INFORMATION TECHNOLOGY, COMPUTER SCIENCE, AND MANAGEMENT

UDC 004.89, 004.032.26 https://doi.org/10.23947/2687-1653-2020-20-4-430-436

Automation of information distribution in adaptive electronic document management systems using machine learning

A. D. Obukhov
Tambov State Technical University (Tambov, Russian Federation)

Introduction. Electronic document management systems (EDMS) are used to store, process and transmit large amounts of information. Automation of these processes is a challenge that requires a comprehensive solution. Its solution will reduce the time and material costs for design and make the transition to a more advanced, adaptive EDMS. The paper is devoted to the development of new methods for automating the process of distributing information in the EDMS. The work objective is to improve the accuracy of the information distribution in the EDMS through moving from analytical or algorithmic solutions to the use of new methods based on machine learning technologies. The application of neural networks in the furtherance of this purpose will also improve the efficiency of software development through automating the analysis and processing of information.
Materials and Methods. A new method of the automated information distribution based on machine learning technologies including a mathematical description of the information distribution rules is proposed. The formulated list of conditions for the information distribution provides the implementation of software based on neural networks for solving the problem of automatic data distribution in the EDMS.
Results. The method of automated information distribution has been tested on the example of the EDMS subject area when solving the problem of analyzing the correctness of information entered by the user. In the course of experimental studies, it was found that the proposed method, based on machine learning technologies, provides better accuracy (8 % higher) and is more efficient (in accordance with the Jilb metrics and cyclomatic complexity).
Discussion and Conclusions. The results obtained confirm the efficiency and accuracy of the method proposed. The presented results can be used to automate the processes of distribution and verification of information in adaptive EDMS, as well as in other information systems. Based on the method developed, it is also possible to solve connected problems: search for duplicates and similar documents, classification and placement by file categories.

Keywords: electronic document management systems, information distribution, automation, adaptability, machine learning.

For citation: A. D. Obukhov. Automation of information distribution in adaptive electronic document management systems using machine learning. Advanced Engineering Research, 2020, vol. 20, no. 4, p. 430–436. https://doi.org/10.23947/2687-1653-2020-20-4-430-436

Funding information: the research is done under grant no. MK-74.2020.9 from President of the Russian Federation.

Introduction. Electronic document management systems (EDMS) are widely used for storing, processing and transmitting information of various types: documents, spreadsheets, graphic and audio information, engineering and accounting documentation, etc. The main functionality for working with documents over the years of the EDMS development has been formed. Further development is aimed at automating human activities, increasing flexibility and reliability, and implementing intelligent decision support modules [1–3].

However, such flexibility and adaptability of the EDMS causes additional time and material costs, increases the complexity of designing and upgrading the system. Therefore, the automation of system design is a challenge for the development of adaptive information systems, including EDMS. This implies many separate tasks for automating the processes of analysis, processing and transmission of documents, the solution of which in totality provides reducing the load on developers during the implementation of adaptive functions in EDMS.

Within the framework of study, the issue on classification of information and its subsequent automatic distribution into specified categories in the EDMS will be considered. This sub-task is one of the most common when organizing electronic repositories and archives, filing cabinets, and filling out document forms. Distribution refers to the placement of data and files in specified positions: by categories, media, directories, and so on [4]. Classical algorithms are not always able to detect an error in the placement of information without involvement of an expert or moderator.

Considering approaches to solving the problem of information distribution, we can distinguish approaches based on the use of machine learning[1]. The application of artificial intelligence technologies for solving classification problems has proved its efficiency in numerous studies and experiments [5-8]. When solving the problem and implementing the information distribution method, machine learning technologies will be used to automate the data classification process.

Based on the analysis of the information movement process, the following urgent tasks can be identified for the classification and distribution of data in adaptive EDMS:

— classification and determination of compliance of the entered data with the category in which the user placed them [9];

— data categorization[2] , and, in case of an error, moving them to the correct categories, or raising a warning to the user [10];

— definition of data duplication by features [11].

