Academicia Globe:
Inderscience Research

ISSN: 2776-1010 Volume 2, Issue 6, June, 2021

PROSPECTS FOR USING ARTIFICIAL INTELLIGENCE TECHNOLOGIES IN DOCUMENT AUTOMATION SYSTEMS

Muminova Sunbula Shaxzodovna
Assistant of the Department of «Providing information Security", Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan.
sunbula.axmedova@gmail.com.

Asatov Ma'rufjon Azimjon o'g'li
Master Student of 1st Degree of the Department of «Providing information Security", Tashkent University of Information Technologies
Named after Muhammad al-Khwarizmi, Tashkent., Uzbekistan.
marufjon.asatov@mail.ru.

ABSTRACT
Automation of the processes of analysis, processing and transmission of information in the development of information systems will reduce the complexity of implementation, time and material costs, free up the resources of developers to solve more complex and creative problems. One of the ways to automate these processes is the use of machine learning methods, however, without a formalized methodological and mathematical apparatus, it is impossible to provide a comprehensive solution to the problem posed.The article discusses the capabilities of artificial intelligence technologies in relation to the familiar and understandable to all tasks of automating the workflow of enterprises. And also, the article discusses the issues of design automation of adaptive electronic document management systems (EDMS).

Keywords: neural network architecture, neural network technologies, design automation, electronic document management systems

INTRODUCTION

The transition to electronic document management, although it speeds up the processing of documents in a radical way, does not change the essence of these processes. The transition from paper journals in office work to electronic ones, replacing the imposition of a resolution "on paper" with the ability to create a task for a document on a tablet device, automatic fixing of an electronic approval journal instead of maintaining a paper approval sheet - this is certainly more convenient, but does not change the essence of the processing process ... However, there are opportunities for cardinal changes in these processes and a corresponding increase in their efficiency, which opens up huge prospects for the development of EDMS. What are they?

On the one hand, when switching to electronic document flow in the information system, day by day information is accumulated about already implemented scenarios of information processing: data on what typical operations were performed by system users when processing documents, what decisions

ISSN: 2776-1010 Volume 2, Issue 6, June, 2021

were made by participants in business processes based on the content of documents, and other information that accompanies these business processes. An important characteristic of this information, in contrast to similar information in paper workflow, is that it is available for machine processing and can serve as a source material for the application of machine learning technologies.
On the other hand, the rapid development of artificial intelligence technologies, and machine learning in particular, has made the use of these technologies available and relatively inexpensive to create specialized solutions. It is they who will ensure the maximum efficiency of work with documents and easily relieve employees of routine operations.

CLASSICAL APPROACHES TO THE IMPLEMENTATION OF ARTIFICIAL INTELLIGENCE SYSTEMS

In the processes of document flow, two of the most laborious operations can be distinguished - the translation of documents from paper to electronic machine-readable form and the search for documents. It is no coincidence that these two areas attracted the attention of developers in the first place.
Recognition and document search technologies are no longer something new and are very widespread, but recently they have acquired a new sound associated with the development of AI technologies. If the traditional recognition tasks were reduced to digitizing individual letters and symbols - full-text recognition, and the limit was the parsing of their semantics, based on the binding of character sets to certain positions in the paper form of a document (form recognition), then now artificial intelligence systems allow to do so much more. For example, to select separate semantic data from the document not in accordance with the binding to the position in the text of the document, but in accordance with their meaning.
Thus, the Compreno platform, developed by the Russian company ABBYY, provides developers with mechanisms that provide the ability not only to translate paper documents into a machine-readable form, but also to extract individual words and related expressions with certain semantics from flat text that is not presented in a structured form. For example, in the text of the contract, attributes and phrases can be highlighted that characterize the subject of the contract, the legal addresses of the counterparties and the names of the responsible persons, the amount of the contract and other structured data. For this, special technologies are used for high-level semantic text analysis based on so-called ontologies (special dictionaries describing certain subject areas). The creation of these ontologies is carried out by linguistic specialists, and the system transfers their knowledge of the described subject areas to the field of computer technology. This example illustrates one of two classical approaches to the implementation of artificial intelligence systems - the so-called top-down approach, which allows high-level psychological processes occurring in a person's consciousness to be simulated in a computer system. These technologies of semantic parsing of text and the selection of individual semantic entities can be used not only for automatic search for attributes (metadata) in documents, but also for solving other problems, for example: intelligent search for documents not based on syntactic analysis (the presence of certain lexical structures and their variations in the text), but based on the

