Methodology for Evaluating the Performance of Websites and its Optimization Using Data Mining Techniques

Olena Sorokivska[a], Halyna Kozbur[a], Iryna Strutynska[a], Liliya Melnyk[a] and Roman Sherstiuk[a]

[a] *Ternopil Ivan Puluj National Technical University, 56, Ruska Street, Ternopil, 46001, Ukraine*

Abstract

The article presents the results of testing and improvement of the developed comprehensive methodology for evaluating websites of authorities and local self-government. The study was conducted with the aim of evaluating the openness of the work of authorities and local self-government, the availability of council decisions, content, ease of use of electronic resources. In the process of research, the authors of the article developed a system of evaluation criteria, as well as conducted an in-depth analysis of the technical parameters of websites.

The possibility of improving the methodology is considered. One of the ways was to reduce the dimensionality of the task, which consisted in reducing the number of sub-indicators and/or main indicators, according to which the evaluation was carried out. Correlation analysis was used to identify indicators that allow merging or removing from the system of criteria.

In order to understand the state and functioning of community websites as a whole, identify problem groups and develop a system of recommendations to improve their work, the data mining technique was applied, namely, cluster analysis. A clustering model has been built that allows you to find hidden structures in the information content of websites of communities (authorities), reveal complex relationships between information resources, and separate groups of communities whose websites have similar characteristics. The results of clustering can be used to make decisions in e-government processes and improve the perception and use of information by citizens. The proposed approach will be of particular value in the case of research scaling.

Keywords 1

Website, e-government, authorities and local self-government, data mining, clustering, correlational analysis

1. Introduction

Today, one of the priority directions for the development of e-government in Ukraine is the creation of a digital ecosystem for territorial communities. Such an ecosystem will help improve the quality of life of residents, develop the economy, and ensure more effective management of the territory. It includes the following components:

1. Electronic government – digital tools and services that allow interaction between government institutions and residents (electronic services, electronic declarations, online consultations and other tools that simplify interaction with government bodies).

CITI'2023: 1st International Workshop on Computer Information Technologies in Industry 4.0, June 14–16, 2023, Ternopil, Ukraine
EMAIL: Soroka220996@gmail.com (A. 1); kozbur.galina@gmail.com (A. 2); strutynskairy@gmail.com (A. 3); liliana.mel0512@gmail.com (A. 4); romsher85@gmail.com (A. 5)
ORCID: 0000-0001-8549-2910 (A. 1); 0000-0003-3297-0776 (A. 2); 0000-0001-5667-6569 (A. 3); 0000-0001-8844-5490 (A. 4); 0000-0001-6253-9421 (A. 5).

2. Digital infrastructure – infrastructure necessary to support digital initiatives, such as high-speed Internet, communication networks, cloud services, remote servers, digital data storage and others.
3. Digital services for citizens – digital tools and services that allow residents of the territorial community to be more active participants in governance, including online citizen meetings, petitions, online voting, and others.
4. Digital services for business – digital tools and services that allow business to be more competitive, including electronic procurement system, online business registration, electronic reporting system and others.
5. Digital educational environment – digital tools and services that allow residents of the territorial community to develop their skills and abilities in the learning process.

The site plays a key role in the formation of the community's digital ecosystem, because it allows a wide range of people to find the necessary information about the activities of the community, decisions of authorities and various services provided by the community. This ensures transparency and openness in the work of local self-government bodies and promotes interaction with the community. The site is also a platform for receiving messages and requests from residents, providing answers to questions, publishing news and announcements, organizing online surveys and consultations, etc.

It can be used to provide electronic document circulation, electronic voting, organization of online consultations with the community, etc. The site can help promote the community, promote the development of tourism, culture and other industries that ensure the development of the local economy and increase the standard of living of residents.

That is why the purpose of this study was to understand the state and functioning of community sites as a whole, to improve the system of criteria for evaluating the effectiveness of websites, to identify successful and problematic groups using the data mining techniques for the further development of a system of recommendations so that websites become more effective.