To solve these problems, it is proposed to develop a method for automated information distribution in adaptive EDMS, which generalizes existing approaches to information classification and is based on the use of machine learning methods for automating data processing.

Materials and Methods. We formalize the main stages of the method of automated data distribution in adaptive EDMS.

The method is based on the formation and training of a neural network for information classification. Therefore, at the first stage of the method implementation, it is required to prepare a set of information objects $X = \{x_1, ...x_N\}$ for training and testing the neural network. The object $x_i \in X$ can be represented by text or numerical information entered by the user in the form fields, or by files uploaded via the EDMS interface [12]. The data preparation process can include normalization, tokenization, lemmatization, and extraction of file properties and attributes. For the collected data, a set of categories $Y = \{y_1, ..., y_M\}$ is predetermined, that is, there is a continuous display $X \to Y$. We approximate this mapping by neural network NN:

$$NN(X) = Y.$$

With a sufficient amount of training data, it is possible to provide required accuracy of the classification. In accordance with the studies presented in [13], at least 50 copies of training data should be provided for each output feature. Then for $\forall x_i \in X$, we get $NN(x_i) = y_j$, where $j = 1..M$.

At the second stage of the method, information is distributed. Let the EDMS have a set of information objects $X = \{x_1, ..., x_N\}$ and a set of corresponding categories $Y = \{y_1, ..., y_M\}$. To distribute information, the following conditions should be checked:

— if $\forall x_i \in X$, $NN(x_i) = y_j$ and object x_i are placed by the user in category y_j, then the position x_i remains unchanged;

— if $\exists x_i \in X$, $NN(x_i) = y_m$ (where y_m — the category of malicious objects), then object x_i is removed from the EDMS, a warning is sent to the system administrator, the user who added object x_i, is blocked;

— if $\exists x_i \in X$, $NN(x_i) = y_j$, but object x_i is entered by the user in the category y_k

($y_k \neq y_j$), then it is required to redistribute the information in the EDMS.

Consider all possible options for redistribution:

1. Category y_j is free (empty), then object x_i is transferred from category y_k to y_j.

[1] Umadevi S, Marseline KSJ. A survey on data mining classification algorithms. In: 2017 International Conference on Signal Processing and Communication (ICSPC): IEEE, 2017. P. 264-268.

[2] Popova ES, Spitsyn VG, Ivanova YuA. Using artificial neural networks to solve the problem of text classification. In: Proc. 29th Int. Conf. on Computer Graphics and Vision "Graphicon". Bryansk, 2019. P. 270-273. (In Russ.)

2. Category y_j is occupied by some object x_q (such that $NN(x_q) = y_j$), then object x_i is transferred to the buffer, the user is given a warning about incorrect input x_i and duplication of information.

3. Category y_j is occupied by some object x_q (such that $NN(x_q) \neq y_j$), then object x_q is transferred to the buffer, and object x_i is transferred from the category y_k to y_j, the user is given a warning about incorrect input x_q.

The stages and conditions of information distribution considered in the framework of the method include the main scenarios for adding information to the EDMS. Through the use of machine learning technologies, it is possible to automate the distribution of information into specified categories.

Research Results. To implement and integrate the method into adaptive EDMS, we will use a microservice approach. The neural network is implemented using the Keras library (Python), then it is imported into a microservice, which can be implemented on the basis of any framework, for example, Flask. Data between the EDMS and the microservice is transmitted through the HTTP Protocol in JSON format, which provides compatibility with any EDMS implementation [14].

As an example of the data distribution task, we will use a document card form in the EDMS with 5 fields: document name; document author; contact information (address) of the author; date of creation; document description. In the course of the pilot study, the input of user data into the form and filling in the information in the specified 5 fields will be simulated. The neural network will classify the entered information and check it for compliance with the specified category.

