The Application of Text Mining Algorithms to Discover One Topic Objects in Digital Learning Repositories

Svetlana Vachkova, Roman Kupriyanov, Ruslan Suleymanov, Elena Petryaeva
Moscow City University
Moscow, Russia
svachkova@mgpu.ru, kupriyanovRB@mgpu.ru, sulejmanovRS@mgpu.ru, petryaeva.elena@gmail.com

***Abstract*—The article considers the process of digital transformation of education and the existing issues in this field. Digital technologies improve the efficiency of backend processes taking place within the unified learning platforms. The article offers possible solutions to the issue of one topic objects search in the learning repositories based on the case study of the Moscow Electronic School – a massive database of learning objects. The solution suggested in the article includes a special text mining model developed to analyze lesson script titles and a programmed algorithm aimed at extracting one topic lesson scripts and visualizing them in two-dimensional space. The efficiency of the developed model is supported by the visualization of 36,644 lesson scripts stored in the E-Library of the Moscow Electronic School.**

I. Introduction

The modern learning environment of Moscow’s schools is extensively integrated with digital space, which shapes a new learning reality and new types of social and educational network relationships. The key role in these processes belongs to the Moscow Electronic School.

The Moscow Electronic School (MES) was launched in 2016 when several educational organizations were connected into one network. Today it is an extensively developing system that will encompass almost all schools of Moscow. The MES requires that every lesson script taught at school is represented by a digital copy in the digital repository and that teachers enroll in professional retraining to acquire competence in digital economics. These requirements are very different from the practices applied in the current education system and call for significant changes of the existing approaches.

The MES provides access to 72,492 teachers, 1,638,275 students, and 1,648,362 parents representing the schools of Moscow and other regions of Russia.

The digital transformation of the education system includes:

- Review or update the objectives and content of education, that must aim at unleashing potential of every student.
- Transition from teaching groups to teaching individuals by changing organization and methods of educational work.
- Review and optimization of the curricular, methodological, organizational and informational content used in the teaching process.
- Description and optimization of the business processes so that they are clear for students and teachers, flexible and scalable.
- Application of capabilities offered by digital technologies to automate business processes and all types of operations with data to raise the efficiency and productivity of teaching and learning activities.

The MES E-Library is a unique repository of learning objects available to students, teachers and parents. There are several definitions about what a learning object is, and we use the following definition: "a digitized entity which can be used, reused or referenced during technology supported learning"[1]. By May 2020, the E-Library contained the following learning objects: 1,413,646 lesson scripts, 154,311 applications, 167,227 tests, 45,321 electronic study guides, and several millions of atomic content.

For several years Moscow City University (MCU) has been conducting research on the MES data. The problems of segmentation and differentiation of collections of learning E-resources have been defined by Remorenko and Grinshkun in “The Frontiers of the Moscow Electronic School” (2017) [2]. The research of 2018–2019 featured analysis of the reasons underlying the demand for lesson scripts [3], network interaction of teachers with digital learning objects in the MES [4]. The crucial role of data in the changes of structure and business processes in education was illustrated using the MES case by Schleicher in “Building a learning culture for the digital world: lessons from Moscow” [5].

A unified data environment will contribute to making the process of teaching clear and systematized by using standardized lesson scripts. Moreover, such educational environment will promote teachers' development including growth of professional and personal competences.

II. Problem statement

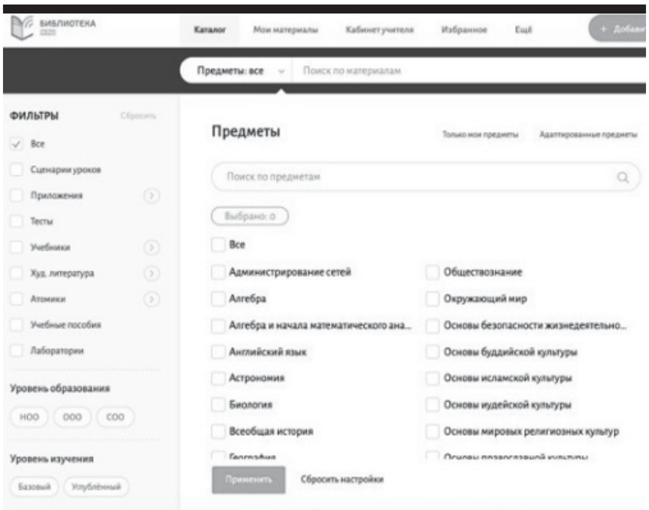
The development and implementation of the MES was aimed at creating a unified digital environment for education, comprising various systems and capable of aggregating teaching and methodological data and making it available for all users. The MES contains a database for teachers that includes a great number of lesson scripts to facilitate the process lesson preparation, as well as repositories of learning objects.

The quality of the created and used lesson scripts becomes of highest importance. The development of digital lesson scripts requires more work, responsibility, accuracy and elaboration compared to analogue lesson scripts. The strict requirements to the quality of uploaded scripts demand involvement of a group of experts including teachers, psychologists, methodologists, software engineers and designers.

