ALGORITHM FOR DETERMINING KEYWORDS FOR DISTRIBUTION OF DOCUMENTS BASED ON MACHINE LEARNING

[1]Marysheva L.T., [2]Abduvalieva Z. A., [3]Sheina N.E.
[1,2]Tashkent University of Information Technologies,
[3]Tashkent State Technical University

https://doi.org/10.5281/zenodo.8436309

Abstract. *The article presents a method for implementing the information environment of a structural unit for document management (such as a department, laboratories, department of work with students) of an educational organization (University, Institute) based on the integration of several systems with an additional intelligent analysis module for the purpose of electronic document management. Methods for document processing and methods for determining keywords are considered. A document type recognition module has been developed. A text recognition method based on keywords has been implemented to develop a data mining module. This article sets the task of developing an intelligent text recognition module and a method for distributing documents to the desired department.*

Keywords: *word weight, OCR, document flow, distribution, expert system, predictive analysis, Uzdoc collaboration software.*

Introduction. A centralized database of information resources and the possibility of document flow between structural divisions significantly increase productivity [1]. The problem of electronic document management is relevant not only in higher educational institutions, but also in other organizations. One of the problems in document management is document identification and distribution to responsible persons or software sections. This article proposes a recognition method and data mining. When working with documents, you have to face such problems as automated processing and distribution of documents that are available for execution. The solution to this problem is to develop an intelligent module that subsequently learns itself and automatically recognizes the document. In order to increase the efficiency of document flow, Uzdoc software was developed. This software recognizes and redistributes documents that are pending execution. The developed module recognizes the type of document, and the learning module, which subsequently determines which category a particular document belongs to, distributes and redirects the document to the desired address. This method of working with documents creates a number of conveniences and automates the process of processing and execution, which leads to efficient and high-quality implementation of decrees by the administration.

Automation of the document flow process has a number of advantages, the main one of which is the time efficiency of the execution and processing of certain decrees. The following is a list of department documents. Basically, in all departments, regardless of the direction and type of training, the same categories of documents and the tasks assigned to the department are identical. There are responsible persons who are responsible for the execution of documents. The main problem is that there is no module that recognizes the document type and categorizes it. This article proposes a solution to this problem. At the Department of System and Practical Programming at TUIT named after Al Khorazmi, research work is being carried out to develop a module for recognition, and specialized software is also being developed. Below are a number of department documents.

Table 1. Documents of the department "System and software":

№	Documents	Type of documents
1	Laws and decrees of the President of the Republic of Uzbekistan and the Supreme Council of the Republic of Uzbekistan.	Document type pdf
2	Orders and instructions of the rector and vice-rectors of the university concerning the activities of departments, and certificates of their implementation.	Document type pdf
3	Annual work plan and minutes of department meetings. Charter of the department.	Document type microsoft word
4	Decisions of the university council, methodological and academic council of the faculty.	Document type pdf
5	State educational standards, standard educational programs.	Document type pdf
6	Monitoring the implementation of individual work plans and calendar plans by teachers of the department.	Document type pdf
7	Assessment of student knowledge (base of written and oral test questions, written and electronic test questions, options for written work, block modules, etc.	Document type microsoft word
8	Reviews and patents of department teachers for scientific works and articles	Document type microsoft jpg

Materials and methods. The research work sets the task of text recognition using keywords to develop data mining modules. The predictive analysis module recognizes a document by keywords, determines which category a particular document belongs to and redirects it to the desired department [2].

To recognize documents, a preliminary expert method for identifying keywords was carried out. Based on keywords, the recognition module determines which category the document belongs to and sends it to the desired department. More frequently occurring words are taken as key words [2]. In documents, you may encounter the fact that the same words can appear in several documents at the same time. In this case, you should use phrases. For example, the word higher education appears in several documents at the same time. In this case, when searching for keywords, the phrase "Higher and secondary specialized education" was taken, and in another document "Law on Higher Education". Also, keywords should be considered a word that defines the essence of the document and is unique for this document. This way the issue of keyword identity is resolved. Below is a table clearly showing the identification of keywords and phrases. [2].

