards do not contain lists of compulsory courses (except for liberal arts and physical education). This gives universities relative freedom in designing curricula, but imposes the responsibility for proper planning of students' learning activities. The ability to compare educational programs can be a useful tool in the implementation of this concept.

In previous research a combination of Bag of Words and Support Vector Machine was proposed for legal document classification [5]. This approach was subsequently criticized for its insufficient accuracy.

The current task has similarities with the previous one, both focusing on Natural Language Processing, however the main distinction lies in the training data. Legal document classification was based on a predefined dataset whereas the analysis of curricula requires data scraping. It is also a similarity detection task rather than a classification task.

To improve the quality of text processing vector embeddings [6] could be used. They represent the parts of speech in the numerical form, preserving the semantics for effective text comparison and summarization. Among the existing models which could be used to produce the embeddings are Word2Vec [6] and BERT [7], both based on neural networks. The key difference is that the former generates context-independent vectors, whereas the latter produces contextualized embeddings [8].

The Graph Theory methods were implemented to evaluate learning plans [9-11]. The model proposed consists of nodes (courses) and edges that measure the distance between vertices. To calculate the distance, the following formula (1) is provided:

$$w(v,u) = \frac{(\sum cred)}{2N} \times \big(cred(v) + cred(u)\big) \times \\ \times |comp(v) \cap comp(u)|, \qquad (1)$$

where: *w* is the *distance* between courses *v* and *u*,
∑*cred* is the overall *sum of credits* in learning plan,
N is the *number of courses* in curriculum,
cred(x) is the *number of credits* for course x,
{comp(x)} is the *set of competences* for course x.

The quality of the curriculum is then evaluated using graph density and modularity. It is assumed that the optimal model should have moderate modularity and high density.

However, the proposed distance measurement is not a proper algebraic metric since *w(x, x)* is not equal to zero and the triangle inequality is violated. Even though the zero distance is interpreted as the absence of link between courses the motivation behind the formula is ambiguous.

One of the references of the previously explored articles is the study [12], which proposes a method for the formation of an individual educational trajectory using a dynamic equation model. The model is applicable to a student who has already chosen a program.

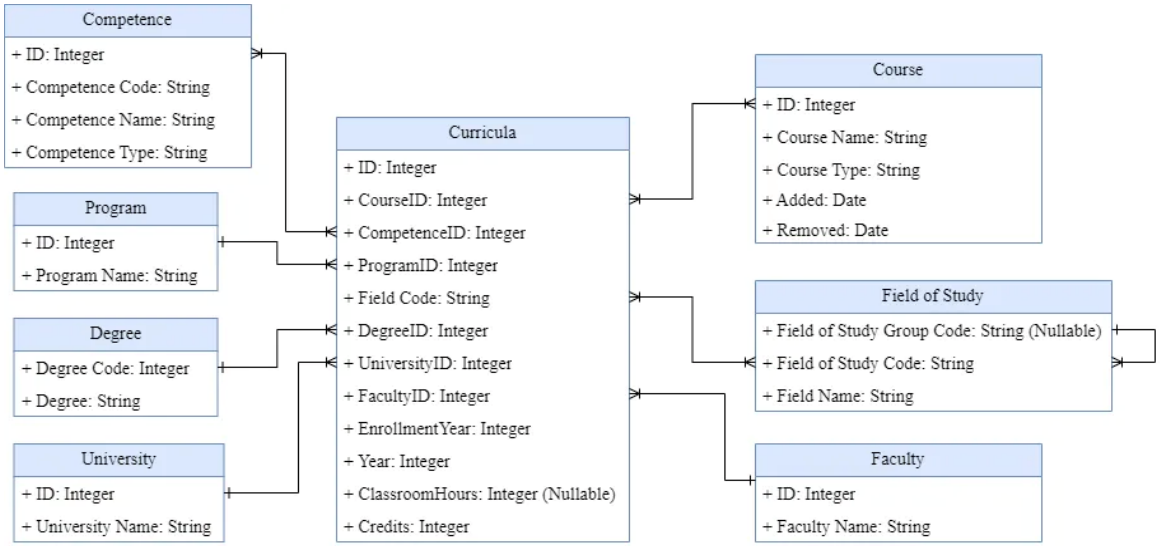

Figure 1. Data Warehouse schema

Therefore, it is not appropriate for the research objectives. Furthermore, the paper does not provide test cases which verify the model's performance.