ISSN: 2776-1010 Volume 2, Issue 6, June, 2021

meaning of the search query; for tasks of automatic classification of the flow of incoming documents: for example, to determine the storage location or launch certain processes of their processing, etc.
The second approach - bottom-up, models intelligence based on analogs of its biological structural elements - the so-called neural networks - and allows you to implement machine learning mechanisms.

USE OF MACHINE LEARNING IN EDMS

Machine learning is a way to identify hidden patterns of making certain decisions based on an array of accumulated data. Simplified machine learning mechanism can be demonstrated with the following example. We have a reference array of documents that are manually categorized based on their content. To apply machine learning technologies, you need to make an assumption about the decision criteria. Let's say documents belong to one category or another, based on the presence of certain keywords in the text, and a certain set of metadata. When an assumption about a decision-making model is formed, a virtual neural network with an undefined internal structure can be generated, which receives data on the presence of keywords in a document instance from a reference array as input, and its output is the assignment of a document to a particular category.
The process is performed within the recommended life cycle, allowing you to structure data processing and analysis projects. This life cycle represents the milestones that are typically completed by projects, often iteratively:

1. Commercial aspect.
2. Obtaining and analyzing data.
3. Simulation
4. Deployment
5. Customer acceptance

Purpose of use:

- Determine the optimal data characteristics for a machine learning model.
- Create an informative machine learning model that most accurately predicts the target.
- Create a machine learning model suitable for the production environment.

Methods of execution:
At this stage, three main tasks need to be solved.

- Designing features. Generate data characteristics from the raw data to make it easier to train your model.
- Training the model. Find the model that most accurately answers the question posed by comparing the success metrics for the models.
- Determine if the model will be suitable for the working environment.

A visual representation of the life cycle of the group data processing and analysis process is shown in Figure 1:

Academicia Globe:
Inderscience Research

ISSN: 2776-1010 Volume 2, Issue 6, June, 2021

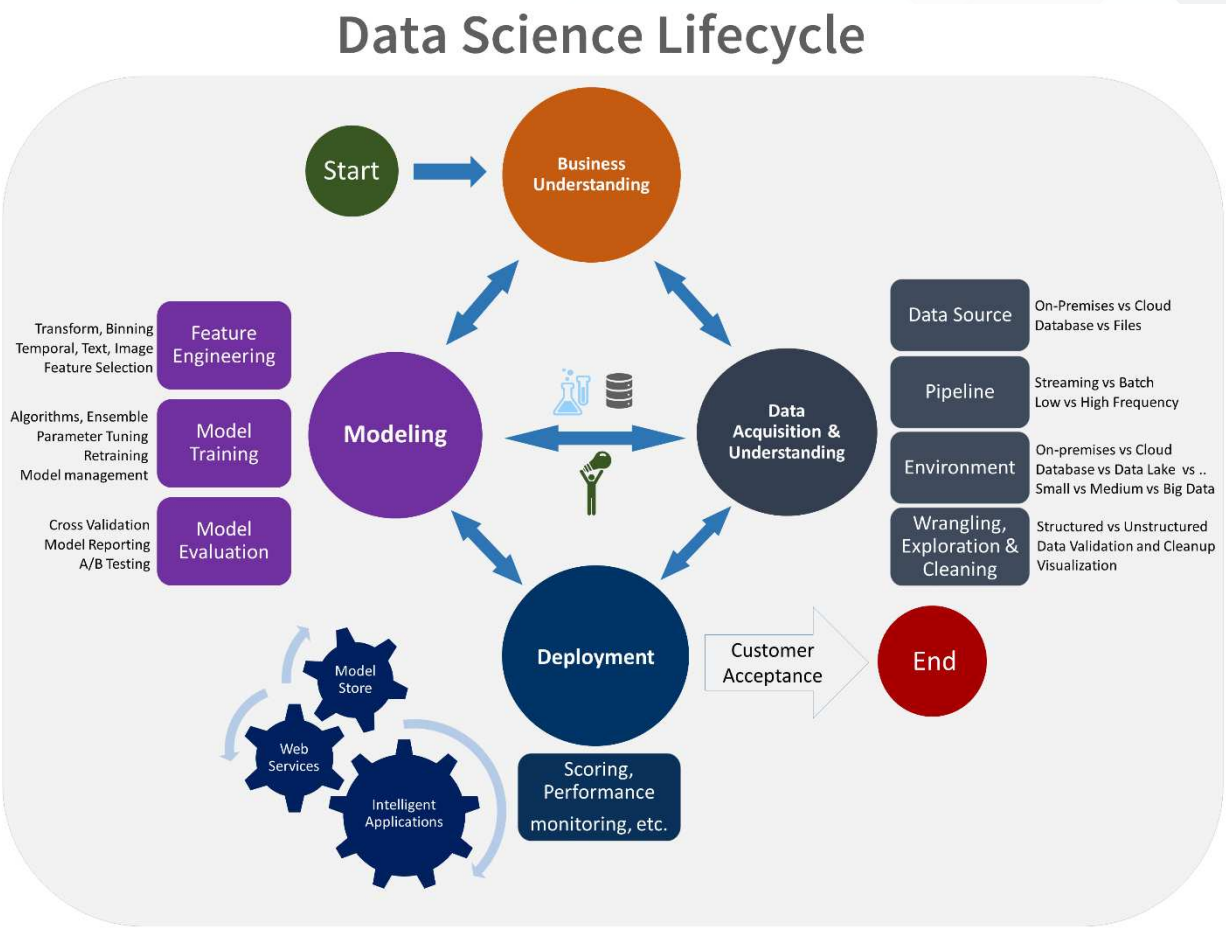

Fig. 1. The life cycle of the data processing and analysis process

METHODS OF TRAINING AND ASSESSMENT OF ITS QUALITY

As mentioned above, the main characteristic of systems developed using machine learning methods is the ability to learn. Depending on the types of problems to be solved, various algorithms are used to implement this key feature. Within the framework of this article, we will consider the type of unsupervised learning, and also define the classes of problems that are suitable for this type.

Initially, the structure of the network is not defined. The learning process is as follows: a specific copy of a document from the reference sample is fed to the input, for which the corresponding category is fixed at the output - as a result, the structure of the neural network is flashed. Starting from a certain step of learning (formation of the network structure), it can already begin to predict the result (in our case, whether the document should be assigned to one category or another). If the assumptions about the decision criteria were made correctly, then as the training sample grows, the probability of an adequate prediction should increase, otherwise it is necessary to change the hypothesis about the criteria.

Formally, the formulation of the unsupervised learning problem can be described as follows. Let X be a set of data - descriptions of some objects. It is necessary to find a set Y consisting of interconnections

Academicia Globe:
Inderscience Research

ISSN: 2776-1010 Volume 2, Issue 6, June, 2021

f: (x, x ') between objects from X (x, x' ε X, f ε Y). The quality of identifying relationships is checked by some metric selected based on the problem being solved. Unsupervised learning is used to solve the following types of problems:

1. The task of clustering.
2. Search for association rules.
3. Reducing the dimension of the data.
4. Data visualization.

To a certain extent, each of the last three tasks is a derivative of the first or its particular case. Let us consider in more detail the formulations of these tasks. The task of searching for association rules means identifying in the feature descriptions of objects (source data) such sets and values of features that are especially often (not by chance often) found in the source data. If we draw an analogy with the first task, then each rule in this case can be represented as a cluster.