2. Related Works

Digital transformation of territorial communities is an important issue that requires attention and research in the era of information society. Researchers, scientists, and practitioners (local government bodies) around the world are exploring the issue of building digital societies and forming certain digital ecosystems of the community. For example, in the study [2], authors Karen Mossberger, Caroline J. Tolbert, and Ramona S. McNeal examine the relationship between the Internet, society, and civic engagement. They argue that access to digital technologies and their use can enhance democratic activity and promote social inclusion, but only if citizens are equipped with the skills and knowledge necessary to navigate the digital landscape. Author Jochen Schmidt, in the study [3], considers the issue of digital transformation of territorial communities through the prism of interaction and location, physical and digital infrastructure in cities and their impact on the digital transformation of territorial communities.

In the work of author Bashuk A.I. [4], a theoretical and applied concept of implementing communication strategies of state authorities in the conditions of the information society is developed. Regarding the importance of a website for communication between government and society, author Michael A. Hollisworth in his work "Building Trust in Government Websites: A Review of the Research Literature on Online Service Quality, Trust and Civic Engagement" [5], investigates how the quality of online services affects trust in government websites and civic engagement.

Research conducted by many foreign and domestic scientists, such as Sarmad Ali, Inayatullah Shah, and Abdul Rauf [6]; Weng Marc Lim, Min Choon Yap, and Voon Hsien Lee [7]; Saeed-Ul Hassan, Hafiz Farhan Ahmad, and Muhammad Salman Mazhar [8]; Maksymenko A.O. [9]; Lypchuk M. [10]; and others, show that a website is an important tool for improving communication between government and the community, especially for providing access to information about the activities of local government, services and programs it provides. Factors that can affect the effectiveness of the website in this context are highlighted. In the study [11] by authors Joshua A. Turner, Gregory S. Dawson, and Robert D. Galliers, the functionality and design of websites of diversified government

institutions around the world were analyzed and an evaluation was given on their compliance with standards for accessibility and usage.

In a study [12] conducted in Ukraine in 2019, it was shown that 80% of local government authorities have their own website, but only 22% of them have full information about their activities and services. This indicates the need to improve the quality and accessibility of information on local government websites.

Regarding the methodology for assessing the websites of local self-government bodies, it is worth noting the works of foreign authors Huseyin Ince, Orhan Dolu, and Mehmet Nafiz Aydin in their research [13], which analyzed the evaluation of local self-government websites from the perspective of citizens. They proposed an evaluation methodology based on five criteria, including information accessibility, website functionality, and effectiveness. In the work [14], authors Jose A. Castillo-Manzano, Mercedes Castro-Nuño, and Juan A. Parra-Domínguez evaluated the quality of local self-government websites in Spain. The authors proposed an evaluation methodology that uses 18 parameters to determine the level of website effectiveness and quality. In the research of authors Wenjing Pan, Rui Guo, and Yuan Cheng [15], a comparative analysis of the quality of local self-government websites in China was conducted. They proposed an evaluation methodology that uses 9 criteria, including accessibility, interaction with citizens, and ensuring confidentiality.

In the works of domestic scientists, issues of the functioning of websites of state authorities and local self-government bodies are investigated. For example, in the work [16] by author Korzh I.F., the functioning of websites was analyzed regarding compliance with current legislation, which regulates issues related to the requirements for their content and information filling [17, 18]. However, more attention should be paid to the issue of monitoring the websites of territorial communities in Ukraine, as the relevance and importance of web resources are extremely high, and research in the relevant context has not been conducted.

3. Proposed Methodology

During September-December 2022, the authors of the article monitored the websites of authorities and local governments of 14 territorial communities (hereinafter referred to as TC) of Ternopil Oblast, as well as the Ternopil Oblast Military Administration. The study was conducted with the aim of evaluating the openness of the work of authorities and local self-government, the availability of council decisions, content, to ease the use of electronic resources.

Among the studied communities were Ternopil, Baikivtsi, Velyki Hayi, Pidhorodne, Zalishchyky, Lanivtsi, Vyshnivtsi, Shumska, Skalat, Chortkiv, Zboriv, Pochaiv, Terebovlya, and Mykulyntsi ones. In the process of research, the authors of the article developed separate evaluation criteria, as well as conducted an in-depth analysis of the technical parameters of the sites. In the evaluation process, each site received a score based on individual parameters (Table 1).

