UDC 004.9

DOI: 10.15587/1729-4061.2022.251689

DEVELOPMENT OF INTELLIGENT ELECTRONIC DOCUMENT MANAGEMENT SYSTEM MODEL BASED ON MACHINE LEARNING METHODS

Madina Sambetbayeva
Corresponding author
PhD, Associate Professor*
Senior Researcher**
E-mail: madina_jgtu@mail.ru

Inkarzhan Kuspanova
PhD Student*

Aigerim Yerimbetova
PhD, Associate Professor, Leading Researcher**
Department of Software Engineering
Institute of Automation and Information Technologies
Satbayev University
Satbayev str., 22 a, Almaty, Republic of Kazakhstan, 050013

Sandugash Serikbayeva
Teacher*

Shynar Bauyrzhanova
PhD Student*

*Department of Information Systems
L. N. Gumilyov Eurasian National University
Satpayev str., 2, Nur-Sultan, Republic of Kazakhstan, 010008
**Institute of Information and Computational Technologies
Shevchenko str., 28, Almaty, Republic of Kazakhstan, 050010

With the daily increase in document flow, as well as the transition to paperless document management around the world, the demand for electronic document management systems is increasing. This significantly requires optimization of these systems in terms of quality document information retrieval and document management. However, research based on statistical methods cannot effectively handle large amounts of data extracted from electronic documents. In this regard, machine learning methods can effectively solve this problem. This paper presents an approach to building a model of an intelligent document management system using machine learning techniques to ensure efficient employee performance in organizations. The authors have solved a number of problems to optimize each of the document management subsystems, resulting in the development of an intelligent document management system model, which can be effectively applied to enterprises, government and corporate institutions. The feasibility and effectiveness of the proposed model of intelligent document management system based on machine learning and multi-agent modeling of information retrieval processes provides maximum reliability and reduced time of work on documents. The obtained results show that with the help of the presented model it is possible to further develop an intelligent document management system that will allow an electronic document to qualitatively go through the whole life cycle of a document, starting from the moment of document registration and finishing with its closing, i.e. execution, which will greatly facilitate the daily work of users with large volumes of documents. At the same time, the paper considers the application of topic modeling methods and algorithms of text analysis based on a multi-agent approach, which can be used to build an intelligent document management system

Keywords: electronic document management system, machine learning, multi-agent technologies, topic modeling

Received date 19.11.2021
Accepted date 17.01.2022
Published date 25.02.2022

***How to Cite**: Sambetbayeva, M., Kuspanova, I., Yerimbetova, A., Serikbayeva, S., Bauyrzhanova, S. (2022). Development of intelligent electronic document management system model based on machine learning methods. Eastern-European Journal of Enterprise Technologies, 1 (2 (115)), 68–76. doi: https://doi.org/10.15587/1729-4061.2022.251689*

1. Introduction

Electronic document management systems are used in various fields. In particular, electronic document management systems play a key role in the structuring of paperwork processes in government agencies, bringing them into a single order, as well as optimizing the work of civil servants by providing effective and seamless access to documents with the function of automating routine operations to track and search for necessary information and the formation of reports on the document flow.

However, every year a very large volume of documents with a regulated time of their processing is processed in public authorities, and the quality and efficiency of document interactions largely determine the efficiency and effectiveness of public authorities. As e-government develops, the number of requests processed can reach several thousand per day. At the same time, processes in government structures are typified, as are documents, so the application of intelligent algorithms will be more effective than in a structure with a complex and unique organizational structure. Machine learning can speed up document processing, prepare all the data necessary for human decision-making and also prevent human error.

The history of document management goes back to the end of the nineteenth century with the invention of the filing cabinet. In 1898, Edwin Granville Seibels developed a vertical filing system in which paper documents are organized in boxes placed in folded cabinets. These cabinets would remain the primary method of document storage in the business world for most of the twentieth century [1].

The history of document management changed dramatically in the 1980s with the increasing availability of com-

Copyright © 2022, Authors. This is an open access article under the Creative Commons CC BY license

puter technology. The development of servers enabled organizations to store documents electronically in centralized mainframes. This was the beginning of electronic document management systems. Meanwhile, the invention of scanners made it possible to convert paper documents into digital documents. The growth of computers has enabled businesses to create and store documents on computers in the office [2].

Modern electronic document management systems make it possible to store large volumes of digital documents centrally. In order to ensure good classification of electronic documents, many electronic document management systems rely on a detailed document storage process including certain elements called metadata. A large number of modern corporations use original storage protocols in their electronic document management systems to enhance information security, which is what makes an electronic document management system so valuable to a business or organization [2].

Every activity is reflected in documents, be it management, finance or production, so document management is a vital system for an organization. Quite often, the automation of document management alone can significantly improve the business processes of an entire organization.

The massive increase in the volume of electronic document flow due to the pandemic has led to an increase in the mechanical single-type work of the employees, managers and document services staff of organizations, who register and respond to a thousand or more documents a day. This has resulted in increased labor and time costs. The relevance of this problem in modern times is to improve the traditional systems of electronic document management, by applying methods of data analysis and machine learning, to optimize the work of employees of the organization and quality passage of the entire life cycle of the electronic document with minimal human intervention in the process.