The amount of uploaded data is huge compared to its quality. One of the pertinent issues that the MES currently faces is the development of a search engine that is capable of efficient search through the databases. We consider this challenge as a way to improve the innovate digital platform.

The MES E-Library provides users with the system of filters (Fig. 1) that includes:

- filters by the type of object (lesson scripts, textbooks, atomics, applications, fiction, study guides, etc.);
- filters by education level (primary general, basic general and secondary general education);
- filters by proficiency level (basic, advanced);
- filters by subject;
- filters by monitored elements of learning content.



Предметы: все

ФИЛЬТРЫ

☑ Все
□ Сценарии уроков
□ Приложения
□ Тесты
□ Учебники
□ Худ. литература
□ Атомики
□ Учебные пособия
□ Лаборатории

Уровень образования: НОО ООО СОО

Уровень изучения: Базовый Углублённый

Предметы

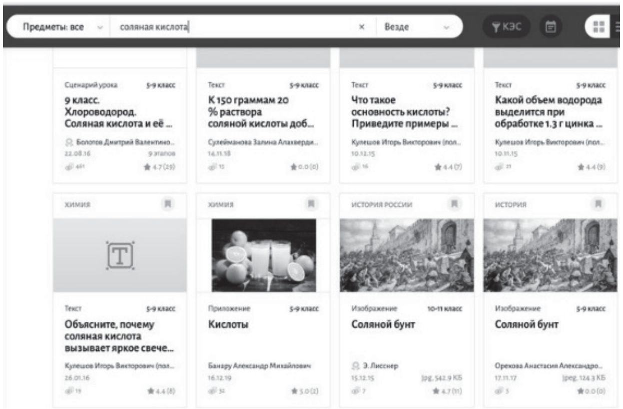
Выбрано: 0

□ Все
□ Администрирование сетей
□ Алгебра
□ Алгебра и начала математического ана...
□ Английский язык
□ Астрономия
□ Биология
□ Всеобщая история
□ География
□ Обществознание
□ Окружающий мир
□ Основы безопасности жизнедеятельно...
□ Основы буддийской культуры
□ Основы исламской культуры
□ Основы иудейской культуры
□ Основы мировых религиозных культур
□ Основы православной культуры

Fig. 1. Teacher’s account in the MES E-Library. Object search

Besides, users can search for objects via the search boxes.

Unfortunately, today the search in the MES E-Library has become quite complicated. The search via the boxes sometimes yields irrelevant objects as results. For example, when searching the MES for “salt acid”, the results include objects referring to the chlorohydric acid and to the Salt Riot – the Moscow uprising of 1648 (Fig. 2); when searching for “magnesium” the results include objects about magnets (Fig. 3).



Предметы: все | соляная кислота

Сценарий урока, 5-9 класс: 9 класс. Хлороводород. Соляная кислота и её ... Болотов Дмитрий Валентино... 22.08.16, 9 этапов, 461, 4.7 (29)

Текст, 5-9 класс: К 150 граммам 20 % раствора соляной кислоты доб... Сулейманова Залина Алахверди... 14.11.18, 15, 0.0 (0)

Текст, 5-9 класс: Что такое основность кислоты? Приведите примеры ... Кулешов Игорь Викторович (пол... 10.12.15, 16, 4.4 (7)

Текст, 5-9 класс: Какой объем водорода выделится при обработке 1.3 г цинка ... Кулешов Игорь Викторович (пол... 10.11.15, 21, 4.4 (9)

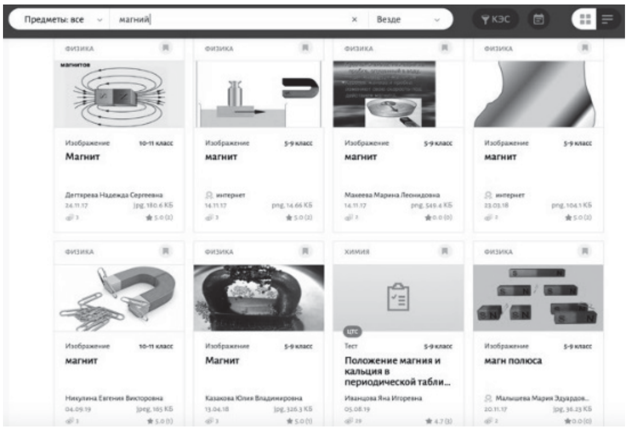
Текст, 5-9 класс: Объясните, почему соляная кислота вызывает яркое свече... Кулешов Игорь Викторович (пол... 26.01.16, 19, 4.4 (8)

Приложение, 5-9 класс: Кислоты. Банару Александр Михайлович, 16.12.19, 34, 5.0 (2)

Изображение, 10-11 класс: Соляной бунт. Э. Лисснер, 15.12.15, jpg, 542.9 КБ, 7, 4.7 (11)

Изображение, 5-9 класс: Соляной бунт. Орехова Анастасия Александро... 17.11.17, jpeg, 124.3 КБ, 3, 0.0 (0)

Fig. 2. The “salt acid” search request



Предметы: все | магний

Изображение, 10-11 класс: Магнит. Дегтярева Надежда Сергеевна, 24.11.17, jpg, 180.6 КБ, 3, 5.0 (1)