IV. CURRICULA STRUCTURE AND DATA WAREHOUSE MODELLING

In Russia, Higher Education is regulated by the respective Federal Law [13] and Education Standards [14]. These documents were analyzed to create the Entity-Relationship model (Appendix A). Then it was transformed to a Data Warehouse Star-Schema (fig. 1).

The following dimensions were chosen: *competence, program, degree, university, course, field of study,* and *faculty*.

It was decided to choose following measures: *ECTS credits* and *contact hours.*

A preliminary analysis shows that the average number of lessons a week decreases as the undergraduate student advances (fig. 2).

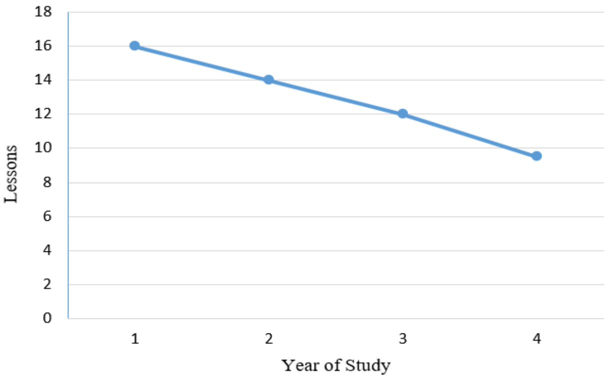

Figure 2. Average number of lessons each year

On average, a first-year student has 16 classes (lectures or seminars) per week, a second-year student has 14 and a third-year student 12 classes per week, and a fourth-year student has 9-10 classes per week.

V. SYSTEM DESIGN

In a nutshell, the system consists of following modules (table 1):

Table 1. Modules and functionality

Module	Functionality
Webpage	User Interaction
API Gateway	Load distribution
Data Access Layer	Request handling and access to learning plans data
Data Warehouse	Storage for learning plans data
ETL Module	Data scraping, processing, transformation to SQL and loading
Machine Learning Model	Advanced analytics

Since the graphical interface is requires for user's convenience, a Single Page Web Application is implemented. In order to access data from the warehouse REST API [15] was developed and deployed in Docker Container [16].

Furthermore Nginx [17] proxy server is used for load distribution since the scalability is required due to the seasonality of university enrollment. ETL module is devised for data scraping. Due to different formats of curricula (pictures, tables, pdf-files) the process is not unified yet. A C4 component diagram is located in Appendix B.

VI. COURSE SIMILARITY MEASUREMENT

In order to find similar courses, the following algorithms might be used:

1. Choose two educational programs and obtain their learning plans.
2. For each course in the learning plans assign the competences it improves, find their description (the code will not suffice) in the educational standard.
3. Using pretrained model, vectorize each competence description.
4. For each course compute the mean vector using formula.
5. Normalize the mean vector to unit length.
6. Find nearest neighbors using L2-norm or cosine similarity.

The course names themselves could be used instead of or combined with competence descriptions for steps 2-4 to find the similar courses.

For the experiment the SBERT model [18] was used. As an example, the Software Engineering and the Business Informatics programs taught at the Perm Campus of HSE University were selected. Their curricula were scraped using Selenium [19] and used for the test (table 2).

The cosine measurements computed for course names and the sets of competences assigned to them are uncorrelated (Pearson and Spearman correlation coefficients equal to -0.11 and -0.13, respectively).

Table 2. Most similar courses

Software Engineering Course	Business Informatics course	Cosine Similarity (names)	Cosine Similarity (comp.)
Top 5 similar by competences			
Group Dynamics	Strategic Management	0.08	0.69
Group Dynamics	Theory and History of Management	0.14	0.69
Group Dynamics	Decision-making	0.04	0.69
Software Design	IT-business infrastructure	0.40	0.65
Databases	Accounting	0.10	0.65
Top 5 similar by names			
Web Programming	Web Programming	1.00	0.44
Safe Living Basics	Safe Living Basics	1.00	0.35
Discrete Mathematics	Discrete Mathematics	1.00	0.10
Research Seminar	Research Seminar	1.00	0.46
Programming	Programming	1.00	0.11

To evaluate the models, expert assessments are used. For that 150 random samples are chosen and ranked manually. Competence-based measurements correlate weakly with expert assessments (Pearson coefficient 0.12), while Name-based measurements show a moderate correlation (Pearson coefficient 0.65).