2. Literature review and problem statement

The work [3] presents the results of research of machine learning techniques to discuss the current challenges in managing scientific workflows in distributed systems. It is shown that there are some potential issues with using machine learning, such as collecting the training data. They described the workflow-level analysis, task-level analysis, infrastructure-level analysis, cross-level analysis, online/offline analysis and training data collection. They believe that new workflow systems will be able to understand the user's previous requests, discover the related data and structure the computations needed to deliver the desired results. But questions related to using machine learning techniques in the scientific workflow space remain under-researched. The reason for this may be the difficulty in analyzing the processes used to produce the results, as well as the difficulty in reproducibility. Nevertheless, it can potentially provide a means of comparison of different scientific methods and their similarities and differences to other approaches. An option to overcome the corresponding difficulties can be the use of machine learning methods. It is this approach that was used in the work [4]. The authors used the concept of support subsystem for decision making based on the application of interactions of the type "User - EDMS - Document"; and technologies of machine learning, which will be used to automate the process of document processing by the example of EDMS design documentation, are formulated. The obtained scientific results will be used for the problem solution of information processing automation in different information systems. However, the work [5] suggested developing an adaptation algorithm using machine learning methods to solve the problem of structural-parametric synthesis of EDMS. It is shown that the main scientific results obtained in the paper include: formalized criteria for adapting EDMS; the algorithm for designing and adapting EDMS; and development of software for adapting EDMS, including a trained neural network and API.

The work [6] presents the results of research of using automated data input from scanned copies of documents of the contract department compared to manual input. To improve the transfer of data from a scanned copy of a document to a corporate database management system in the energy industry, it is proposed to use machine learning, presented in the form of a neural network. It is shown that machine learning allows data classification for analyzed documents, which ensures the selection of the correct template when generating an electronic document. The choice of tools for developing a software module for data extraction is substantiated and the principle of its operation is described. But questions related to multiple text classification machine learning models are not reflected. The reason for this may be the difficulty in terms of using classification methods on a large volume of documents. The approach of using classification methods was applied in the work [7].

The work [7] used Optical Character Recognition (OCR) to create and evaluate multiple text classification machine learning models, including both "bag of words" and deep learning approaches. They evaluated the system on three different levels of classification using both the entire document as input, as well as the individual pages of the document. Also, they compared the effects of different text processing methods. This model distinguished between clinically-relevant documents and not clinically-relevant documents with an accuracy of 0.973; between intermediate sub-classifications with an accuracy of 0.949; and between individual classes with an accuracy of 0.913. However, the paper only provides a comparison of classification methods, but does not provide a semantic map for subsequent document categorization. The reason for this may be difficulties associated with providing a semantic map for document classification. Moreover, this approach was used in the work [7].

The work [8] describes the results of research of a proprietary approach based on the use of a semantic map as a feature reduction tool for document classification. They investigated the impact of this approach on the quality of document classification and describe its application to the implementation of document categorization. But questions related to using agent technologies are not considered. The reason for this may be difficulties associated with providing the agent technologies for documents, which makes the relevant research impractical. The approach of developing the new architecture by using agents was used in the work [9].

The work [9] presents the results of research of a new concept of knowledge classification integrated on the cognitive agent architecture, so as to speed up its inference process. They described the new architecture and the agent will be able to select only the actionable rule class, instead of trying to infer its whole rule base exhaustively. But questions related to using multi-agent technologies and topic modeling are not considered. The reason for this may be objective difficulties associated with the costly part of developing the architecture with multi-agent technologies for documents. An option to overcome the corresponding difficulties can be to develop a model of document management systems using machine learning, topic modeling and multi-agent technologies.

3. The aim and objectives of the study

The aim of the study is to build a model of an intelligent electronic document management system using artificial intelligence algorithms, which will enable users to optimize the passage of the entire life cycle of the document with minimal human intervention in the process.

The created technologies and tools will provide modeling of information extraction processes and creation of automatic text processing systems based on the ontology of the subject area and linguistic knowledge represented by subject dictionaries and fact models.

The objective of the study is to develop a model, methods and effective algorithms that provide information extraction from documents, compiling intelligent text analysis and optimizing the entire business process of document management.

In order to achieve the objective, the following tasks had to be accomplished:

– to perform a comparative analysis of existing electronic document management systems on the information systems market;

– to optimize each of the electronic document managegment system's subsystems using machine learning methods;

– to construct a model of intelligent document manaagement system to ensure the efficient work of the staff of organizations.

4. Materials and methods

Machine learning is a field of computer science, which aims at training computers to learn and act without explicit programming. Specifically, machine learning is an approach to data analysis that involves building and adapting models that allow programs to "learn" from experience. Machine learning involves building algorithms that adapt models to improve their ability to form predictions or further actions on similar procedures that have been embedded in the model [10].

Electronic Document Management Systems (hereinafter referred to as EDMS) are automated information and reference systems designed to automate the technological processes of preparation, registration, centralized structuring, storage, archiving, search and processing of documents, control of execution, authorization of access to them, issuing and distribution of documents, extracting information from documents and its analysis, obtaining knowledge from the accumulated information, decision-making support [11].