Table 1

Criteria for scoring sites of territorial communities of Ternopil region

Assessment Point	Assessment Criteria
0 – «No information»	This type of information is not presented on the community's official website
1 – «Partially filled»	This type of information is presented partially or not in full
2 – «Information is complete»	This type of information is presented in its entirety volume

The system of indicators for each of the 15 sites contained five elements:

$$\{w_{k1},\ w_{k2},\ w_{k3},\ w_{k4},\ w_{k5}\},\ \ k = 1,2,\dots,15,$$

where

w_{k1} – is the overall rating of the *k*-th site (in points) according to indicator 1 "Availability of important general information on sites";

w_{k2} – overall evaluation of the *k*-th site (in points) according to indicator 2 "Availability of important documentation on sites";

w_{k3} – overall evaluation of the $\underline{k}$-th site (in points) according to indicator 3 "Availability of additional information on sites";
w_{k4} – overall evaluation of the k-th site (in points) according to indicator 4 "Availability of network functions and site efficiency (e-democracy)";
w_{k5} – the overall assessment of the k-th site (in points) according to indicator 5 "Technical requirements for the site".
The final evaluations W_k of the websites were obtained according to the formula:

$$W_k = \sum_{i=1}^{5} w_{ki}, \quad k = 1,2,\dots,15. \qquad (1)$$

Each of the five elements (criteria) of the developed evaluation system contained a number of subcriteria. The maximum possible value of the final assessment for each site is 108 points (Table 2).

Table 2
Summary table of evaluation of sites by main indicators

Name (center) of the territorial community (authority)	Website denotation	Site assessment by indicators, in points					Final assessment, W_k, in points	Final assessment W_k, in % of the maximum
		w_{k1}	w_{k2}	w_{k3}	w_{k4}	w_{k5}		
Ternopil OMA	S_1	10	13	19	8	12	62	57,4%
Ternopil CC	S_2	19	25	23	12	15	94	87,0%
Baykivtsi	S_3	14	15	22	13	18	82	75,9%
Velikiy Hlybochok	S_4	13	6	18	13	17	67	62,0%
Zalishchyky	S_5	13	8	17	10	11	59	54,6%
Lanivtsi	S_6	14	16	18	14	14	76	70,4%
Vyshnivets	S_7	12	11	13	12	11	59	54,6%
Shumsk	S_8	14	20	20	14	18	86	79,6%
Skalat	S_9	14	18	14	14	14	74	68,5%
Pidgorodne	S_{10}	15	5	12	11	16	59	54,6%
Chortkiv	S_{11}	13	26	14	14	14	78	75,0%
Zboriv	S_{12}	9	11	14	12	18	64	59,3%
Pochayiv	S_{13}	8	22	13	2	16	61	56,5%
Terebovlya	S_{14}	12	14	13	7	12	61	53,7%
Mykulyntsi	S_{15}	11	10	16	14	17	68	63,0%

According to the results of the analysis, Ternopil City Council, Shumsk Territorial Community and Baikivtsi Territorial Community were among the three leaders in terms of the main site assessment indicators. The results of the research were brought to the leaders of all communities to optimize the work of the sites and borrow the experience of progressive communities in the formation of a digital ecosystem.

An important task of this study, in addition to directly evaluating the sites, was to test the very methodology of site research. Improving the methodology will increase the efficiency of further research and may be useful for scaling the research to individual regions or the country as a whole.

4. Results

4.1. Analysis of the criteria system for sites assessment

A total of 54 sub-indicators were used in the assessment of the state and functioning of the sites in five sections of the assessment based on the main indicators. In order to analyze the effectiveness of the developed system of indicators for the sites' performance assessment, the columns with ratings, of summary table 2 were used for each of the communities for each of the main indicators. One of the

ways to improve the methodology was to reduce the dimensionality of the task, which consisted in reducing the number of sub-indicators and/or main indicators by which the assessment was conducted. We note that reducing the size of the criterion system should not lead to the loss of important information about the operation of websites.