It is evident that comparison of courses by name gives an imprecise match due to lack of detail, as it favors the general courses (such as History, Law and Economics) or the course with the same name. The name may change over the course of time, but the model would mark them as two different courses. In similar way, it could be taught under various names (e.g., on several educational programs). Though it might be used to provide the overall rough estimate of resemblance.

For this the ratio between the quantity of similar courses (we consider the course similar if the cosine similarity between respective embeddings is higher than 0.65) and the total number of courses for two programs could be computed.

It should be noted that the competence-based approach is inadequate if the data contains inconsistently assigned competences. For example, the above-mentioned Business Informatics' Linear Algebra course is said to develop "positioning products in the global marketplace" (which certainly does not reflect the essence of Linear Algebra).

VII. CONCLUSION

The scientific novelty of the study lies in the use of information technology to analyze curricula. It should be stressed that the project is an investigation into an area of study which is still under-researched. Preliminary research shows that the average number of lessons decreases with the year of study.

Natural language processing methods have been used to compare courses using their names and competences assigned to them.

There were 200 learning plans scraped in purpose of the research. It is planned to scale the project using data from other universities, such as Moscow State University. In future work the other data (such as descriptions of courses and historical data) and methods (such as expert systems) could be used to identify similar courses.

Future work should focus on developing an intuitive user interface based for exploratory search [20]. A guide summarizing the domestic education system and its regulation could be created to improve the user experience.

VIII. USE OF ARTIFICIAL INTELLIGENCE

The online translator DeepL [21] was used to find and correct errors. While most of the text was handwritten, artificial intelligence was used to translate and paraphrase already written text.

REFERENCES

[1] Kırdök O., Harman E. High School Students' Career Decision-making Difficulties According to Locus of Control. *Universal Journal of Educational Research, vol 6, No. 2,* pp. 242-248, 2018. doi:10.13189/ujer.2018.060205.

[2] Kılıç S., Günal Y. University Students' Career Decision Regret: A Mixed-Method Research. *IJERE,* 2023, vol. 8, No. 3, pp. 521-531. doi:10.24331/ijere.1257601.

[3] Prakhov I., Rozhkova K., Travkin P. Osnovnye strategii vybora vuza i bar'ery, ogranichivayushchie dostup k vysshemu obrazovaniyu [Main strategies for choosing a higher education institution and barriers limiting access to higher education]. HSE University, Moscow, Rep. 17. Available at: https://www.hse.ru/data/2022/01/25/1755651210/ib_17_2021.pdf, accessed 01 Apr. 2024 (In Russian).

[4] Kelchevskaya N., Shirinkina E. Integraciya obrazovatel'nyh i professional'nyh standartov v usloviyah reformirovaniya: problemy i puti [Integration of educational and professional standards under conditions of reform: problems and ways of solution]. *Universitetskoe upravlenie: praktika i analiz* [University Governance: Practice and Analysis], issue 1, 2018. pp. 16-25. DOI 10.15826 (In Russian).

[5] Nasu Iu., Lanin V. Development of Legal Document Classification System Based on Support Vector Machine. *Trudy ISP RAN/Proc. ISP RAS*, vol. 35, issue 2, 2023. pp. 49-56. DOI: 10.15514/ISPRAS-2023-35(2)-4.

[6] Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. 2013. Available at: arXiv:1301.3781, accessed 01 Apr. 2024.

[7] Devlin J.; Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018. Available at: https://doi.org/10.48550/arXiv.1810.04805, accessed 01 Apr. 2024.

[8] Shen Y., Liu J. Comparison of Text Sentiment Analysis based on Bert and Word2vec. In Conf. *IEEE 3rd ICFTIC,* 2021, pp. 144-147. Available at: https://ieeexplore.ieee.org/document/9647258, accessed 01 Apr. 2024.