The EDMS are designed to improve the efficiency and transparency of workflow processes and perform the following main functions:

– receipt, processing, registration and storage of incoming and outgoing correspondence, including internal correspondence;

– receipt, processing, registration, storage and redirecetion of references of individuals and legal entities;

– receipt, processing, registration and storage of organinzational and administrative documents;

– the process of giving instructions (resolutions) and exercising control over the execution of documents;

– the process of reading, creating, approving, signing, finalizing, executing, closing and storing documents;

– the process of electronic document exchange;

– the process of formation and storage of reports.

However, every year a very large volume of documents with a regulated processing time is processed in public authorities, and the quality and efficiency of document interactions largely determine the efficiency and effectiveness of public authorities. The functions of EDMS in governmental bodies are not limited to internal management. A large volume is occupied by external document management – communication with citizens and organizations on the provision of public services. Due to the development of e-government, the number of processed requests can reach several thousand per day. At the same time, processes in government structures are typified, as are documents, so the application of intelligent algorithms will be more effective than in a structure with a complex and unique organizational structure.

Machine learning can accelerate the processing of documents, prepare all the data necessary for human decision-making and prevent human error [12].

Through the use of machine learning algorithms, a document can go all the way from registration to the formation of the document itself with minimal human intervention in the process.

The existing EDMS are automated information and reference systems designed to automate the following technological processes, both intra-agency and interdepartmental:

– processing of electronic documents (hereinafter referred to as EDs); ensuring control mechanisms of ED execution;

– authorizing access to EDs;

– providing access to published documents.

The system processes documents created and used in the records management of the state bodies.

To ensure these objectives, EDMS provides the following functions: preparation, registration, storage, archiving and retrieval of EDs at the departmental level, and routing at the interagency level; provision of a transport environment for interagency exchange of documents [13].

The EDMS consists of the following subsystems:

– a subsystem for processing internal and external correspondence;

– a subsystem for the preparation and approval of draft documents;

– a subsystem for supporting regulatory and reference information (NRIS);

– a subsystem for the preparation of reporting information;

– a subsystem of interaction with temporary archive base;

– a subsystem for the conversion of documents in hard copy to an electronic format;

– a subsystem of the system administration;

– a subsystem ensuring the interaction with the EDMS center.

The functional structure of the EDMS subsystems is shown in Fig. 1.

This functional structure is suitable for traditional electronic document management systems and requires significant modification and optimization to build a new model of an intelligent document management system.

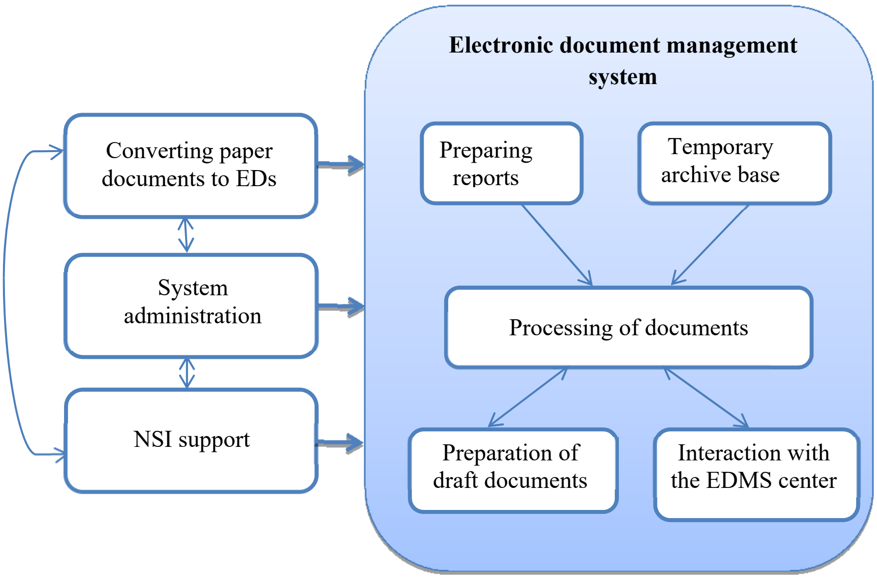
Electronic document management system
Converting paper documents to EDs
System administration
NSI support
Preparing reports
Temporary archive base
Processing of documents
Preparation of draft documents
Interaction with the EDMS center

Fig. 1. Functional structure of electronic document management system subsystems

5. Approaches to building an intelligent document management system model

5. 1. Comparative analysis of existing electronic document management systems on the information systems market

On the market of information systems, a number of electronic document management systems are represented, but they can provide only typical automation of paperwork. Description of document management systems implemented in the market of the country is presented in Table 1.

It follows from the above table that the existing systems in the country provide full automation of paperwork. But with the daily increase in the volume of documents with regulated time of their processing, the quality and efficiency of document interactions largely depend on the promptness and efficiency of organizations. The functions of electronic document management systems are not limited to internal administration, the large volume is occupied by external document management communication with citizens and organizations on the provision of services. At the same time, the application of intelligent algorithms will be more effective in increasing the productivity of each user and the organization as a whole.