For convenience, we divide all input data into 4 categories. First category: laws and decrees of the President of the Republic of Uzbekistan. Second category Ministry of Higher and Secondary Education of the Republic of Uzbekistan, Ministry of Digital Technologies of the Republic of

Uzbekistan. The third category is a certificate of orders, instructions and their implementation by the rector and vice-rectors of the university. Each of these categories consists of one folder. The fourth category is department documents, curriculum, syllabus, department reports, information about graduates and students. The fourth category includes 23 folders.

Table 1. Fragment of keyword combinations.

№	Folder name	Frequency of occurrence of keywords or phrases	Keywords
1	References on the orders, tasks and their implementation of the university rector and vice-rectors regarding the activity of the department (5 years Article 14a, Article 65)	X_1	Higher and secondary special education, information, rector, vice-rector
		X_2	faculty, dean, students
		X_3	Ministry of Development, order',
		X_4	technologies, communication
		X_5	Secretary, report, decision
2	Annual agenda and minutes of department meetings	X_6	Work plan, agenda, attended, attended, number, plan
		X_7	excerpt, day, chair, order, spoke
		X_8	accepted, Agreed. agreed,
		X_9	systematic and applied programming, university, meetings, I confirm,
		X_{10}	faculty, software engineering, director, teachers, BMI, students

For 26 inputs that consist of specific documents. For each input data, several keywords are defined. In general, for 26 inputs, we have identified 557 keywords that can be used to recognize inputs into categories.

Based on the above data, an algorithm is constructed that determines the input data stream to determine keywords. For 26 folders, the input data consists of:

$$I - x_i : \bar{x}_i = \frac{1}{k_1}\sum_{i=1}^{k_1} x_i \ , \ at \ \ x_i = 4; \ \bar{x}_1 = 1$$

$$II - x_i : \bar{x}_i = \frac{1}{k_2}\sum_{i=k_1}^{k_2} x_i \ , \ at \ x_i = 4; \ \bar{x}_2 = 2 \ (1)$$

$$III - x_i : \bar{x}_i = \frac{1}{k_3}\sum_{i=k_2}^{k_3} x_i \ , \ at \ \ x_i = 4; \ \bar{x}_3 = 3 \ (2)$$

$$\dots$$

$$26 - -x_i : \bar{x}_i = \frac{1}{k_2}\sum_{i=k_1}^{k_2} x_i \ , \ at \ x_i = 4; \ \bar{x}_4 = 4 \qquad (3)$$

At $\bar{x}_1 \neq 0$ the input data belongs to the first category, for x¯_2≠0 the input data belongs to the second category, etc., for x¯_26≠0 the input data belongs to the 4th category. Starting from the

4th folder to 26, all documents belong to category 4. This way we can divide all input data into categories. This implies:

$$\bar{x}_4 = (\bar{x}_4 + \bar{x}_5 + \bar{x}_6 + \bar{x}_7 \dots + \bar{x}_{26}) \qquad (4)$$

We define a vector string x¯_1,x¯_2,x¯_3,x¯_4 and, based on the length of this vector, determine the category of the input string as follows: calculate the Euclidean length of the vector x¯

$$|x| = \sqrt{\bar{x}_1^2 + \bar{x}_2^2 + \bar{x}_3^2 + \bar{x}_4^2} \qquad (5)$$

Thus, if |x¯|≥4, then the input flow belongs to the 1st category, if 3≤|x¯|≥4, then it belongs to the 3rd category, if 2≤|x¯|≥3, then it belongs to 2nd category, if 1≤|x¯|≥2 belongs to the first category. By applying this recognition method to elements of the first category, you can determine what type of document the input string of the 1st category belongs to. Thus, if some input string x¯is expected, then by the length of the vector of the string x¯, it is possible to determine the category of this string, which allows the data to be processed by the necessary specialists.