To create a set of source data, we use the generator based on open database of Russian names, cities and countries, as well as sequences of Russian words and figures. This will enable to get constructions that correspond to real data, but do not store personal data of real people. Text data will be processed using a tokenizer to convert it to a numeric format. The maximum sentence length is limited to 20 words. We generate a training sample. We will generate 10,000 elements for each category. We will also add 50,000 items with incorrect and erroneous data. Thus, we get an array of 100 thousand elements.

The relationships between input attributes and categories are shown on the heatmap in Fig. 1. The map provides visualization of the correlation matrix and performing a visual analysis of the data to determine how the parameters affect each other and the output variables [15].

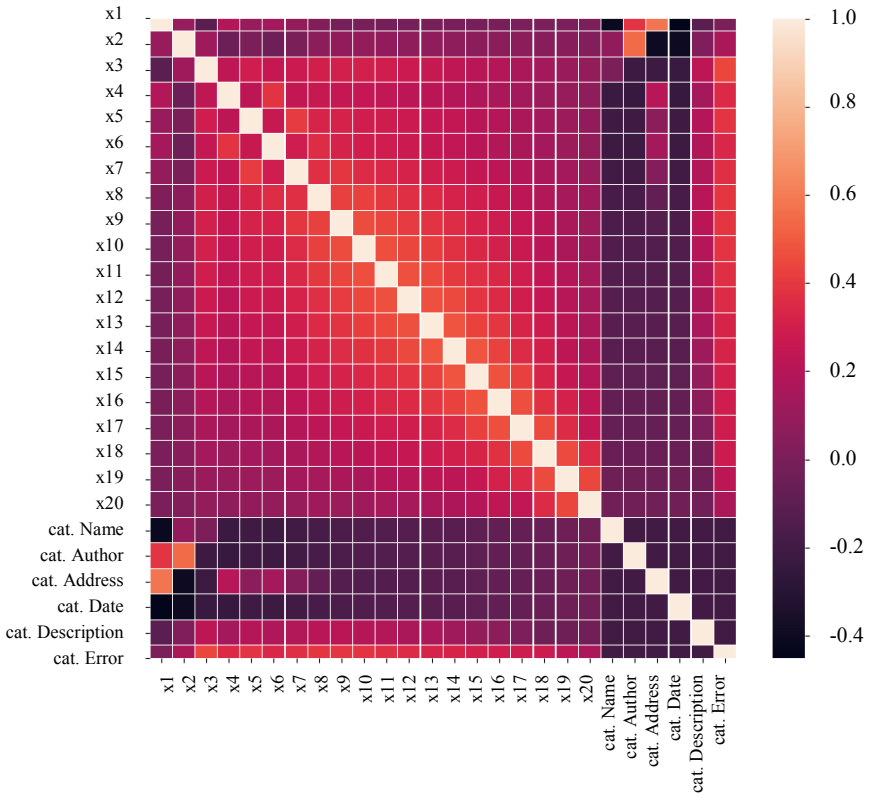

Fig. 1. Raw data heatmap

The network learning process is shown in Fig. 2. The final accuracy of the neural network on the test set after 5 epochs amounted to 97.8%.