The task of reducing the dimension of data is as follows. There is a large (much larger) volume of feature descriptions of objects. Moreover, this volume is due to the impressive number of dimensions of the feature space. It is necessary to present the same data in a space of a lower dimension, while minimizing the loss of information. Grouping by clusters will be one of the options for solving the problem. The task of data visualization is, in fact, a special case of the previous one: its goal is to present the initial data in a displayed space, that is, a space of dimension 2 or 3. As follows from the above, unsupervised learning in one way or another comes down to clustering. Therefore, to assess the quality of training in this way, as a rule, clustering quality metrics are used. Moreover, when choosing them, it is taken into account that these metrics should not depend on the initial data, but only on the partitioning results. All quality assessments can be divided into external and internal. The former use external information about the true division of objects into clusters, the latter rely only on a set of initial data, that is, these metrics can work with an unlabeled sample when the true division of objects into groups is not known in advance. And it is with their help that the optimal number of clusters is determined. Let's take as an example the metrics allocated by ODS [2]:

1.Adjusted RandIndex (ARI). This metric belongs to the group of external ones: it is assumed that the true labels of objects are known (for example, set by an expert), but it does not depend on the values of the labels themselves. Only from dividing the sample into clusters.

The measure is calculated as follows: let n be the number of objects in the sample, a - the number of pairs of objects with the same labels and located in the same cluster, b - the number of pairs of objects with different labels and located in different clusters. Then you can calculate the proportion of objects for which the initial and resulting partitioning are consistent:

$$\mathbf{RI} = \frac{\mathbf{2(a+b)}}{\mathbf{n(n-1)}}$$

Academicia Globe:
Inderscience Research

ISSN: 2776-1010 Volume 2, Issue 6, June, 2021

The resulting value is called RandIndex (RI) and expresses the similarity of two different clusterings of the same sample. For this index to give values close to zero for random clustering for any n and the number of clusters, it is necessary to normalize it, that is, to obtain the AdjustedRandIndex:

$$\mathbf{ARI} = \frac{\mathbf{RI} - \mathbf{E}[\mathbf{RI}]}{\mathbf{max}\ (\mathbf{RI}) - \mathbf{E}[\mathbf{RI}]}$$

where E is the mathematical expectation.
The ARI measure is symmetrical and does not depend on permutations and label values. In fact, this index is a measure of the distance between different sample partitions. Its range is [-1,1]. Its intervals can be interpreted as follows: for "independent" partitions into clusters - negative values, for random partitions - close to zero, for similar partitions - positive values, moreover, ARI = 1 indicates the coincidence of partitions ...

2.Adjusted MutualInformation (AMI). This metric is similar to the previous one: it is also symmetric and does not depend on the values and permutations of labels. The entropy function is used to determine it. Sample splits are interpreted as discrete probabilities: the probability of being assigned to a cluster is equal to the fraction of objects in it. As in the previous case, for this metric it is necessary to calculate a special index - MutualInformation (MI). This value is defined as mutual information for two distributions corresponding to the division of the sample into clusters. This can be interpreted as a piece of information common to both partitions: how much information about one of them reduces the uncertainty about the other. This index is calculated using the following formula:

$$\mathbf{MI}(\mathbf{X}, \mathbf{Y}) = \mathbf{\Sigma}_{\mathbf{y}\in\mathbf{Y}}\mathbf{\Sigma}_{\alpha\in\mathbf{X}}\mathbf{p}(\mathbf{x}, \mathbf{y})\mathbf{lo}\,\mathbf{g}\left(\frac{\mathbf{p}(\mathbf{xy})}{\mathbf{p}(\mathbf{x})\mathbf{p}(\mathbf{y})}\right)$$

where p (x, y) is the joint probability distribution function for X and Y; p (x) and p (y) are the limiting probability distribution functions for X and Y, respectively.
The area of the AMI index value is the range [0,1]. It is interpreted as follows: values close to zero indicate the independence of the partitions, and those close to one indicate their similarity (or coincidence with AMI = 1).
3. Homogeneity, completeness, V-measure. These metrics consider sample splitting as discrete distributions and are determined using the entropy function and conditional entropy.
Homogeneity determines how much each cluster consists of objects of the same class, and is calculated using the formula:

$$\mathbf{h} = \mathbf{1} - \frac{\mathbf{H}(\mathbf{C}|\mathbf{K})}{\mathbf{H}(\mathbf{C})}$$