To identify a possible correlation between the 5 main site assessment criteria, pairwise values r_{ij} of Pearson correlation coefficients were calculated for all possible combinations of criteria:

$$r_{ij} = \frac{\sum_{k=1}^{15}(w_{ki}-\overline{w_i})(w_{kj}-\overline{w_j})}{\sqrt{\sum_{k=1}^{15}(w_{ki}-\overline{w_i})^2 \sum_{k=1}^{15}(w_{kj}-\overline{w_j})^2}}, \quad . \quad i = \overline{1,5}; \; j = \overline{1,5} \tag{2}$$

where

$\overline{w_i}$ – is the average value (in points) obtained according to the i-th criterion;

$\overline{w_j}$ – is the average value (in points) obtained by the j-th criterion, both for a sample of 15 sites.

Note that the value of the correlation coefficient does not depend on the order of the selected criteria $r_{ij} = r_{ij}$, therefore, the correlation matrix is symmetric for all pairs of criteria. All elements of the main diagonal $r_{ii} = 1$. All obtained values of pairwise correlation coefficients are entered in the correlation matrix (Table 3).

Table 3

The value of pairwise correlation coefficients in the form of a correlation matrix

Name of the assessment indicator	1. Availability of important general information	2. Availability of important documentation	3. Availability of additional information	4. Availability of network functions and site efficiency (e-democracy)	5. Technical require-ments for the site
1. Availability of important general information	1,0000				
2. Availability of important documentation	0,2047	1,0000			
3. Availability of additional information	**0,5035**	0,2409	1,0000		
4. Availability of network functions and site effectiveness (e-democracy)	**0,5171**	-0,0495	0,3065	1,0000	
5. Technical requirements for the site	0,0050	0,0404	0,2563	0,2632	1,0000

The values highlighted in the table (greater than 0.5) indicate the presence of a positive linear correlation between:

1) 1st and 3rd criteria (0.5035);

2) 1st and 4th criteria (0.5171).

This indicates that positive changes in the values of one of the indicators lead to positive changes in the average values of the other. A comparative analysis of the sub-indicators included in the specified pairs of criteria will make it possible to make a decision on merging the main criteria that showed a positive correlation, or removing one criterion from the pair that showed a positive correlation from the system of criteria.

A detailed analysis of the content of the sub-indicators of the 1st and 3rd criteria showed that the more a community or authority places important general information on its website, the more additional information is presented on it. A detailed analysis of the content of the sub-indicators of the 1st and 4th criteria showed that in communities where important general information is posted in a sufficient amount, as a rule, there are quite developed network functions and the indicators of the site's effectiveness are significantly higher. So, the research results showed that the developed methodology can be improved by consolidating the criteria system (up to 4 indicators), combining criteria 1 and 3 into the group "Availability of important general and additional information on the sites".

4.2. Clustering of sites based on assessment results

In order to understand the state and functioning of community sites as a whole, to identify successful and problematic groups, as well as to develop a system of recommendations so that websites become more effective, the data mining technique was applied, namely the clustering of sites according to the values of 5 indicators, similar to what made by the authors in [20].

To build the clustering model, the Cluster option of the Data Mining add-on in MS Excel 2010 was used. The clustering model was obtained by the Scalable K-means method, which allows the selection of the desired number of K clusters and the possibility of working with large volumes of data [21]. The K-means method is a well-known unsupervised machine learning method that combines elements with the smallest distances into one cluster, and at the same time assigns elements with the largest distances to different clusters. The k-means algorithm calculates the squared Euclidean distances between elements in a cluster and the cluster mean, and converges on a final set of K clusters when that sum reaches its minimum value [22].

The Euclidean distance d_E between the two sites S_n and S_m with sets of estimates $W_n = (w_{n1}, w_{n2}, w_{n3}, w_{n4}, w_{n5})$ and $W_m = (w_{m1}, w_{m2}, w_{m3}, w_{m4}, w_{m5})$ respectively, is calculated by the formula:

$$d_E(S_n; S_m) = \sqrt{\sum_{i=1}^{5}(w_{ni} - w_{mi})^2}\,, \tag{3}$$

where

S_n, S_m – sites, the Euclidean distance between which is calculated;

w_{ni} – the value of indicators for the site S_n, obtained according to 5 criteria;

w_{mi} – the value of indicators for the site S_m, obtained according to 5 criteria.