Изображение, 5-9 класс: магнит. интернет, 14.11.17, png, 14.66 КБ, 3, 5.0 (2)

Изображение, 5-9 класс: магнит. Макеева Марина Леонидовна, 14.11.17, png, 549.4 КБ, 2, 0.0 (0)

Изображение, 5-9 класс: магнит. интернет, 23.03.18, png, 104.1 КБ, 2, 5.0 (2)

Изображение, 10-11 класс: магнит. Никулина Евгения Викторовна, 04.09.19, jpeg, 165 КБ, 1, 5.0 (1)

Изображение, 5-9 класс: Магнит. Казакова Юлия Владимировна, 13.04.18, jpg, 326.3 КБ, 3, 5.0 (1)

Тест, 5-9 класс: Положение магния и кальция в периодической табли... Иванцова Яна Игоревна, 05.08.19, 29, 4.7 (3)

Изображение, 5-9 класс: магн полюса. Малышева Мария Эдуардов... 20.11.17, jpg, 36.23 КБ, 2, 0.0 (0)

Fig. 3. The “magnesium” search request

The huge amount of objects and the inaccuracy of the semantic analysis system increase the time teachers spend in search for specific objects to use at a lesson. Sometimes search results offer 50 and more pages of learning objects even in case filters are applied. As a result, teachers often choose to upload their own content in the MES E-Library than to use the available data. This increases the amount of repeated content, clutters the database, and complicates the search. Thus, a vicious circle is closed.

In 2019–2020 the MES has been subject to changes connected with the development of the Topical Framework to structure the learning objects by topics and didactic items. The Topical Framework is generated within the MES. However, the amount of the existing learning objects in the MES E-Library is so huge that it is quite impossible to manually assign the objects to the Topical Framework. There is a need for a tool that will streamline the mapping of the MES E-Library objects and the Topical Framework, and reduce the amount of manual work.

III. Background literature

The issue of searching one topic objects in the MES E-Library arose in September 2019 due to the increase in amount of uploaded learning objects and the requirement to assign them to the Topical Framework. One topic objects could be defined as a group of two or more learning objects which are combined based on a single learning issue. The research included the literature review of the existing approaches to the application of semantic analysis algorithms for solving problems of this kind.

The MES database represents a system identical to library systems. To search for data in a digital library one often uses the search by keywords. This approach is actively used since it is one of the simplest and efficient ways to range the search results provided that the wording is clearly coined. The processes of search and extraction of keywords are extensively developed. For example, the methods of machine-learning and neural network training are used to automate the process of keyword extraction from text corpus [6, 7].

The classical method of searching by keywords does not always provide relevant results, especially when a keyword has several meanings or a key phrase is not clearly elaborated. In this case one uses exploratory search which functions in two aspects: the problem context and the search context [8].

To improve the search engine and to adapt the search results to users' requests, the object domain can be described by ontology-based models that include the terms of the object domain and the rules of their usage [9]. The formalized description of the object domain can be used to construct classes, objects and correlations between them. The model-theoretic approach can be used to extract key meanings from the text data and build object domain models [10]. Besides the ontology-based models, the approaches based on Deep Learning are actively developed as they can improve the quality of data search by extracting new attributes and metadata [11]. There are also algorithms for constructing multi-dimensional classification of data resources by using the faceted classification and the system model ontology terms as facets. Such algorithms enable to adequately identify and select data resources in correlation with the system model of the business process run by advanced software applications [12].

The assessment of a search engine aimed at focused search can be conducted by an expert in a specific field who can sometimes be an end user as well. This approach enables to evaluate the actual relevance of search results and receive feedback. The relevance of search results can be defined by various metrics that often include Mean average precision (MAP), normalized discounted cumulative gain (nDCG), mean reciprocal rank (MRR), precision and recall [13].

To search and extract new knowledge from semi-structured data, semantic analysis is used, which also serves as foundation to build new algorithms aimed at improving the quality of text processing. Semantic correlations can be established by enriching the Bag-of-words model and further improvement of its method of data representation [14]. The method of recognizing named entities can reduce the amount of irrelevant data that is offered to the user in case of a request with ambivalent meaning [15]. The method of mining the background knowledge from external sources can be used to enrich the attributive space that describes the object domain [16]. The method of creating keyword sequences enables to define the semantic kernel of the text corpus that is used to find the most coherent texts [17].

The topic model enables to extract implicit categories from the text corpus whereas each text can be allocated to a number of categories and the degree of allocation is defined by the calculated probability [18]. The topic model can be also applied to users or authors. In this case users are grouped into categories based on common interests and their interests are defined by viewed or published documents [19].

Also, methods of working with graph models (graph embedding techniques) that use the representation of graph nodes in vector space and are focused on determining the similarity of topics are widely used. In [20], the most popular techniques for working with graph models are presented, as well as a comparative analysis of the effectiveness of their applicability on several data sets.