Academicia Globe:
Inderscience Research

ISSN: 2776-1010 Volume 2, Issue 6, June, 2021

where K is the result of clustering, C is the true partition, H is the entropy function. Wherein

$$\mathbf{H(X) = -\Sigma_{i=1}^{n} P(x_i) \log_b P(x_i)}$$

$$\mathbf{H(X|Y) = \sum_{i,j} p(x_i y_j) \log \frac{p(y_i)}{p(x_i y_j)}}$$

Completeness measures the extent to which objects of the same class belong to the same cluster and is determined by the formula

$$\mathbf{c = 1 - \frac{H(K|C)}{H(K)}}$$

These measures take values in the range [0,1]. They can be interpreted as follows: the larger the value, the more accurate the clustering. These measures are not symmetric and are not normalized, in contrast to those considered earlier, and therefore they depend on the number of clusters. According to the resource [2]: “Random clustering will not give zero rates for a large number of classes and a small number of objects. In these cases, ARI is preferred. However, if the number of objects is more than 1000 and the number of clusters is less than 10, this problem is not so pronounced and can be ignored. " To take into account the values of both quantities, a symmetric V-measure is introduced, showing how similar the clustering is to each other. Its calculation takes place according to the formula

$$\mathbf{\upsilon = 2 \frac{hc}{h + c}}$$

If we formed the correct criteria and had a sufficient number of learning acts (elements of the reference sample), then after the end of the "training" we will receive a structure that will with a high degree of probability refer an arbitrary document containing certain keywords and metadata to one or another categories. In the limit, the accuracy will be absolutely the same as if it was done by an expert who formed a reference sample “manually”, and even more: a machine learning-based system is free from flaws and does not allow random errors. A rich variety of tools allows today to form different versions of artificial intelligence systems using neural networks and machine learning for applied problems.
In practice, the most widespread are two AI tasks - classification and regression, as well as all kinds of their derivatives. The task of classification is reduced to assigning an object to one or another class from the final list (the example that we considered above). The regression problem differs in that as a result of the algorithm operation, an object is assigned one or more numerical parameters: for example, you can build an algorithm for predicting the preparation time of a response to an incoming document, depending on its content and other parameters, based on the accumulated base of precedents.

Academicia Globe:
Inderscience Research

ISSN: 2776-1010 Volume 2, Issue 6, June, 2021

So, the options for using AI technologies in the field of workflow are endless, and it is obvious that as technologies improve, more and more new ones will appear. Today, AI is most often used to solve three types of tasks - intelligent document search, automatic classification and automatic extraction of attributes (metadata) from the text of documents. Modern EDMS, as a rule, offer ready-made or customizable solutions for these tasks.

Automatic generation of document metadata allows you to automate a wide variety of document processing scenarios: for example, automatic registration of documents in the system, autorun of certain processing processes, assignment of those responsible for the processing of processes, assignment of due dates, etc. There are also less common applications of AI technologies in the EDMS, focused on specific processing procedures - and then we will consider various examples of the use of AI technologies that seem interesting and promising to us, and some of them are already being implemented in the framework of pilot projects.

CONCLUSION

Obviously, the need to use artificial intelligence technologies in the above examples is a consequence of the lack of structure in the processed content. Given a sufficient amount of metadata describing various attributes of a document, these same tasks could be solved based on strict formal processing rules, but the elimination of unstructured content from an organization's business processes is a distant future.

Artificial intelligence technologies open up special prospects in the field of corporate process automation and optimization. Here are some examples of AI's capabilities in process management:

- Automatic assignment of the duration of manual processing steps in business processes. The process control system can itself predict the optimal timing of certain stages based on the accumulated information about their labor intensity.
- Selection of an approval route for a document based on its content, taking into account the workload of personnel and the competence of employees.
- Planning the time of completion of the process and determining its planned metrics based on the accumulated information about the precedents.
- Predicting violations of planned deadlines for processes and individual tasks, optimizing processes in the course of their execution - changing the priority of unfinished tasks, deadlines, automatic delegation of tasks taking into account the workload of employees and their competence, etc.
- Completion of assignments in the event of a critical violation of deadlines, generation of approval results and assignment reports based on precedents.
- Revealing hidden regulations, typical scenarios of document processing based on the accumulated history of free routing. The system can analyze typical processing methods and generate process templates.
- And this is not a complete list of possible applications of AI technologies in the field of process management.