Table 1

Comparative analysis of existing electronic document management systems

System	Description
1	2
InDocs	A solution for document management and automation of document-oriented business processes. Designed for government and commercial organizations (enterprises) of various sizes. InDocs allows automating work with various types of documents, both within the local network, and in territorially distributed structures with the complex scheme of information flows
AlmexECM	The platform is implemented in Java. Combined with OOP and OOD (object-oriented programming and design), TDD (test-driven development) and a number of powerful frameworks and libraries (Spring, ORM Hibernate), the AlmexECM platform is a powerful and stable workflow automation tool. The platform uses Apache Thrift technology to write integration code in any modern programming language
Thesis Softline	SoftLine offers solutions based on leading ECM platforms, ready for rapid implementation, for retail, banks, insurance organizations, industrial and construction companies, the oil and gas sector and government agencies
Synergy workflow	Software for organizing a holistic and systematic approach to managing the organization's internal processes, namely: Control of orders and protocols; Organizational and administrative document management; Personnel records and workflow; Performance discipline report
DIRECTUM QAZ	Specially created for the Kazakhstan market, the Directum QAZ system covers the whole range of tasks on automation of internal electronic document circulation and specific business processes in companies of various industries. The system with built-in intelligent mechanisms supports ES from the National Certification Authority of the Republic of Kazakhstan (NCA RK), is localized into the Kazakh language and registered by the Ministry of Justice of the Republic of Kazakhstan (certificate of state registration No. 157 dated 13.08.2018)
E1 Evfrat	The E1 Evfrat system is a powerful tool for the automation of business processes and document flow for companies of all types and sizes. The system efficiently solves tasks both within a small structure, such as an office, a department, a unit or a local organization as a whole, and within a geographically distributed organization with a complex scheme of information flows. The system has a leading position in the DMS, BPM and EMC system classes

Continuation of Table 1

1	2
EDMS Docsvision	Docsvision is a full-featured EDMS/ECM platform that enables a wide variety of solutions for automating business processes and document processing tasks
Documentolog	Documentolog provides the full lifecycle of all electronic documentation within the current business processes of the company. Automation of business processes of any organization in accordance with the internal regulations of the organization and approved regulations of the Republic of Kazakhstan
TENGRIDOC®	The system allows you to manage the creation, execution and approval of documents. The system is adaptable to any type of company, supports the exchange of documents with geographically distributed offices and automates both the work of individual departments (e.g. clerical department) and the company as a whole
Euredoc	EDMS is designed to organize paperless workflow and corporate document management technology, which ensures the movement of documents at the customer from the time they are created or received, through completion and execution or sending, as well as providing information support for organizational and administrative activities
Integro	Automation of the entire document "life cycle" (from project creation to document write-off and archiving)
RealSoft	Based on a platform IBM Lotus Notes/Domino IBM Messaging and Collaboration Solutions (Collaboration Solutions Messaging and Collaboration), Collaboration Solution Portal (Collab Sol Portal)

5. 2. Optimizing each of the electronic document management subsystems using machine learning techniques

In order to implement intelligent workflow, the above subsystems need to be improved using machine learning.

The internal and external correspondence processing subsystem performs the following basic operations:

1. Document registration.

In this operation, the primary details of the document are defined and the subsequent course of its processing is determined, i.e. to whom the document should be transmitted for consideration. The document can be passed not only directly to the next stage of document processing, but also in case the possible executor of this document is already known, directly for execution.

2. Definition of resolutions for the document.

In the course of this operation, the document is determined in accordance with the hierarchical structure of the organization – the document goes from the chief executive to the final executor.

3. Putting a document under control.

This operation can be performed after the document is registered in the EDMS or at any moment of the document processing. At the moment of this operation, the responsible executor is determined and the deadline for the document execution is set.

4. Execution of instructions for this document.

According to resolutions, the execution of this document is going on. If the execution requires the creation of a response document, it is carried out in the form of preparation of a draft document and its approval.

5. Withdrawal from control.

Withdrawal of control is carried out according to the results of the document execution, if this document requires confirmation.

The following tasks are supposed to be implemented in this subsystem:

– cluster data analysis. Before applying machine learning algorithms to documents, given the large volume of data, it is necessary to partition them into clusters. This will solve problems such as finding duplicates, searching for related/similar documents, etc., and will also allow building an algorithm for more accurate prediction of document attributes;

– prediction of document attributes. Any electronic document is accompanied by a set of attributes (author, subdivision, document type, executor, etc.) to be filled in for its further processing, as well as for further document search and report generation. As a matter of fact, the process of document processing depends entirely on the set of attributes: for example, documents from a specific addressee and on a specific topic (the categories mentioned above) must be processed by a specific unit and according to quite specific rules. At present, the processing of each document is 100 % manual. But, given the structured nature of this information, the same rules can easily be taught to a machine learning algorithm. By "ingesting" a good database in which documents are structured according to organization rules, machine learning algorithms will be ready to independently predict new attributes and processing routes for new documents, as well as predict the number of days it takes to complete a task and identify the performer. In order for algorithms to learn how to do this with high accuracy, a database of structured and not-so-structured data of huge volumes is needed [14];

– autofill at registration. Based on the contents of the text, the system should automatically fill in the necessary data in the document card, determine the relation with other similar documents or correspondence, and suggest the addressee of the message itself, based on the data on the fulfillment of similar matters. And according to the same principle, the document processing times are to be determined;

– automatic abstracting. Manual abstracting (creating a brief, meaningful "excerpt" from the full document) is a difficult, time-consuming job, so it is also advisable to use automatic abstract generation tools. The first publications on the subject of automatic text abstracting methods date back to 1958. Since then a large number of methods have been developed and the quality of the results has improved. The main tasks of automatic abstracting in an EDMS are to highlight the main information in a document and to avoid duplication.