Despite the available intellectual algorithms for text data search, the MES as a specific repository requires its own heuristic algorithm [21] to be trained on the MES data, as well as support of metadata – structured data that provides comprehensive description of the content and its key elements [22]. The implementation of this solution is aimed at helping experts to mark up documents and identify thematic groups or micro groups within the subject area, where a convenient visual representation of data and the proximity of elements is available. The marked-up data can serve as a training data set for classification algorithms when new objects are subsequently assigned to one of the groups. The use of probabilistic thematic models is also a possible solution to the tasks set. However, the lack of knowledge about the approximate number of nested clusters and subclusters will not allow to use such solutions, as they will still need to be seriously refined.

IV. Methodology

As has been previously mentioned, at the time of this research, the MES E-Library contained 1,413,646 lesson scripts including 44,527 moderated scripts. The analyzed data comprised 36,644 lesson scripts since only those lesson scripts that had undergone moderation were selected with exclusion of the lesson scripts with identical titles.

Table I shows an example of raw data used for analysis.

TABLE I. EXAMPLE OF ANALYZED DATA

Script_id	Subject	Education level	Topic
34105	Environmental Studies	Primary general	Geographic map and plan
34147	Mathematics	Primary general	Numbers from 10 to 20. Grade 1. Lesson 1.
34214	Environmental Studies	Primary general	Land navigation. Compass
286028	Russian language	Primary general	“Consolidating knowledge on noun cases. “ grade 3
60449	Reading of literature	Primary general	4 gr._V.F.Odoyevsky A town in a tobacco box

We had a hypothesis suggesting that one topic lesson scripts must have similar titles. To test this hypothesis, we developed a model for intellectual analysis of lesson script titles and programmed an algorithm able to extract one topic lesson scripts and visualize them in two-dimensional space. All algorithms have been programmed using Python 3.7 and the following libraries:

- pymssql – extracting data from DB;
- pandas, numpy, colorsys – processing data, generating colors and graphs;
- gensim – word embedding;
- scikit-learn – for dimensionality reduction algorithms;
- plotly, seaborn, matplotlib – work with graphs.

The elaborated model can be conventionally divided into three blocks:

- data pre-processing;
- defining similarity of lesson script titles;
- visualization of results.

The stage of data pre-processing included lemmatization and text stemming, changing the letter case, removing special symbols, such as commas, full stops, dashes, etc; removing conjunctions and prepositions; reducing sentences to word arrays so that every sentence is represented by a set (array) of words.

To define similarity of lesson script titles, we applied the word2vec method [23, 24]. This method is based on training of an artificial neural network using a large text corpus. In this research we conducted training of the algorithm based on 36,644 lesson script titles. The algorithm is aimed at generating a vocabulary for a text corpus and calculating word embeddings based on context similarity of words. Thus, words with close meanings are represented by similar vectors, which are defined by cosine similarity:

$$similarity = \cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}, \quad (1)$$

where *A, B* – word embeddings, *n* – set of real numbers.

By applying the word2vec algorithm we created word embedding for every word within the 50-dimensional space. However, to define similarity of lesson script titles, we had to create embedding for every lesson script title. To calculate embedding of a lesson, all its word embeddings were extracted and added up using the array addition method. As a result, each lesson script title received its embedding in 50-dimensional space.

To visualize the results in the form convenient for human interpretation, we had to reduce the vector dimensionality while maintaining the distance between the vectors. This task was solved by application of t-distributed stochastic neighbor embedding algorithm (t-SNE) [25]. t-SNE is an algorithm for non-linear dimensionality reduction. To maintain the distance between the two data points in space, the algorithm assigns the following similarity measure:

$$p_{j|i} = \frac{exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k\neq i} exp\,(-||x_i - x_k||^2/2\sigma_i^2)}, \quad (2)$$

where x_i – point in the original data space, $\|x_i - x_k\|$ – Euclidean distance between two data points.

The formula (2) shows the proximity of data point x_j to data point x_i with Gaussian distribution around x_i and preset variance σ_i^2. The variance is defined so that the data points located in the high density areas have smaller variance than the data points located in the low density areas.

The t-SNE algorithm aims to reflect $\{y_1, y_2 \ldots y_n\}$ of similarities p_{ij} in d-dimensional space (with $y_i \in R^d$). To do this, it uses the formula:

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k\neq m} (1 + ||y_k - y_m||^2)^{-1}}, \quad (3)$$

whereas q_{ii} equals zero, and

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}. \quad (4)$$

Thus, to solve the task of reducing the vector dimensionality while maintaining the distance between the vectors, the following steps must be performed:

1) Calculating similarity for every two vectors of lesson script titles within the original space using the Gaussian kernel (2, 4).

2) Defining similarity for every two vector reflections in the visualization space using the t-distributed kernel (3).

3) Constructing a target function – cosine similarity (1) and reduce it by the gradient descent method.

After all transformations, we received embedding for every lesson script title in two-dimensional space. This embedding can be easily visualized by the two-dimensional graph where every lesson script is represented by a data point with coordinates (x, y).

V. Results

The visual representation of one topic lesson scripts helps to estimate the correspondence of topics to study subjects (assigning one topic lesson scripts to one-subject clusters), analyze the topical structure of the MES E-Library, as well as perform search of cross-disciplinary lesson scripts which may be elaborated in further research.