The result of building a clustering model for 3 clusters, namely the number of elements in each cluster and the main characteristics of the clusters, is shown in Figure 1. The cluster analysis made it possible to distinguish groups of community sites based on their similarity (10 participants, 3 participants, 2 participants).

The 1st cluster included 10 sites that received average scores according to the criterion "Availability of important documentation" (43.66% of sites received scores of 15-19 out of 30 possible) and above average scores according to the indicator "Availability of important general information on sites" (57.1% of sites received 14 points out of 20). 42.86% of the participants of the first cluster received the maximum points according to the indicator "Technical requirements for the site", the same percentage of sites received a high score according to the criterion "Availability of network functions and effectiveness of the site (e-democracy)" (14 points out of 16 possible).

The 2nd cluster included 3 sites that received low points according to the indicator "Availability on sites of important documentation" (81.28% of sites received 5-9 points out of 30 possible) and average points according to the criterion "Availability on sites of important general information" (2 sites received 13 points out of 20 possible, 1 site – 15 points).

The 3rd cluster included 2 sites, the main difference from other evaluation participants is the presence of important documentation (the sites received 20-26 points out of a possible 30).

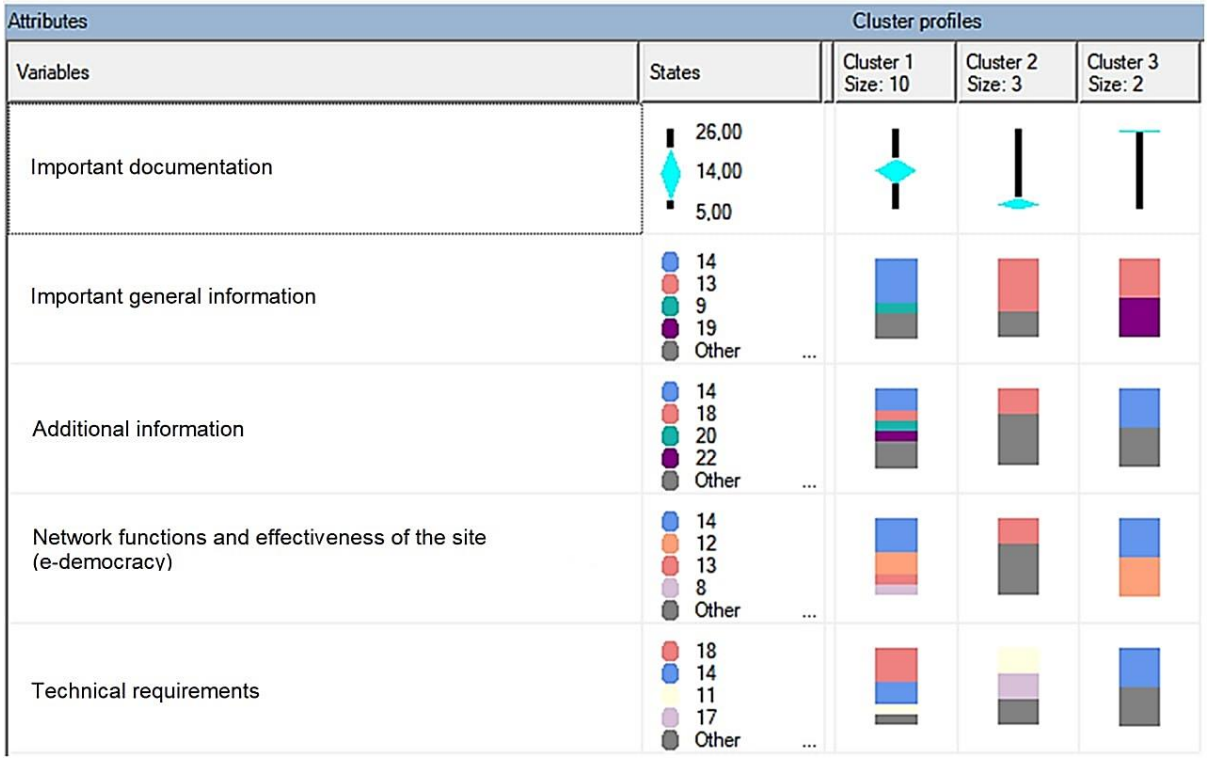

Figure 1: Main characteristics of the 3 site clusters

The built clustering model allowed us to find hidden structures in the information content of community (authority) websites and to separate groups that have similar characteristics. The results of clustering can be used to identify complex relationships between information resources and further develop a system of recommendations to improve the perception and use of information by citizens. In addition, it can be useful in reducing data processing costs, since clustering can help identify the most important characteristics of information resources that will be used to make decisions in e-government processes.

The proposed approach will be of particular value in the case of research scaling. In this case, it is necessary to justify the choice of the clustering method, its settings, as well as methods of determining the quality of the built clustering model.

5. Conclusions

The proposed methodology for evaluating community websites when used can have several key effects on the development of the e-government system:

Improving interaction with the community. Evaluating website performance will help to understand how the community is interacting with the site and whether the website is meeting their needs. Based on this information, authorities and local governments can identify shortcomings and improve the functionality of the site to ensure better interaction with the community.

Attracting new users. If a website is not effective in engaging with the community, new users may not find the information they need and may not return to the website. Evaluating website performance will help identify these issues and improve site functionality to attract new users.