In the subsystem of preparation and approval of draft documents, the following main operations are carried out: to automate the processes of preparation and agreement of documents, the conventional type of document – "project document", characterized by its own card and having its own unique registration number, is introduced.

This subsystem provides linkage to the following subsystems:

1. Internal and external correspondence processing subsystem.

The processes of drafting and approving documents are mainly considered as one of the operations in the course of document processing. Interaction between subsystems is performed by defining an assignment in the subsystem of internal and external correspondence, processing an assignment, which becomes the basis for document creation.

A document created without an explicit assignment definition (initiative) is sent to the processing subsystem for registration after the approval and signature operations.

2. The subsystem of work with NSI, for request of elements of directories, which are used for filling of requisites of documents.

Reference books are used for the determination of values of details of draft documents, therefore during the processing of documents, their registration, enquiry of data of reference books from the NSI subsystem is carried out [15].

In this subsystem, the following tasks are supposed to be implemented:

– intelligent routing. The system in a few seconds.

The system should automatically determine a project approval route on the basis of its content, find related documents and draw up a document resolution. As a result, the approval process is accelerated and errors in document production are reduced. It is also possible to see which steps or employees slow down the work. At the same time, the intelligent routing can take the workload of employees into account – if the reviewer's task limit is exceeded, a colleague gets involved in the approval process;

– preparing a response template. Often the same type of requests are received.

Very often, the same kind of inquiries are received. In the preparation of the answering document project to these inquiries, the system must automatically search similar documents and prepare the typical template of the answering document;

– intelligent search in the system.

The system should allow setting up intelligent search of the documents.

In the subsystem of support of the normative-reference information, the following basic operations are carried out:

This subsystem provides access to the system's NSI. The NSI used in the system is classified into:

1. NSI used in interdepartmental documents exchange.

For this class of directories, centralized maintaining and storing are provided.

2. The directories of a given public authority.

These directories are not used in interdepartmental exchange and are used to determine the details of documents of the state authority itself.

This subsystem provides other subsystems included in the EDMS with the necessary data.

The following tasks are to be implemented in this subsystem: smart directory. Implementation of smart guide, semantic search for words [16].

In the subsystem for the preparation of reporting information, the following basic operations are carried out.

This subsystem is designed for getting statistical information about documents that are processed in the system.

In addition to providing standard statistical reporting forms, there is the opportunity to create templates of reporting forms that reflect the specifics of the respective public authority, using a high-level report builder.

The subsystem for managing report templates and generating report forms provides the following functions:

1. Receipt of report forms using available report templates.

2. Creation of report templates and determination of a schedule for generating report forms available to users.

This subsystem interacts with the subsystem of documents processing and NSI to obtain necessary information to provide reporting information.

In this subsystem, it is expected to implement the following tasks: generation of intelligent reports. Intelligent report generation for overdue documents in real time, report generation for each executor with detailed workload of each executor, report generation for control documents.

The subsystem of interaction with the departmental archive of EDMS performs the following main operations:

– this function ensures the transfer of data (documents finalized in the previous year's records management) to the EDMS departmental archive;

– this subsystem interacts with the subsystem of documents processing and NSI to obtain the necessary information and transfer documents to the temporary archival repository.

The subsystem for converting paper documents into electronic records.

This subsystem ensures the automation of conversion of paper documents to electronic format. The scanning subsystem allows storing electronic documents for different databases and browsing the list of scanned documents.

The following tasks are to be implemented in this subsystem: text recognition in scanned documents, application of the machine learning method to scanned texts.

The system administration subsystem performs the following basic operations:

– the functions of the System Administrator include managing the settings of the relevant subsystems and access to the functions provided by the system, as well as viewing the audit logs of the users' work. In addition, the system provides auditing of operations performed by the System Administrator himself, except for the determination of access rights to system objects;

– system administration functions fall outside of this subsystem: backup, system installation, system parameter settings, etc.

The administration subsystem interacts with the other systems included in the EDMS to determine the settings of these subsystems, ensure security and audit operations performed by the users of the subsystems.

In order to implement the above tasks, it is proposed to use the actively developing and currently relevant area – topic modeling.

Topic modeling is a method of constructing a model of a collection of text documents, which determines what topics are related to each of the documents. The Topic Model of a collection of text documents identifies which topics are covered by each document and which words (terms) constitute each topic. The currently most popular topic modeling methods can be divided into two main groups: algebraic and probabilistic (generative).