Fig. 4 shows the visualization of 36,644 lesson scripts stored in the MES. Each study subject is depicted by a certain color. The figure shows that the majority of lesson scripts is grouped into clusters according to the study subjects they belong to. We can notice that Mathematics and Geometry are located close to Physics and Information Technology, while Information Technology Programming are located far from mathematics due to specific terminology. It is important to note, that the allocation of segments was carried out manually, in accordance to the visualization obtained. But, the process of cluster allocation can be automated if these algorithms are implemented in information systems or software of digital learning repositories. However, there is a number of lesson scripts that are not included in the study subject clusters. We assume that such lesson scripts belong to cross-subject (interdisciplinary) area.

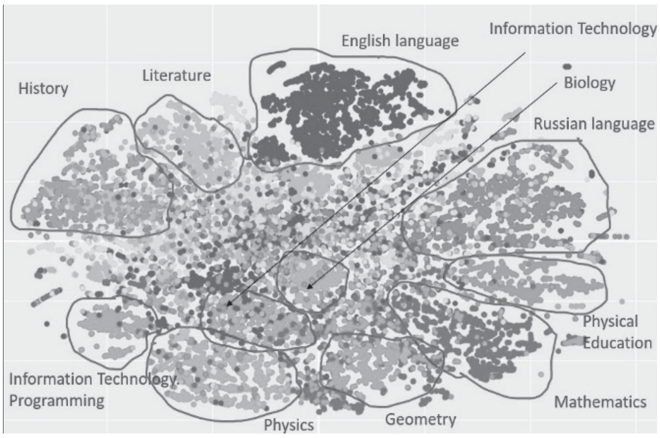


Fig. 4. Visualization of lesson scripts in the MES based on similarity of lesson subjects

In the primary general education database we discovered the following interdisciplinary topics: "The Great Patriotic War", "The world around me", "Sounds and letters", "Moscow studies". In the basic general education the interdisciplinary topics included: "General classification of plants", "Collection of poems about Moscow by Marina Tsvetayeva", "Neurotechnologies", "Healthy lifestyle. Diseases and maladies", "Citizens and their lifestyle", "Occupational safety", "War". For example, the semantic cluster "Occupational safety" included the titles of lessons in Physical Education, Technology, Computer Science. The semantic cluster "War" included the lesson scripts in World History, History of Russia, Literature, Music, for example: "The consequences of war: revolution and collapse of empire", "Russia in the World War I", "The genre of song during the war years".

We discovered, that the algorithm creates semantic clusters based on common topics, as well as on common words. For example, the algorithm included into one cluster the scripts that begin with the words: "What is..." in the lessons "What is computer software?" (Information technology), "What is electricity?" (Physics), "What are rules?" (Social studies). This cluster includes lessons in Technology, Fine Arts, Russian language, Algebra, English language, etc.

The lesson scripts in Geography, History of Russia, Social studies, Biology, Physical education were included into one cluster defined by the keyword "Russia".

The semantic analysis showed that the lesson scripts are grouped by the type of activity reflected in the title.

The multicoloured semantic clusters include lesson scripts for various subjects that have titles with the word "General": English language, Algebra, Biology, World history, Geometry, Information technology, Mathematics, Literature, Social studies, Russian language, Physics, Physical education. The same trend was discovered for the semantic cluster uniting the lessons for monitoring and testing of knowledge. This cluster included various subjects that have titles with the words "evaluation", "final", "entrance", "testing", "practical", "self-study", "laboratory work", etc. For example, "Entrance work in Geometry", "Literature. Grade 5. Evaluation work", etc.

Another semantic cluster was created for the lesson scripts that have titles with the phrase "advisory class", for example, "Advisory class 'Children of War'", "The lesson of victory. Advisory class", etc.

The word "game" was the kernel for the lesson script cluster including "Games with ball" (Physical education), "Mathematical games" (Mathematics), "Ancient Olympic games" (World history). The word "system" was the kernel for the cluster including such subjects as Information technology "Information systems", Geography "Water systems", Astronomy "Planets of the Solar system", History "The system of international relations in the 20th century".

The algorithm sometimes created absurd clusters, for example, it grouped some of the lesson scripts with the titles "Vitamins" and "Lichenes".

The conducted analysis resulted in understanding that we needed to develop a model for algorithm training, and test its functioning with regard to the quality of search results.

The quality of the suggested model of one topic lesson script search has been evaluated by experts represented by the MCU teachers of specific disciplines. The expert evaluation has been conducted by 9 experts on subjects taught at the primary general education level (Russian language, Reading of Literature, Mathematics, Environmental Studies, Physical Education, Music, Fine Arts, Technology, English language); and 19 experts on subjects taught at the basic general education level (Russian language, Literature, English language, French language, German language, Chinese language, Mathematics, Algebra, Geometry, Information Technology, History, Social Studies, Geography, Chemistry, Biology, Physics, Physical Education, Music, Fine Arts). The experts assessed the quality of correlation of lesson script titles within the algorithm-generated semantic clusters (table II, table III) and indicated the typical mistakes in the functioning of the algorithm aimed at its improvement.