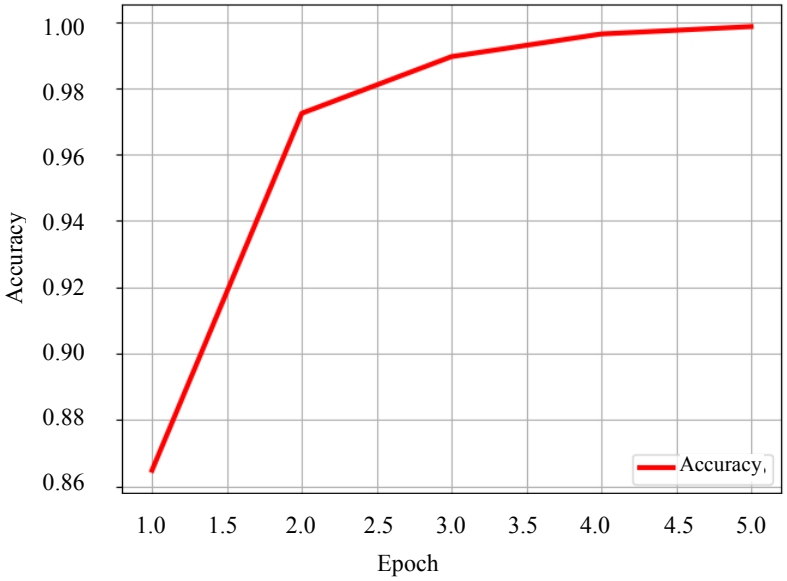

Fig. 2. Classification neural network learning process

The number of errors did not exceed 7 % for the "Author" category and 1 % for the "Description" category. In other categories, incorrect data was recognized in 100% of cases.

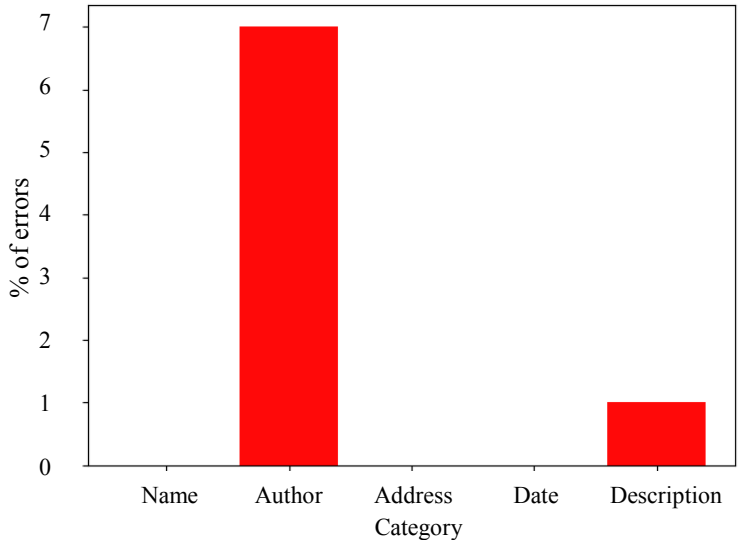

Fig. 3. Testing the neural network for incorrect data entry

The second experiment (Fig. 4) consists in entering data from j categories to i ($j \neq i$). A set of 100 elements for each category i consisting of elements of other categories (25 for each category) is formed. The graph shows the number of tests in which the neural network incorrectly recognized data from other categories as corresponding to the current category. In the second experiment, 1 % of errors was allowed in the "Name" category. The remaining categories were worked out without errors.

Next, we compare the accuracy of the proposed method (hereinafter referred to as the "neural network method") with the classical solution to the problem of classification and distribution of information (referred to as the "classical method"). For the classical method, the following results were obtained: 16 % of errors in the "Name" category, and 25 % — in the "Description" which is much worse than indicators of the neural network method (Fig. 5). On average, the use of neural network methods provides accuracy increase by 8 %.