The algebraic models include the standard Vector Space Model (VSM), Latent Semantic Analysis (LSA) and among the probabilistic ones, the most popular are probabilistic LSA (pLSA) and Latent Dirichlet Allocation (LDA), as well as additive regularization of topic models (ARTM) based on these algorithms [17].

During the study, it was decided to use Latent Dirichlet Allocation as this method avoids the drawbacks of pLSA, such as "over-learning" and lack of pattern in generating documents from the set obtained by topics, which significantly improves the final sample.

Latent Dirichlet Allocation is a generating model used in machine learning and information retrieval that allows explaining the results of observations using implicit groups, so that it is possible to identify the reasons for the similarity of some parts of the data. For example, if the observations are words collected in documents, it is argued that each document is a mixture of a small number of topics and that the occurrence of each word is related to one of the topics in the document [17].

5. 3. Building a model of an intelligent document management system to ensure efficient work of employees in organizations

The model of an intelligent document management system is developed on the basis of new methods and algorithms of text analysis based on machine learning, as well as multi-agent approach [14]. The novelty of machine learning methods is the simplified setup and development of information retrieval systems and facilitated switching of the whole document management system to a new subject area. And multi-agent text analysis and information extraction consist of using two types of agents. Lexical agents correspond to the objects of the subject area found in the text, and cognitive-linguistic agents detail these objects and establish connections between them. Cognitive-linguistic agents are associated with various cognitive operations associated with complex linguistic structures and different types of linguistic information processing. All agents act in parallel and independently. The agents extract information from the text in the form of ontology-based structures (facts, objects, relations). The result of their actions will be the information represented by the network of agents, where each agent forms an object or an instance of a relation corresponding to some class of ontology [18].

The proposed approach is in line with the current trends of modern research devoted to the automatic processing and analysis of large volumes of heterogeneous data presented in natural language. Most traditional systems of text analysis are organized on the basis of purely sequential architecture, which is easy to use and manage, but the efficiency of such systems is very low. The sequential architecture implies the sequential analysis of text at different linguistic levels: graphematic, morphological, syntactic and semantic.

Recently, semantically-oriented approaches to text analysis have been developed and systems with non-sequential architecture have been created [19], when the analysis is performed in parallel at all levels on the basis of various semantic resources such as thesauri and ontologies. Using the multi-agent approach makes it possible to create good alternatives to text analysis systems with sequential architecture. The peculiarity of the approach is the representation of the developed system with the help of autonomous entities – agents, which have the ability to interact with the environment and other agents. In the process of this interaction, the functioning of the system takes place. Traditionally, the advantage of the multi-agent approach is the paralleling of the system functioning process, due to the independence of the agents and their ability to interact with each other, as a result of which the tasks of the system are solved locally and therefore significantly accelerate the result [20].

Having analyzed all the existing EDMSs in the country and the unresolved document management problems and machine learning algorithms, a model of intelligent EDMS (Fig. 2) is proposed to further implement a self-learning, self-developing and self-regulating document management system.

The implementation of the above model is expected to improve the document routing mechanism, statistical processing of routing schemes (approval statistics, delegation, bottlenecks, time delays), reconfiguration of the document route, development of document operations system, formation of knowledge base, work with large document flows, recognition of documents, etc.

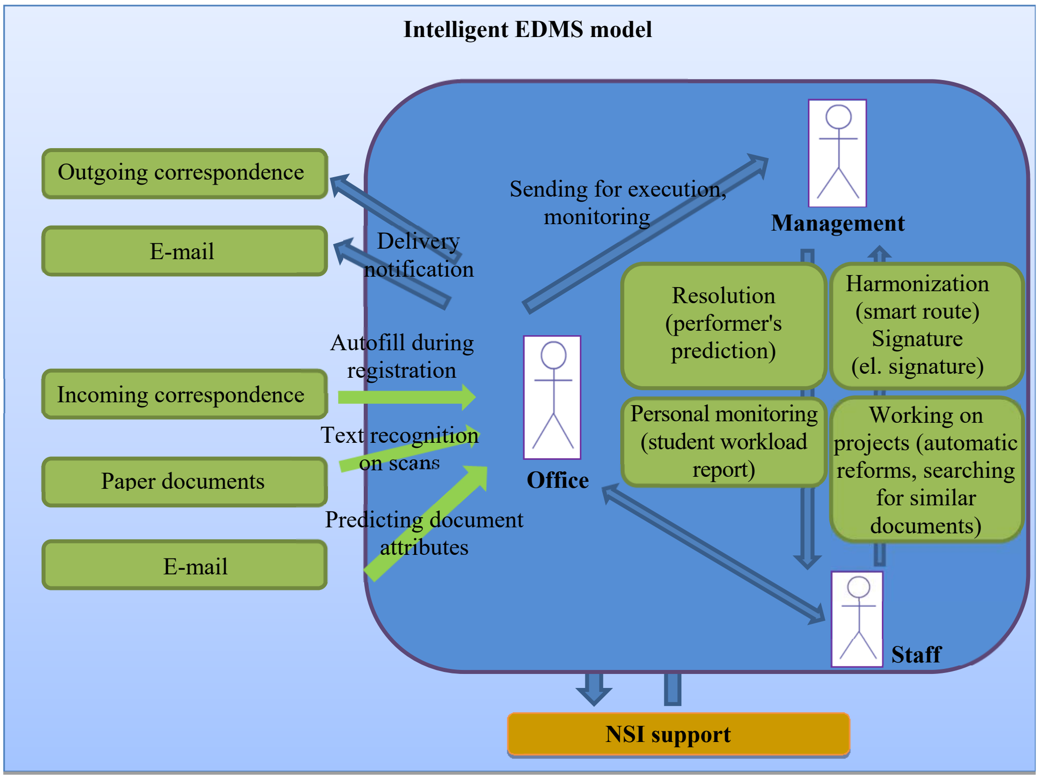
Intelligent EDMS model
Outgoing correspondence
E-mail
Incoming correspondence
Paper documents
E-mail
Delivery notification
Sending for execution, monitoring
Management
Autofill during registration
Text recognition on scans
Predicting document attributes
Office
Resolution (performer's prediction)
Harmonization (smart route) Signature (el. signature)
Personal monitoring (student workload report)
Working on projects (automatic reforms, searching for similar documents)
Staff
NSI support