[9] Zykova T., Kytmanov A., Khalturin E. Ob analize uchebnogo plana podgotovki bakalavrov v oblasti informacionnyh tekhnologij [On the analysis of the curriculum for training bachelors in information technology]. In Conf. *New Educational Strategies in the Modern Information Space,* St. Petersburg, 2022, pp. 338-343. Available at: https://www.elibrary.ru/download/elibrary_49318813_30209023.pdf, accessed 01 Apr. 2024 (In Russian).

[10] Zykova T., Kytmanov A., Khalturin E. Analiz formirovaniya kompetencij uchebnogo plana obrazovatel'noj programmy [Analysis of the formation of competences of the educational program curriculum]. In Conf. *Transformation of Mechanics and Mathematics and IT-Education in Terms of Digitalization,* Minsk, 2023, vol. 1, pp. 176-181. Available at: https://www.elibrary.ru/download/elibrary_54184628_52556588.pdf, accessed 01 Apr. 2024 (In Russian).

[11] Zykova T., Kytmanov A., Khalturin E., Noskov M. O sravnenii grafovyh modelej uchebnyh planov podgotovki inzhenerov v oblasti informacionnyh tekhnologij [On the Comparison of Graph Models of Curricula for Training Engineers in the Field of Information Technology]. In Conf. Informatization of Education and E-learning Methods: Digital Technologies in Education, Krasnoyarsk, 2023, pp. 1105-1109. Available at: https://www.elibrary.ru/download/elibrary_54778536_72713754.pdf, accessed 01 Apr. 2024. (In Russian).

[12] Mitsel A., Cherniaeva N. Methods for control over learning individual trajectory. IOP Conf. Ser.: Mater. Sci. Eng., 2015, vol. 91, № 012069. doi:10.1088/1757-899X/91/1/012069.

[13] Federal Assembly of Russia. (2012). No. 273-FZ, Federal Act on Education. Available at: https://www.consultant.ru/document/cons_doc_LAW_140174/, accessed 01 Apr. 2024. (In Russian)

[14] FGOS. Federal Education Standards (Russia). Available at: https://fgos.ru/, accessed 01 Apr. 2024. (In Russian)

[15] Fielding R. *Architectural Styles and the Design of Network-based Software Architectures,* Ph.D. dissertation. University of California, Irvine. 2000. Available at: https://ics.uci.edu/~fielding/pubs/dissertation/top.htm, accessed 02 Apr. 2024.

[16] Docker (2023). Docker: Accelerated Container Application Development. Available at: https://www.docker.com/, accessed 01 Feb. 2024.

[17] Nginx (2023). Nginx. Available at: https://www.nginx.com/, accessed 01 Feb. 2024.

[18] SBERT. (2023). SentenceTransformers Documentation. Available at: https://www.sbert.net/, accessed 01 Apr. 2024.

[19] Selenium. (2023). Selenium Webdriver. Available at: https://www.selenium.dev/documentation/webdriver/, accessed 01 Apr. 2024.

[20] Ruotsalo T., Peltonen J., Eugster M., Głowacka D., Floréen P., Myllymäki P., Jacucci G., Kaski S. Interactive Intent Modeling for Exploratory Search. *ACM Transaction on Information Systems*, vol. 36, issue 4, article 44, 2018. 46 pages. DOI:10.1145/3231593.

[21] DeepL. (2023). DeepL Translator. Available at: https://www.deepl.com/translator, accessed 01 Apr. 2024.