Therefore, function (2) can be used to recognize normal strings constructed using keywords and phrases. Next, based on keywords and key phrases, an intelligent module is developed, which self-learns based on manually compiled data. Keyword extraction programs are often used to capture the main idea of a document and create or select keywords.

Step 1. Preliminary familiarization with the text and identification of keywords. Dividing text into sentences. Dividing a sentence into words, terms

Step 2. Calculate keyword weight. More frequently occurring words or phrases have a higher weight.

Step3. For each keyword, a weight is determined. The largest weight starts from 40 to the smallest weight equal to 10. Accordingly, priority is determined by the weight of the word.

Step 4: Eliminate duplicate keywords. One method is data compression. But in this case, a different method was used. For more accurate text recognition, not words are used, but key phrases and phrases. This eliminates repetition of keywords.

Step 5: Sort keywords by weight.

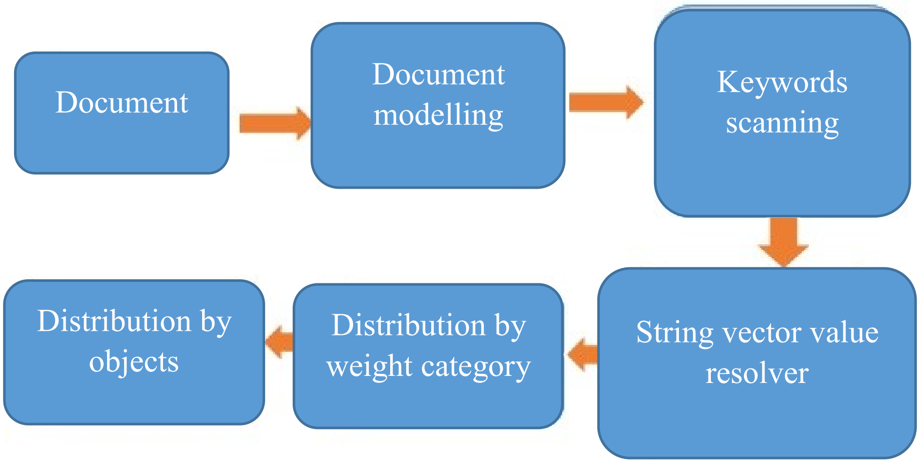

Figure 1.1. General scheme of recognition and execution

In this scheme, the document is submitted for execution. The document enters the software, which works online, through which you can exchange documents with other departments. The software has a special module for sending and transmitting documents to the desired department.

After processing, the document is sent to the required department. After receipt, you should determine the type of document, pdf, word, jpeg, etc. Once identified, a conversion process is implemented using technology to identify and convert scanned handwritten or printed text characters into machine-recognizable characters. OCR technology is implemented directly into Uzdoc software. The next step is scanning keywords. Keywords are scanned to determine which category (folder) the document belongs to. The document will be distributed by keywords [1]. those. Each keyword has its own weight. In this case, the determination of weight categories is done manually. At the next stage, the scanned document is transferred to the responsible person for execution. For example, if this is a syllabus, work program, etc. then it goes to the teacher Sharipov.B for execution, if the topics of the dissertation works then N. Khodzhieva, etc.