Fig. 2. Intelligent model of electronic document management systems

6. Discussion of experimental results of studying intelligent electronic document management system model based on machine learning methods

As a result of a comparative analysis of existing traditional electronic document management systems, it was revealed that none of them can be presented as a smart document management system. That accordingly requires the improvement and enhancement of traditional electronic document management systems in terms of using advanced training methods. In this regard, a model of an intelligent electronic document management system has been developed, which is explained by the introduction of new modules to improve the business processes of electronic document management (Fig. 2). Thus, the methods for business process optimization of EDI systems, such as search of duplicates, search of related/similar documents, and algorithm for more accurate prediction of document attributes, automatic registration of documents, intelligent routing, preparation of response letter templates and generation of intelligent reports are presented.

As for text analysis, based on the research performed, the task of resolving ambiguity in text analysis and extracting information from it has been formulated. The result of text analysis is a set of found facts represented in the form of text coverage of information objects of a given subject area. The ambiguity of the text at this level is manifested in the presence of conflicting relations between objects, where each conflict, in fact, generates a separate version of the object coverage of the text. In contrast to the work [7], where there are only three different levels of classification, which use the whole document as input data, and individual pages of the document, our proposed problem of conflict resolution is to resolve all ambiguities so that the system is conflict-free and yet retains as many objects and relations as possible. To solve this problem, a number of algorithms will be used to resolve text ambiguities as part of a conflict resolution system for a multi-agent text analysis system.

A verification method is proposed for families of distributed and multi-agent systems generated by a context-dependent network grammar of a special kind. This verification method can be applied to verify the properties of multi-agent conflict resolution systems. The set of ontology instances found in the process of text analysis can be represented in the form of Scott information system with an inference relation in the form of a set of information relations. The obtained Scott information system generates a multi-agent system whose agents resolve lexical and semantic ambiguities and is the correct algorithm for ambiguity resolution.

On the basis of the research, it is proposed to develop a technological environment for the creation of subject-oriented systems of information extraction from texts on the basis of a specialized multi-agent platform. And also the implementation of the design of technological environment, which includes three main components: a dictionary subsystem, a module of genre typing and a module of multi-agent text analysis.

The originality of the approach proposed in the study, in contrast to the work [9], where the architecture is built where the agent can select only the class of valid rules, instead of trying to exhaustively derive the entire rule base, is that the information extraction will be performed using machine learning algorithms, and multi-agent algorithms that implement text analysis from the interaction of agents: information agents, matched to the domain entities, and control agents, calculating the characteristics of the object. Another feature of the developed technology is preliminary modeling of the processes of text analysis, in which the model of information extraction is presented in the form of an attributed graph with the given properties.

The main advantage of this approach is that the created technology significantly reduces the time of work on documents, ensures optimal document flow, and provides transparency (explainability) of the results obtained by the system for the user. Also the development of a vocabulary subsystem, which performs lexical analysis of the text and identifies meaningful terms on the basis of subject-oriented dictionaries. Development of genre typing module, performing genre analysis of the text on the basis of genre patterns. On the basis of the specialized multi-agent platform, you can create a simplified implementation of the algorithms of multi-agent text analysis. It will include an agent initialization system, a conditionality checker, a component for searching for appropriate control models, and implementation of agent interaction protocols.

This paper builds an assumed model of an intelligent electronic document management system, which may lead to difficulties in the development of the system itself and the choice of intelligent analysis methods. However, all the shortcomings of the study can be eliminated with the addition of new research results.

The development of this research may lead to the abandonment of traditional electronic document management systems in favor of intelligent ones. However, one may face experimental difficulties when dealing with a large volume of electronic documents.

7. Conclusions

1. A comparative analysis of existing electronic document management systems revealed that none of them can be positioned as a smart document management system. They have huge shortcomings and require improvement by using methods of data mining and multi-agent systems. The use of these methods will build documents into a single business process, thereby optimizing the function of working with documents.