In addition to these general cases, various specific applications of the described technologies can be found in each subject area. Here are some notable implemented examples:

ISSN: 2776-1010 Volume 2, Issue 6, June, 2021

- automatic standard control (checking the compliance of design and technological documents with the formal requirements of the quality management system) in design and engineering organizations;
- search for judicial precedents in the systems of management of claims and claims work;
- search for typical responses to requests in Service Desk services and contact centers;
- automatic audit of compliance with regulations for the use of documents and search for traces of possible malicious actions in security management systems;and much more.

In conclusion, it is worth noting that companies that actively use electronic document management systems and plan their development should pay attention to situations when even working with electronic documents becomes time consuming and leads to repetitive routine actions. Most likely, a solution is possible that will take the process to a new level through the integration of AI technologies - this is carried out within the framework of project development. Obviously, we are at the very beginning of using artificial intelligence in the field of document flow, but individual projects and ready-made solutions already today demonstrate the practice and prospects of using these technologies.

REFERENCES

1) Nemchinova EA, Plotnikova NP, Fedosin SA Preparation and processing of reference text information for classification using artificial neural networks // Nonlinear World. 2019.Vol. 17.No. 2.S. 27-33.
2) Solomentsev Ya. K., Chochia PA Application of neural networks for diagnostics of the type and parameters of image distortions // Information processes. 2020.Vol. 20.No. 2.S. 95-103.
3) Vinokurov AV Parametric method of processing video information based on the use of neural networks as a mechanism for adapting the size of images to the bandwidth of the communication channel // Industrial ACS and controllers. 2017. No. 6.S. 36-39.
4) Kislitsyn EV, Panova MV, Zhernakov RS Principles of application of neural network technologies in the analysis of big data // Prospects of science. 2017. No. 9, pp. 7-10.
5) Vitenburg EA The architecture of the software complex for intelligent decision support in the design of the security system of the enterprise information system. Bulletin of Cybernetics. 2019. No. 4.S. 46-51.
6) Gainullin RN, Rakhal Ya., Rizaev IS, Sharnin LM. Forecasting of business processes based on neural networks // Bulletin of Kazan Technological University. 2017.Vol. 20. No. 3.S. 121-124.
7) Danilov AD, Mugatina VM Solving the problem of optimization of regression testing using a neural network approach // Modeling, optimization and information technologies. 2020.Vol. 8. No. 1.S. 35-36.
8) Obukhov A.D., Krasnyansky M.N. Neural network architecture of information systems // Bulletin of the Udmurt University. Mathematics. Mechanics. Computer science. 2019.Vol. 29. Iss. 3.S. 438-455.

ISSN: 2776-1010 Volume 2, Issue 6, June, 2021

9) Obukhov A., Krasnyanskiy M., Nikolyukin M. Algorithm of adaptation of electronic document management system based on machine learning technology // Progress in Artificial Intelligence. 2020. P. 1-17.
10) Krasnyanskiy M., Ostroukh A., Karpushkin S., Obukhov A. Formulation of the Problem of Structural and Parametric Synthesis of Electronic Document Management System of Research and Education Institution // Global Journal of Pure and Applied Mathematics. 2016.Vol. 12. No. 3.P. 2395-2409.
11) Stefanova NA, Kurbangeldyev D. Estimation of the cost of software development // Actual problems of modern economy. 2020. No. 1.S. 67-72.33
12) Odilovich, O. A., Umirzokovich, T. F., & Turdibaevich, K. R. (2021). Increasing the Efficiency of Higher Education Personnel Training Management in Uzbekistan. Annals of the Romanian Society for Cell Biology, 9251-9264.