TABLE II. THE EXPERT EVALUATION OF THE MODEL OF ONE TOPIC LESSON SCRIPT SEARCH OF THE PRIMARY GENERAL EDUCATION LEVEL LESSONS

№	Subject	An amount of lesson scripts	True	False	True rate, %
1	Mathematics	1887	962	926	50,98
2	Reading of Literature	1334	647	687	48, 50
3	Fine Arts	381	184	197	48,29
4	English language	1308	623	685	47,62
5	Technology	572	247	325	43,18
6	Music	258	110	147	42,63
7	Environmental Studies	1635	667	953	40,79
8	Russian language	1943	710	1233	36,54
9	Physical Education	837	251	586	30

TABLE III. THE EXPERT EVALUATION OF THE MODEL OF ONE TOPIC LESSON SCRIPT SEARCH OF THE BASIC GENERAL EDUCATION LEVEL LESSONS

№	Subject	An amount of lesson scripts	True	False	True rate, %
1	Physical Education	1262	874	388	69,25
2	German language	519	335	184	64,55
3	Chemistry	230	141	89	61,3
4	Mathematics, Algebra, Geometry	1573	822	691	56,07
5	Physics	1759	928	684	52,76
6	Geography	1418	680	738	47,95
7	Chinese language	47	22	25	46,8
8	French language	289	116	173	40,13
9	Social Studies	783	297	486	37,93
10	Russian language	2504	935	1569	37,34
11	Music	263	80	183	30,42
12	Biology	712	194	518	27,24
13	English language	3041	584	2457	19,2
14	Fine Arts	443	85	358	19,19
15	Information Technology	3075	264	2811	8,58
16	Literature	1759	142	1617	8,07
17	History	849	51	798	6

The experts noted that the most frequent errors emerged when the algorithm included lesson scripts into clusters on the basis of:

- similar words, one-root words, interference with prepositions;
- similar lemmas;
- similar word meaning;
- proper nouns;
- generalizations;
- prepositions;
- abbreviations;
- arithmetic symbols;
- Roman letters;
- similar letter groupings.

The algorithm also incorrectly grouped lesson scripts if the titles had many meanings or ambivalent meanings, such as the title of a popular fable "The Dragonfly and the Ant".

The big number of errors was due to the inability of the algorithm to cluster incomplete titles or distinguish the lexical and grammatical categories.

Based on the results of the analysis, we conducted the work on the extraction of keywords relevant for every subject, listing stop-words that must not interfere with the work of the algorithm, training of the algorithm using the MES Topical Framework (the general classification of learning topics and content elements for school subjects).

After this work had been finished, we conducted another semantic analysis of the lesson scripts and received new results.

The Fig. 5–9 show the clusters and individual lesson scripts for study groups in Information Technology (Fig. 5, 6) and Literature (Fig. 7, 9). The Figures clearly illustrate that the developed algorithm can define both clusters at the subject level, and micro-clusters (topics) within the subject field.

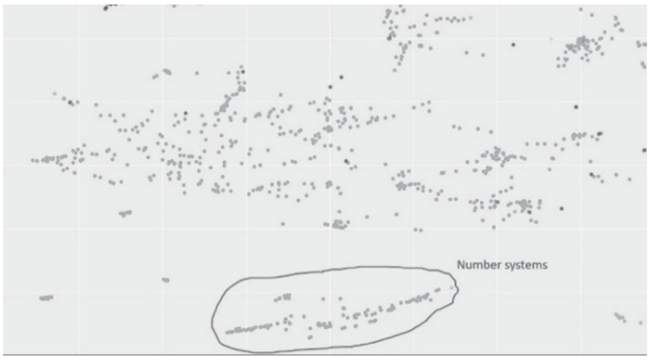


Fig. 5. Information Technology Cluster – number systems

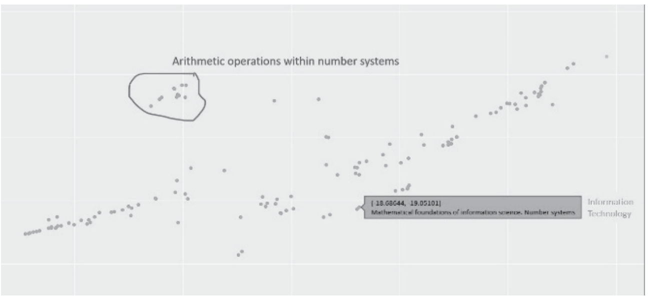


Fig. 6. Information Technology Cluster – number systems. Subcluster – arithmetic operations within number systems