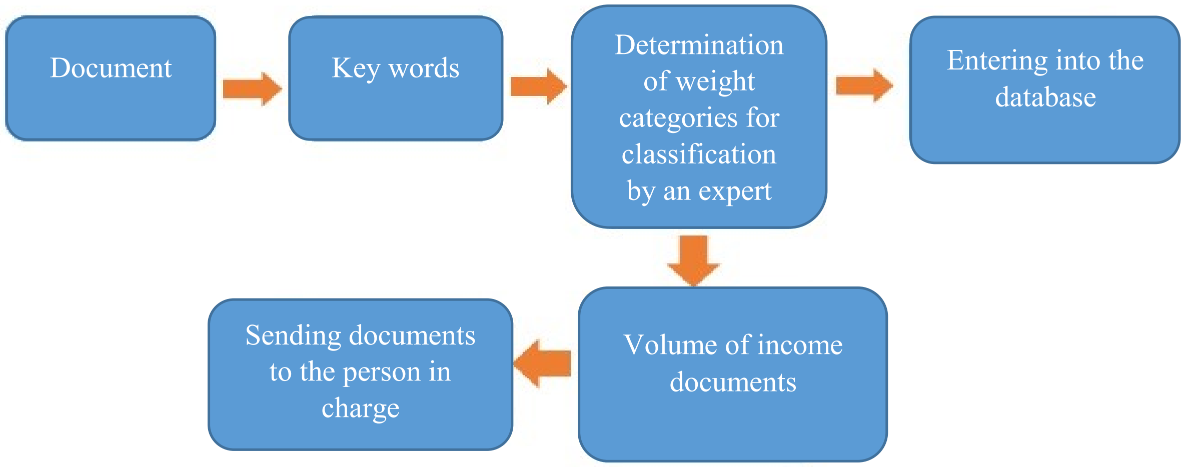
Document
Key words
Determination of weight categories for classification by an expert
Entering into the database
Volume of income documents
Sending documents to the person in charge

Figure 1.2. Algorithm for manual distribution by keywords

In this scheme, the document is submitted for execution. Keywords are defined manually [1].

Keywords are determined in the following way: the most frequently occurring words and words that define the essence of the document.

Once you have identified your keywords, you should categorize them. At this stage, keywords are recognized and weight categories are determined for classification by an expert, and further training takes place on its basis.

Certain keywords will be entered into the database. The documents are distributed to the responsible person. After distribution

1. Document for execution
2. Keywords
3. Recognition of keywords and determination of weight categories for classification by an expert, training occurs.
4. Certain keywords are entered into the database for further training of the software module and automatic determination of the category by keywords
5. Documents are distributed to the responsible person, in the program itself, automatically.
6. Volume of documents received for execution.