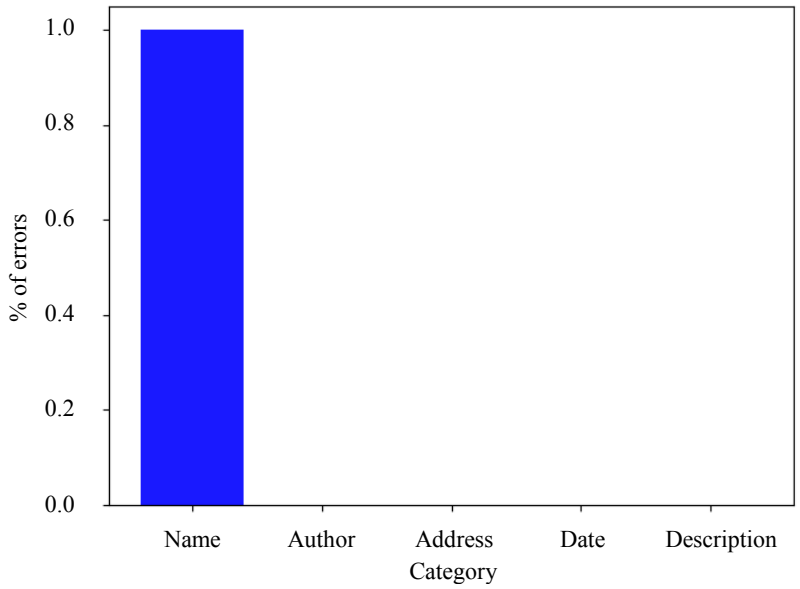

Fig. 4. Testing the neural network for incorrect data entry

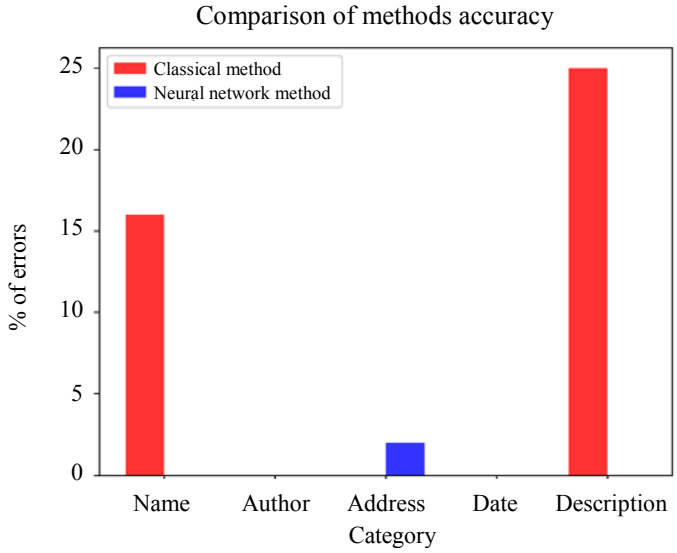

Fig. 5. Comparison of accuracy of neural network and classical methods

Next, we will compare the methods by metrics of software implementation complexity. The cyclomatic complexity expressed by McCabe number (total and averaged) [16], as well as the complexity according to the Jilb metrics [17], are used as metrics. The final results are presented in Table 1.

Table 1

Comparison of classical and neural network methods

Metrics	Classical method	Neural network method
Cyclomatic complexity (total)	26	22
Cyclomatic complexity (averaged)	A (3.71)	A (1.37)
Jilb metrics	0.8	0.15

Thus, the complexity of the neural network method in terms of cyclomatic complexity and Jilb metrics is lower. The accuracy of the developed method is almost 8% higher. It is worth noting that the complexity of implementing the classical and neural network methods of data classification and distribution is relatively small, so, they can be considered comparable. However, the obtained results on classification accuracy confirm the high efficiency of the neural network approach.

Discussion and Conclusions. The paper sets the task of automating the processes of data classification and distribution in adaptive EDMS. The proposed method of automated data distribution includes the formalization of the classification and distribution of data, the use of machine learning technologies to automate the solution to the problem. The formulated list of information distribution conditions provides implementing the software based on neural networks that solves the problem of automatic data distribution.

To test the developed method, experimental studies were conducted on the basis of generated data on documents in the EDMS. The accuracy of the trained neural network was about 98 %. Additional tests have shown that the neural network can detect incorrectly distributed data in almost 100 % of cases, and, under the worst conditions, the error did not exceed 7 %. Thus, the efficiency and accuracy of the proposed method is confirmed. The developed method, in comparison to the classical implementation based on algorithmic support, shows the following positive effect: an increase in average accuracy by 8 %, a decrease in the complexity of the implementation.

The presented results can be used to automate the processes of distribution and verification of information in adaptive EDMS, as well as in other information systems. In addition, on the basis of the proposed method, it is possible to solve related tasks: search for duplicates and similar documents, classification and placement by file categories.

References

1. Kuznetsova EV. Aktual'nye problemy ehlektronnogo dokumentooborota v organakh vlasti [Topical problems of electronic document management in the bodies of power]. Management Issues. 2013;4:73–77. (In Russ.)