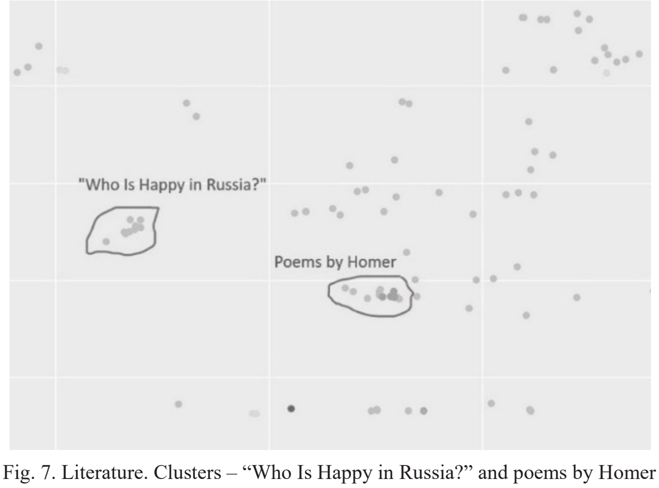


Fig. 7. Literature. Clusters – "Who Is Happy in Russia?" and poems by Homer

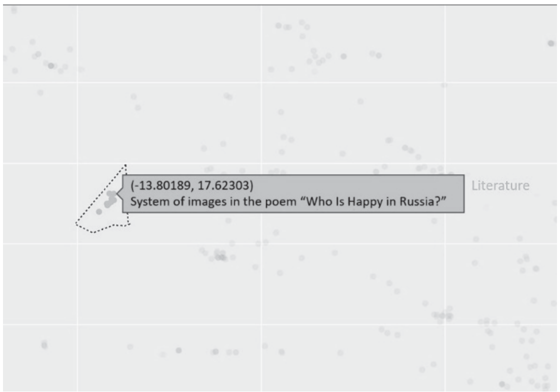


Fig. 8. Literature. Clusters – “Who Is Happy in Russia?” Subcluster – the system of images in the poem “Who Is Happy in Russia?”

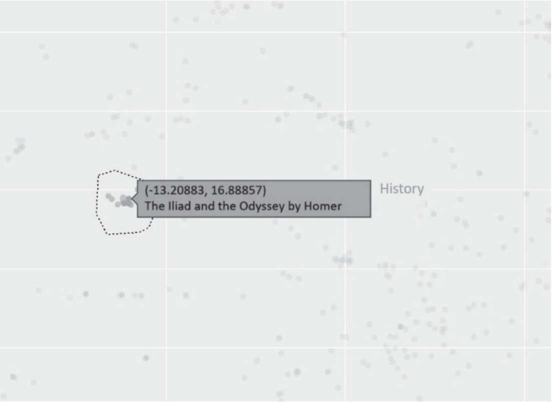


Fig. 9. Literature. Clusters – poems by Homer. Subclusters – the Iliad and the Odyssey

After the training, the algorithm could define topic subclusters within semantic clusters. The analysis of the data regarding Physical education in primary school clearly illustrates this. Figure 10 shows the semantic cluster "Active games with sports elements".

The area 1 includes lesson scripts with the titles: "Basketball elements. Active game. Five throws", "Active games with basketball elements", "Active games with football elements". The area 2 includes lesson scripts with titles: "Active games with basketball elements, Grade 2. 3 lesson scripts", "Active games with basketball elements, Grade 3", "Active games with basketball elements". The area 3 includes lesson script titles with the words: "Active games with volleyball elements". The area 4 includes lesson script titles with the words "Active games with ball pitching elements", "Active games with track and field athletics elements". The area 5 includes lesson script titles with the words "Active games with football elements".

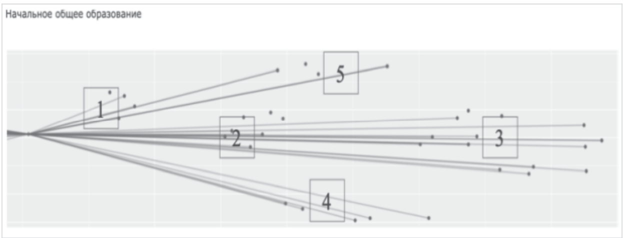


Fig. 10. Semantic cluster "Active games with sports elements"

It is worth noting that since the suggested model is based on the principle of neural network training, the quality of extracting one topic lesson scripts may be improved by training the algorithm on larger text corpus.

VI. Conclusion

The results of the conducted research suggest that the problem of data search in the learning repositories may be solved by applying text mining algorithms. The model developed by this research can be used to solve practical issues of searching one topic lesson scripts stored in the MES E-Library. The research outcomes may advance the frontiers of analyzing the MES learning objects, as well as contribute to the processes of digital transformation of the modern school and education in Russia and other countries.

Moreover, the development and training of the text mining algorithm based on the MES learning objects may further specify the requirements to the Topical Framework of the whole school curriculum, and to the approaches of data structuring based on Big Ideas. The research outcomes may be also used to structure the learning data in various informational educational systems applied in school education. This can facilitate integration of the learning objects uploaded into the MES from different databases and exchange of learning objects between the school databases. As a consequence, this may result in improving the quality of learning content, defining various types of content, acceleration of resources with high teaching potential, removing ‘junk’ learning content.

The further research will be aimed at improving the quality of the model functioning, enriching the methods of its application, implementation of automatic clustering algorithms and comparative analysis of the results of the developed method with the most common approaches: LDA, doc2vec, graph embedding techniques, etc. The marked up by experts data set will allow to use quantitative quality indicators such as precision, recall and F1-measure of algorithms' performance in comparative analysis.

Acknowledgment

This research was supported by the Moscow Department of Education.