Assessment of achievements and effectiveness. Evaluating the effectiveness of the website will help to understand the development of the community and its effectiveness in achieving its goals. For example, if a website is successful in promoting community initiatives and programs, then this may indicate the success of the community.

Reputation improvement. An effective website can help improve the reputation of a community, particularly if the website contains useful information for citizens. Evaluating the performance of a website can help identify weaknesses and improve it to ensure a more positive reputation.

The possibility of improving the methodology was considered. The results of the correlation analysis of the criteria for evaluating the effectiveness of websites showed that the developed methodology can be improved by consolidating the system of criteria (up to 4 indicators), combining criteria 1 and 3 into the group "Availability of important general and additional information on the sites".

For a deeper understanding and possible typification of community approaches to website occupancy and maintenance, a website clustering model was constructed using the K-means method. According to the results of clustering, 3 groups of communities (10, 3 and 2 participants) were singled out, the content and effectiveness of which websites are similar. In the case of scaling up the research, the results of clustering can be used to identify complex relationships between information resources and further develop a system of recommendations for improving the perception and use of information by citizens. In this case, it is necessary to justify the choice of clustering method, its settings, as well as methods of determining the quality of the built clustering model.

The proposed methods of improving the methodology for evaluating the effectiveness of websites (correlation analysis of the main evaluation criteria and cluster analysis of community websites) can be useful for reducing data processing costs, as clustering can help identify the most important characteristics of information resources that will be used for decision-making in e-government processes. In further work, it is interesting to consider the issue of optimization of the methodology for researching the effectiveness of websites using various techniques and methods of intellectual data analysis.

6. References

[1] Sorokivska, O., Panukhnyk, O., Fedotova, Y., & Popadynets, N. Communicative interaction between the public and the government authorities in the conditions of the pandemic: analysis of challenges and opportunities. Journal of Innovations and Sustainability (2022) 6(3). https://doi.org/10.51599/is.2022.06.03.07

[2] Mossberger, K., Tolbert, C. J., & McNeal, R. S. Digital Citizenship: The Internet, Society, and Participation. MIT Press, 2016.

[3] Schmidt, J. The Digital Transformation of Cities: The Interplay between Location, Physical Infrastructure and Digital Infrastructure (2019).

[4] Bashuk, A. I. Communication Strategies of State Authorities in the Information Society: Monograph. Kamianets-Podilskyi: TOV "Drukarria "Ruta". ISBN 978-617-7626-64-9 (2019).

[5] Hollisworth, M. A. Building Trust in Government Websites: A Review of the Research Literature on Online Service Quality, Trust and Civic Engagement. (2017).

[6] Sarmad Ali, Inayatullah Shah, and Abdul Rauf. Impact of E-Governance Websites on Good Governance Practices: An Empirical Study of Pakistan. Journal of Information Technology Management (2015). Vol. 26, No. 4.