APPENDIX A. ER-DIAGRAM OF CURRICULA

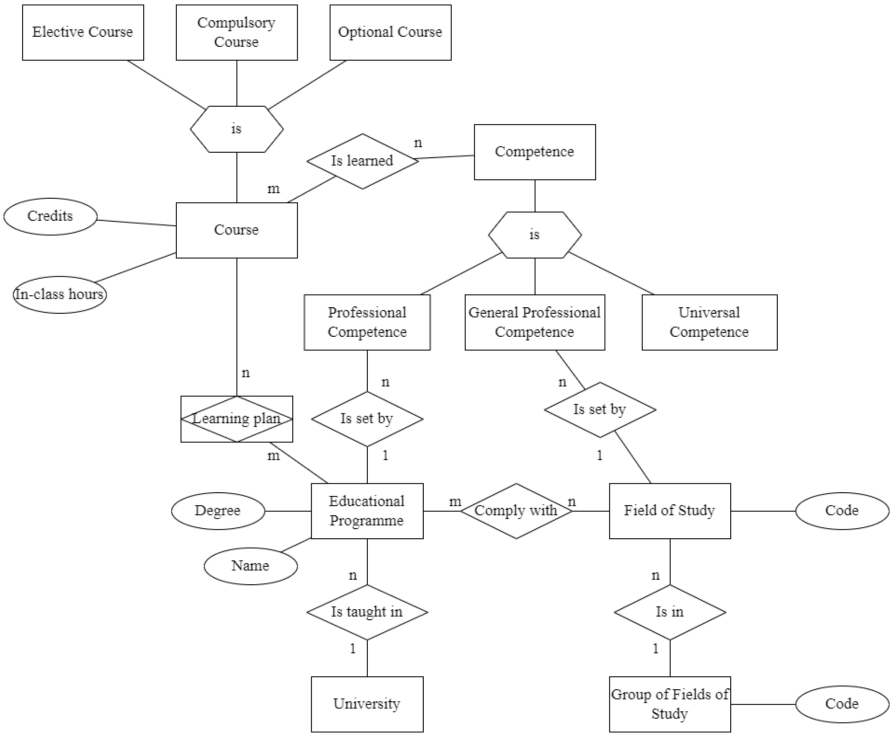

Figure A1. ER-Diagram of a Learning Plan

APPENDIX B. CONTAINER DIAGRAM

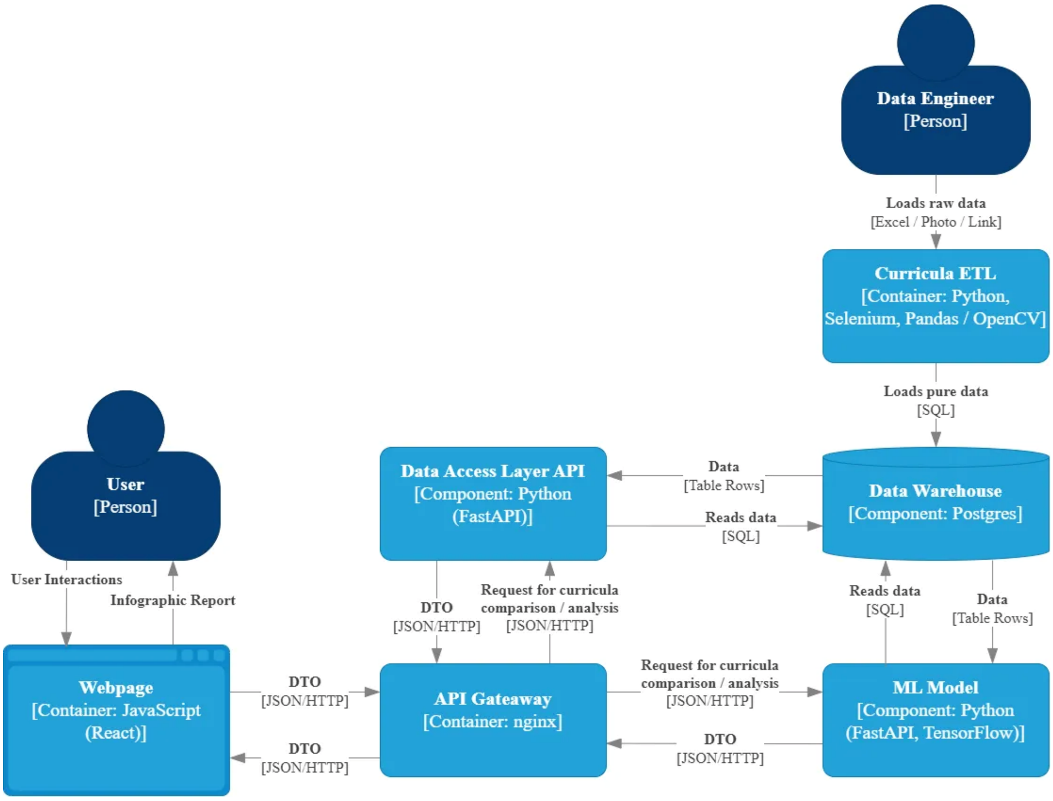

Figure B1. Container Diagram (C4)