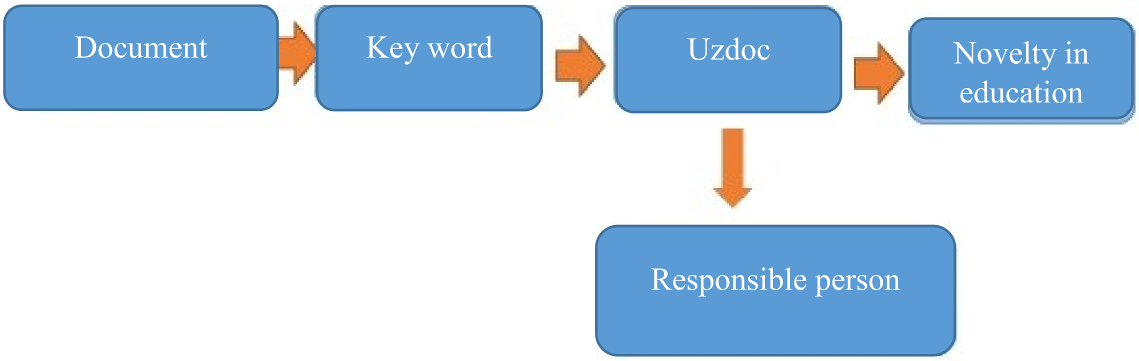

Fig1.3. Algorithm for automatic detection of document topics

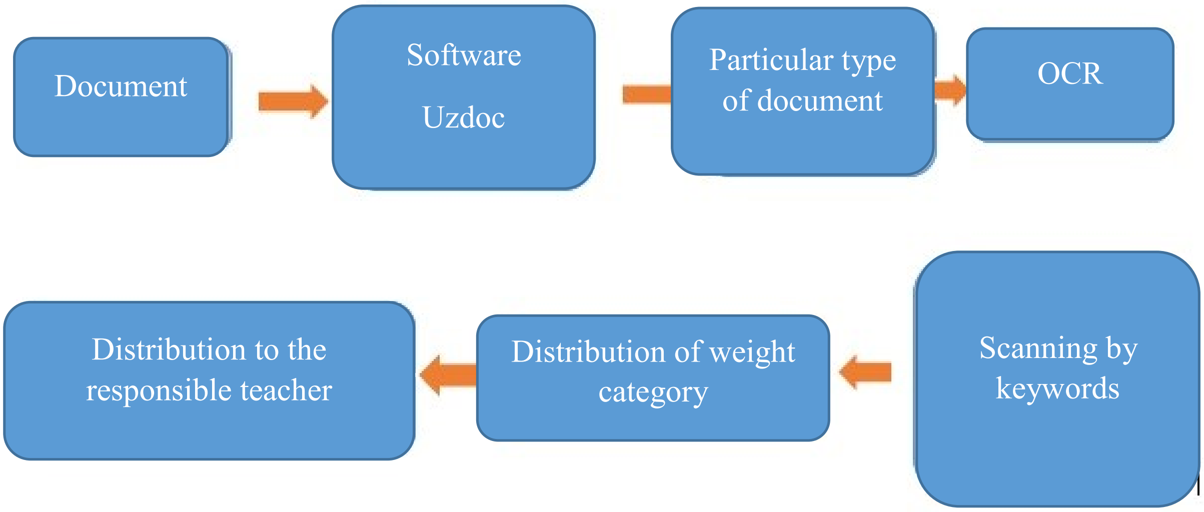
Document
Software
Uzdoc
Particular type of document
OCR
Scanning by keywords
Distribution of weight category
Distribution to the responsible teacher

Figure 1.4. Algorithm of the intelligent software module UzDoc

Next, the already distributed keywords are scanned. Which category (folder) does the document belong to?

Documents are automatically processed in the program itself.

1. Document for execution
2. Keywords
3. Recognition of keywords and determination of weight categories for classification by an expert, training occurs
4. Keywords are entered into the database
5. Documents are distributed to the responsible person, in the program itself, automatically
6. Volume of documents received for execution.

Results. As a result of experiments conducted on documents, an intelligent module was developed for recognizing and distributing documents among departments. Based on the obtained key phrases, an automatic document recognition module was built [1]. The resulting algorithm allows you to build an intelligent module that is capable of recognizing documents by keywords. An algorithm for constructing Uzdoc software has been proposed. Manual definition of keywords was performed. Based on it, an automatic document recognition module has been developed. Uzdoc software has a number of advantages compared to similar programs. The advantage is that it is aimed at processing and executing documents of the department and is designed in such a way that it covers all areas of study and type of activity in a higher educational institution.

Conclusions and Conclusion. As a result of the research, a module was developed for the Uzdoc software, which responds and recognizes the document using keywords and sends the document; these can be files of various types containing data on the categories of documents given above [2]. The process of recognizing and categorizing information is a complex and difficult task. The process of manually creating and identifying key phrases has a number of its own characteristics. This takes quite a lot of time. Depending on the type of document, it is forwarded to the responsible person who reviews the file and executes it. After execution, it is entered into the database. In this way, a database is created, which will later be a source for the training system.