[7] Weng Marc Lim, Min Choon Yap, and Voon Hsien Lee. The Impact of Government Websites' Service Quality on Citizens' Satisfaction and Trust. Journal of Computer Information Systems (2016). Vol. 56, No. 3.

[8] Saeed-Ul Hassan, Hafiz Farhan Ahmad, and Muhammad Salman Mazhar. Citizens' Perception of Local Government Websites and Their Impacts on E-Governance: Evidence from Pakistan. Sustainability (2020). Vol. 12, No. 6.

[9] Maksymenko, A.O. The role of information content of websites of united territorial communities in the process of forming center-periphery interactions. Business Inform. (2020). Vol.10, 233-240. DOI: 10.32983/2222-4459-2020-10-233-240.

[10] Lypchuk, M. Internet representation of the government as a mechanism for dialogue interaction and a key element of democratization of public administration. Effectiveness of Public Administration. (2016). 1/2(1), 188-196. URL: http://www.lvivacademy.com/vidavnitstvo_1/edu_46/fail/ch1/23.pdf

[11] Joshua A. Turner, Gregory S. Dawson, and Robert D. Galliers. Government Websites: An Overview of their Functionality and Design. URL: https://granicus.com/government-website-design/

[12] Monitoring of the Information Content on the Official Websites of Central and Local Executive Authorities in the First Half of 2019. URL: https://www.kmu.gov.ua/news/derzhkomteleradio-oprilyudniv-rezultati-monitoringu-veb-sajtiv-organiv-vikonavchoyi-vladi
[13] Huseyin Ince, Orhan Dolu та Mehmet Nafiz Aydin. Assessment of Local Government Websites in the Age of E-Government: An Analysis from the Citizen's Perspective. (2021).
[14] Castillo-Manzano, J.A., Castro-Nuño, M., & Parra-Domínguez, J.A. Evaluating the quality of local government websites in Spain. Revista de Administración Pública, (2017). 53-88.
[15] Pan, W., Guo, R., & Cheng, Y. Evaluating Local Government Websites: A Comparative Study in China. International Journal of Public Administration, (2020), 43(11), 963-975. DOI: 10.1080/01900692.2019.1684975.
[16] Korzh, I.F. Websites of state authorities and local governments: mechanisms for accessing public information. (2018) (9-16). DOI:10.37750/2616-6798.2018.2(25).270655.
[17] Order of the State Committee for Information Policy, Television and Radio Broadcasting of Ukraine and the State Committee for Communications and Informatization of Ukraine on the functioning of websites of executive authorities (2002). URL: http://zakon5.rada.gov.ua/laws/show/z1022-02.
[18] Order of the State Committee for Information Policy, Television and Radio Broadcasting of Ukraine and the State Committee for Communications and Informatization of Ukraine on the approval of the Procedure for information filling and technical support of the Unified Web Portal of executive authorities and the Procedure for the functioning of websites of executive authorities (2002). URL: http://zakon3.rada.gov.ua/laws/show/z1021-02.
[19] O. Sorokivska, I. Strutynska, L. Melnyk. A Comprehensive Report Based on the Sociological Study Results: Challenges and Opportunities of Communicative Interaction Between the Public and the Authorities, TNTU, 2022.
[20] Strutynska I., Kozbur, H., Dmitrotsa L., Bodnarchuk I., Hlado O. (2019). Small and Medium Business Structures Clustering Method Based on Their Digital Maturity. 278-282. 10.1109/PICST47496.2019.9061464.
[21] Bahmani, B., Moseley, B., Vattani, A., Kumar, R., & Vassilvitskii, S. Scalable k-means++. Proceedings of the VLDB Endowment, (2012) 5(7), 622-633. Retrieved from https://theory.stanford.edu/~sergei/papers/vldb12-kmpar.pdf.
[22] Microsoft Clustering Algorithm Technical Reference (2022). Retrieved from https://learn.microsoft.com/en-us/analysis-services/data-mining/microsoft-clustering-algorithm-technical-reference?view=asallproducts-allversions.