2. Zhong RY et al. Intelligent manufacturing in the context of industry 4.0: a review. Engineering. 2017;3(5):616–630. DOI: 10.1016/J.ENG.2017.05.015

3. Xu D, et al. Enhancing e-learning effectiveness using an intelligent agent-supported personalized virtual learning environment: An empirical investigation. Information & Management. 2014;51(4):430–440. DOI:10.1016/j.im.2014.02.009

4. Kuznetsov SD, Poskonin AV. Raspredelennye gorizontal'no masshtabiruemye resheniya dlya upravleniya dannymi [Distributed, horizontally scalable data management solutions]. Proceedings of ISP RAS. 2013;24:327–358. (In Russ.)

5. Krasnyansky MN, Obukhov AL, Solomatina EM, et al Sravnitel'nyi analiz metodov mashinnogo obucheniya dlya resheniya zadachi klassifikatsii dokumentov nauchno-obrazovatel'nogo uchrezhdeniya [Comparative analysis of machine learning methods for solving the problem of classification of documents of a scientific and educational institution]. Proceedings of Voronezh State University. Series: Systems Analysis and Information Technologies. 2018;3:173–182. (In Russ.)

6. Karampidis K, Papadourakis G. File type identification-computational intelligence for digital forensics. Journal of Digital Forensics, Security and Law. 2017;12(2):6. DOI: 10.15394/jdfsl.2017.1472

7. Kim D, et al. Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. Information Sciences. 2019;477:15–29.

8. Zheng J, et al. Hierarchical neural representation for document classification. Cognitive Computation. 2019;11(2):317–327. DOI:10.1007/s12559-018-9621-6

9. Bodström T, Hämäläinen T. State of the art literature review on network anomaly detection with deep learning. In book: Internet of Things, Smart Spaces, and Next Generation Networks and Systems. Springer, Cham; 2018. P. 64–76. DOI: 10.1007/978-3-030-01168-0_7

10. Datta S, Das S. Near-Bayesian support vector machines for imbalanced data classification with equal or unequal misclassification costs. Neural Networks. 2015;70:39–52. DOI: 10.1016/j.neunet.2015.06.005

11. Irolla P, Dey A. The duplication issue within the Drebin dataset. Journal of Computer Virology and Hacking Techniques. 2018;14(3):245–249. DOI: 10.1007/s11416-018-0316-z

12. Goldberg Y. Neural network methods for natural language processing. Synthesis Lectures on Human Language Technologies. 2017;10(1):1–309. DOI: 10.2200/S00762ED1V01Y201703HLT037

13. Beleites C, et al. Sample size planning for classification models. Analytica chimica acta. 2013;760:25–33. DOI:10.1016/j.aca.2012.11.007

14. Obukhov A, Krasnyanskiy M, Nikolyukin M. Algorithm of adaptation of electronic document management system based on machine learning technology. Progress in Artificial Intelligence. 2020;9:287–303. DOI: 10.1007/s13748-020-00214-2

15. Bazgir O, et al. Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks. Nature Communications. 2020;11(1):1–13. DOI: 10.1038/s41467-020-18197-y

16. Luo A, et al. A Structural Complexity Metric Method for Complex Information Systems. JSW. 2019;14(7):332-339. DOI: 10.17706/jsw.14.7.332-339

17. Smirnov AV. Metody otsenki i upravleniya kachestvom programmnogo obespecheniya [Methods of software quality assessment and management]. Izvestiya SPbGETU “LETI”. 2019;2:20–25. (In Russ.)

Submitted 29.10.2020
Scheduled in the issue 27.11.2020

About the Author:

Obukhov, Artem D., associate professor of the Automated Decision Support Systems department, Tambov State Technical University (106, Sovetskaya St., Tambov, 392000, RF), Cand.Sci. (Eng.), associate professor, ResearcherID: M-9836-2019, ScopusID: 56104232400, ORCID: http://orcid.org/0000-0002-3450-5213, Obuhov.art@gmail.com

The author has read and approved the final manuscript.