References

[1] Rehak, D. R., Mason, R. Keeping the learning in learning objects, in Littlejohn, A. (Ed.) Reusing online resources: a sustainable approach to e-Learning. Kogan Page, London, 2003. (pp.22-30).

[2] V.V. Grinshkun and I.M. Remorenko, “Frontiers of Moscow Electronic School”, *Informatics and Education*, vol. 7, 2017, pp. 3-8.

[3] S.N. Vachkova, V.K. Obydenkova, A.A. Zaslavskiy, and S.V. Kats, “About causes for the “Moscow E-School” lessons scripts relevance”, *Vestnik of Moscow City University*, *Pedagogy and Psychology Series,* vol. 1 (51), 2020, pp. 8-24.

[4] E.D. Patarakin and S.N. Vachkova, “Network analysis of collective operations on the digital education units”, *Vestnik of Moscow City University, Pedagogy and Psychology Series,* vol. 4 (50), 2019, pp. 101-112.

[5] A. Schleicher, OECD Education and Skills Today, Building a Learning Culture for the Digital World: Lessons from Moscow. Web: https://oecdedutoday.com/learning-digital-world-technology-education-moscow.

[6] Adebayo Kolawole John, Luigi Di Caro, and Guido Boella. 2016. A Supervised KeyPhrase Extraction System. In Proceedings of the 12th International Conference on Semantic Systems (SEMANTiCS 2016). Association for Computing Machinery, New York, NY, USA, pp. 57–62.

[7] Rabah Alzaidy, Cornelia Caragea, and C. Lee Giles. 2019. Bi-LSTM-CRF Sequence Labeling for Keyphrase Extraction from Scholarly Documents. In The World Wide Web Conference (WWW '19). Association for Computing Machinery, New York, NY, USA, pp. 2551–2557.

[8] Ya.R. Nedumov, S.D. Kuznezov, *Reseach Seach for Scientific Articles*, Proceedings of the ISP RAS, 2018, pp. 171-198.

[9] C. Obeid, I. Lahoud, H. Khoury, and P.A. Champin, *Ontology-based Recommender System in Higher Education*, 2018, pp. 1031-1034.

[10] J. Chen and J. Gu, "Developing educational ontology: a case study in physics", *in Proc. ICETC Conference*, 2018, pp. 201-206.

[11] I. Safder, S.U. Hassan, A. Visvizi, T. Noraset, R. Nawaz, and S. Tuarob, "Deep learning-based extraction of algorithmic metadata in full-text scholarly documents", *Information Processing & Management*, vol. 57, issue 6, 2020, 102269.

[12] G.G. Kulikova, V.V. Antonova, M.A. Shilina, and A.Z. Fakhrullina, "Structuring Domain Content for Further Data Mining: an example of Forming Structured Content for Education-and-Production Activity", *Information and Control Systems*, vol. 2, 2016, pp. 95-100.

[13] T.A. Nakamura, P.H. Calais, D.d.C. Reis, and A.P. Lemos, "An anatomy for neural search engines", *Information Sciences*, 2018, p. 480.

[14] A.B. Nugumanova, I.A. Bessmertny, P. Pecina, and E.M. Baiburin, "Semantic relations in text classification based on bag-of-words model", *Software & Systems*, vol. 2 (114), 2016, pp. 89-99.

[15] K. Komiya, M. Suzuki, T. Iwakura, M. Sasaki, and H. Shinnou, "Comparison of Methods to Annotate Named Entity Corpora", *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 17, pp. 1-16.

[16] A. Weichselbraun, P. Kuntschik, and A. Brasoveanu, "Mining and Leveraging Background Knowledge for Improving Named Entity Linking", *in Proc. WIMS '18 Conf*, pp. 1-11.

[17] L. Yang, K. Li, and H. Huang, "A new network model for extracting text keywords", *Scientometrics*, vol. 116, 2018, pp. 1-23.

[18] H. Xiong, Y. Cheng, W. Zhao, and J. Liu, "Analyzing scientific research topics in manufacturing field using a topic model", *Computers & Industrial Engineering*, vol. 135, 2019, pp. 333-347.

[19] S. Jung and W. C. Yoon, "An alternative topic model based on Common Interest Authors for topic evolution analysis", *Journal of Informetrics*, vol. 14, issue 3, 2020, 101040

[20] P. Goyal and E. Ferrara, "Graph embedding techniques, applications, and performance: A survey", Knowledge-Based Systems, vol. 151, 2018, pp. 78-94.

[21] M.E. Shwartzman (Ed.), *The Methodology Recommendation on Repository Development*. Moscow: Vashe Tzifrovoe Izdatelstvo, 2018, p. 34.

[22] S.A. Kozlovskiy (Ed.), Open Library. The Recommendations for Libraries on the Publishing Works and Open Licenses Usage in Open Access Regimes, Web: https://goo.gl/jD3J5E.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space", *in Proc. Workshop at ICLR*, 2013, pp. 1-11.

[24] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic Regularities in Continuous Space Word Representations", *in Proc. NAACL HLT*, 2013, pp. 746-751.

[25] L.J.P. van der Maaten and G.E. Hinton, *Journal of Machine Learning Research*, vol. 9, Nov. 2008, pp. 1-48.