2. Each subsystem of the electronic document management system is analyzed. Methods for optimizing each of these subsystems are proposed. The peculiarity of the proposed methods is the extraction of information with the help of machine learning algorithms, and multi-agent algorithms that implement text analysis from the perspective of interaction between agents. Another feature of the developed technology is the pre-modeling of the text analysis processes, in which the model of information extraction is represented in the form of an attributed graph with the given properties.

3. The results of the work allow recommending the developed model of an intelligent system of electronic document management for realization of the given optimization in state structures or in organizations, where the system of electronic document management is implemented. The main advantage of this approach is that the technology created on the basis of this model significantly reduces the time of work on documents, ensures the optimal flow of documents, as well as provides transparency of the results obtained by the system for the user.

Acknowledgments

This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No AP08857179).

References

1. Lapshina, S. N. (2012). Architecture of Enterprise. Yekaterinburg: UrFU.
2. Alpaidin, E. (2017). Machine learning: the new artificial intelligence. Moscow: Alpina Publisher, Publishing Group "Tochka", 208. Available at: https://cdn1.ozone.ru/multimedia/1017469342.pdf
3. Deelman, E., Mandal, A., Jiang, M., Sakellariou, R. (2019). The role of machine learning in scientific workflows. The International Journal of High Performance Computing Applications, 33 (6), 1128–1139. doi: https://doi.org/10.1177/1094342019852127
4. Obukhov, A., Krasnyanskiy, M., Nikolyukin, M. (2019). Implementation of Decision Support Subsystem in Electronic Document Systems Using Machine Learning Techniques. 2019 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon). doi: https://doi.org/10.1109/fareastcon.2019.8934879
5. Obukhov, A., Krasnyanskiy, M., Nikolyukin, M. (2020). Algorithm of adaptation of electronic document management system based on machine learning technology. Progress in Artificial Intelligence, 9 (4), 287–303. doi: https://doi.org/10.1007/s13748-020-00214-2
6. Levina, T., Rodionov, A., Farkhutdinov, R. (2020). Software module for extracting data from electronic documents. 2020 International Conference on Electrotechnical Complexes and Systems (ICOECS). doi: https://doi.org/10.1109/icoecs50468.2020.9278492
7. Goodrum, H., Roberts, K., Bernstam, E. V. (2020). Automatic classification of scanned electronic health record documents. International Journal of Medical Informatics, 144, 104302. doi: https://doi.org/10.1016/j.ijmedinf.2020.104302
8. Kostkina, A., Bodunkov, D., Klimov, V. (2018). Document Categorization Based on Usage of Features Reduction with Synonyms Clustering in Weak Semantic Map. Procedia Computer Science, 145, 288–292. doi: https://doi.org/10.1016/j.procs.2018.11.061
9. Chemchem, A., Alin, F., Krajecki, M. (2018). Deep Learning and Data Mining Classification through the Intelligent Agent Reasoning. 2018 6th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW). doi: https://doi.org/10.1109/w-ficloud.2018.00009
10. Holzinger, A., Kieseberg, P., Tjoa, A. M., & Weippl, E. (Eds.) (2018). Machine Learning and Knowledge Extraction. Lecture Notes in Computer Science. Springer, 372. doi: https://doi.org/10.1007/978-3-319-99740-7
11. Edinaya sistema elektronnogo dokumentooborota gosudarstvennyh organov (ESEDO). Available at: https://www.nitec.kz/index.php/post/edinaya-sistema-elektronnogo-dokumentooborota-gosudarstvennyih-organov-esedo
12. Aliev, V. S., Chistov, D. V. (2011). Business planning using the Project Expert program (full course). Moscow: INFRA-M, 432.
13. Eremeev, M., Vorontsov, K. (2019). Lexical quantile-based text complexity measure. Proceedings of Recent Advances in Natural Language Processing. Varna, 270–275. Available at: https://aclanthology.org/R19-1031.pdf
14. Ataeva, O. M. (2016). An information model of LibMeta semantic library. Software & Systems, 4, 36–44. doi: https://doi.org/10.15827/0236-235x.116.036-044
15. Semantic Web. Available at: https://www.w3.org/standards/semanticweb/
16. Weitzel, D., Bockelman, B., Brown, D. A., Couvares, P., Würthwein, F., Hernandez, E. F. (2017). Data Access for LIGO on the OSG. Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact. doi: https://doi.org/10.1145/3093338.3093363
17. Linev, A. A. (2014). Modern EDMS: From Document Management to Efficiency Management. Deloproizvodstvo, 1. Available at: https://www.top-personal.ru/officeworkissue.html?314
18. Challenger, M., Tezel, B., Alaca, O., Tekinerdogan, B., Kardas, G. (2018). Development of Semantic Web-Enabled BDI Multi-Agent Systems Using SEA_ML: An Electronic Bartering Case Study. Applied Sciences, 8 (5), 688. doi: https://doi.org/10.3390/app8050688
19. Jensen, A. B., Villadsen, J. (2020). GOAL-DTU: Development of Distributed Intelligence for the Multi-Agent Programming Contest. Lecture Notes in Computer Science, 79–105. doi: https://doi.org/10.1007/978-3-030-59299-8_4