REFERENCES

1. Abduvalieva Zebiniso Abdulhamidovna1 , Marisheva Larisa Timofeevna1 and Latipova Nodira Xalimovna1 , Sheyna Nataliya Evgenevna2. STRUCTURE AND FUNCTIONAL FEATURES OF DOCUMENT MANAGEMENT SYSTEMS ON THE EXAMPLE OF THE DEPARTMENT. JOURNAL OF NORTHEASTERN UNIVERSITY
2. Ricardo Campos a,e,∗ , Vítor Mangaravitee,g , Arian Pasquali e , Alípio Jorge b,e , Célia Nunes c,f , Adam Jatowt d. YAKE! Keyword extraction from single documents using multiple local features. 2019 Elsevier Inc. All rights reserved.
3. Isaeva, M., Yoon, H., Y.: Paperless university — How we can make it work?. In: 15th International Conference on Information Technology Based Higher Education and Training (ITHET). pp. 1–8 (2016).
4. [Luo, H., Fan, Y., Wu, C.: Overview of Workflow Technology. J. Softw. 11, 78-82
5. Fan, Yusun.: Base on Workflow Management Technology. Beijin:Tsinghua University Press, 32, (2001)
6. Chen, Hong-na, Zu, Xu, Zhou, Feng: On the Developing Situation, Research Content and Trend of Workflow Technology. Journal of Chongqing Instiute of Technology. 20(2), 65-69 (2006)
7. Li, Zhao, Qing, Li, Farong, Zhong: A Visual Modeling Framework of Workflow Systems Based on CCS. Semantics, Knowledge and Grid. Fifth International Conference. pp. 200-207 (2009)
8. Dinesh, P, Mital, Goh, Week, Leng School of Electrical & Electronic Engineering Nanyang Technology University Nanyang Avenue, Singapur 2263. Text segmentation for automatic document processing. Rosemont, IL, USA. pp. 132-133 (1995)
9. [8] Jian, LI, Na, YU. Research on Document Management System Based on Streaming Storage Technology. International Conference on Industrial Engineering and Engineering Management. pp. 558-562 (2011)
10. Ki Won Lee, Jeong Seok Kang and Hong Seong Park. DocT – Document Management and Testing Tool for Robot Software. International Conference on Ubiquitous Robots and Ambient Intelligence. IEEE. pp. 413-416 (2014)
11. Kruchinin S., V., Bagrova E. V., Chair of Economics, Management and Science Noyabrsk Institute of Oil and Gas (branch) TIU in Noyabrsk. Systems of Electronic Document Management in Russian Education. Pros and Cons. IEEE. pp. 628-630 (2019)

12. Li Sui, Gengchen Shi, Ping Song School of Aerospace Science and Engineering Beijing Institute of Technology Beijing, Xingyu Yuan Prosten Technology Holdings. Design and Implementation of ISO Document Management System. International Conference on Computer Science and Software Engineering. (2008)
13. Li Kai-Qi1,2 Diao X ing-Chun2 Weng Nian-Feng1,2 Li Ge1,2. Research on the Management Framework of Personal Desktop Document DataSpace. International Conference on Information Science and Engineering. pp 959-962 (2009)
14. Yong Wang Bi-yan Sun Fei Cheng Department of Information Engineering, Archives of WenZhou University R&D Center. Electronic-Document-Based Management Process Model for Image Archives in Universities. International Conference of Information Technology, Computer Engineering and Management Sciences. Pp 57-60 (2011)
15. Mirjana A. Andric, Member, IEEE and Wendy Hall. Using Metadata for Information Retrieval in Document Management Systems. Belgrade, Serbia. pp. 1093-1096 (2005)
16. Tasi Tyrväinen, Tero Päivärinta. On Rethinking Organizational Document Genres for Electronic Document Management. Hawaii International Conference on System Sciences. pp. 1-10 (1999)
17. Katarzyna Styk, Jakub Liszcz, Klaudia Drobek. Basic Project Management Documentation Based on the Example of the Student Project AGH Lean Line. pp. 45-49 (2019)
18. W.W. Cohen, Y. Singer, “Context-sensitive learning methods for text categorization”, ACM Transactions on Information Systems (TOIS), 1999, vol. 17, №. 2, pp. 141-173.
19. E. Stamatatos, N. Fakotakis, G. Kokkinakis, “Automatic text categorization in terms of genre and author”, Computational linguistics, 2000, vol. 26, №. 4, pp. 471-495.
20. M. Radovanović, M. Ivanović, “Document representations for classification of short web-page descriptions”, International Conference on Data Warehousing and Knowledge Discovery, Springer, Berlin, Heidelberg, 2006, pp. 544-553.
21. F. Sebastiani, “Machine learning in automated text categorization”, ACM computing surveys (CSUR), 2002, vol. 34, №. 1, pp. 1-47.
22. S. Wang, L. Jiang, C. Li, “Adapting naive Bayes tree for text classification, Knowledge and Information Systems, 2015, vol. 44, №. 1, pp. 77-89.