UDC 004.912
DOI: 10.15587/1729-4061.2022.265961

ANALYSIS OF THE STRUCTURE OF WEB RESOURCES USING THE OBJECT MODEL

Stanyslav Dykhanov
Postgraduate Student*
Natalia Guk
Corresponding author
Doctor of Physical and Mathematical Sciences, Professor, Head of Department*
E-mail: natalyguk29@gmail.com
*Department of Computer Technology
Oles Honchar Dnipro National University
Gagarina ave., 72, Dnipro, Ukraine, 49010

The methodology for analyzing the structure of a web resource using an object model, which is based on the description of the page in HTML and using style sheets, has been proposed. The object of research is a web resource page, the model of which is depicted as a DOM tree. Data on the structural elements of the tree are supplemented with information about the styles of the design of the pages. To determine the similarity of pages, it is proposed to apply a criterion that takes into account the structural and stylistic similarity of pages with the corresponding coefficients. To compare page models with each other, the method of aligning trees will be used. Editing distance is used as a metric, and renaming operations, deleting, and adding a tree node is used as editing operations. To determine the similarity in styles, the Jaccard metric is used. To cluster web pages, the k-means method with a cosine distance measure is applied. Intracluster analysis is carried out using a modification of the Zhang-Shasha algorithm. The proposed approach is implemented in the form of an algorithm and software using Python programming language and related libraries. The computational experiment was performed to analyze the structure of individual websites existing on the Internet, as well as to group pages from different web resources. The structure of the formed clusters was analyzed, the RMS similarity of elements in the middle of the clusters was calculated. To assess the quality of the developed approach for the tasks under consideration, expert partitioning was built, the values of accuracy and completeness metrics were calculated. The results of the analysis of the structure of the web resource can be used to improve the structure of the components of the web resource, to understand the navigation of users on the site, to reengineer the web resource

Keywords: web resource, DOM tree, tree editing distance, similarity in structure and style

Received date 08.08.2022
Accepted date 12.10.2022
Published date 30.10.2022

***How to Cite**: Dykhanov, S., Guk, N. (2022). Analysis of the structure of web resources using the object model. Eastern-European Journal of Enterprise Technologies, 5 (2 (119)), 6–13. doi: https://doi.org/10.15587/1729-4061.2022.265961*

1. Introduction

The Internet is a dynamic environment where significant amounts of information are stored, the volume of which is constantly and rapidly increasing. Given the amount of data and the variety of its subject matter, it is almost impossible to manage it using conventional tools, so there is a need to structure the information. The most common approach for analyzing the structure of websites is the clustering of resource pages by attributes. Semantic knowledge can be used as attributes, for example, data on the semantic core of a web resource or keywords in user requests. To determine the linked segments of a web resource, data on links between pages in the form of links or user paths between resource pages is used. Also interesting for the analysis are the statistical characteristics of pages – the frequency of visits, the time the user stays on the page, the length of the path between pages. The result of clustering the pages of a web resource makes it possible to identify groups of pages that are close in content and apply this information to organize managed extraction of data from web pages, to improve the convenience of obtaining search results. To provide personal recommendations to users, one can use separate linked segments of the web resource. Elimination of the found inconsistency in the organization of the structure will increase the speed of indexing a web resource by search engines. Also, the results of the analysis can be used to check the integrity and reliability of content, analyze the structure of the components of a web resource.

However, general approaches to automate the analysis of the structure of a web resource have not been sufficiently investigated, which predetermines the relevance of theoretical and practical research on this topic.

2. Literature review and problem statement

It is known that most modern websites are dynamic in nature, so linking data models with templates for visualization is used to display them. Data is usually stored in structured data warehouses, and its display to the user is created by combining data with some structures – templates. In general, research and comparison of web resource structures with each other is widely used when searching for inauthentic duplicate sites [1], to build search engines, to improve the comprehensibility of the location of pages and site navigation.

Most web pages on the Internet are represented in Hyper Text Markup Language 5 (HTML5) format. The basis of HTML are the tags with which the markup of documents is carried out, HTML tags make it possible to represent the document in a structured form. Based on HTML pages, the browser builds a structural model of the site – Document Object Model (DOM). This model is a tree, in the vertices of which references and attributes of the element are stored, and the descendants of such an element are HTML tags nested in it and texts with information (content).

There are several options for the structure of the DOM site. The first model is represented by a list, the elements of which are expressions for determining the place of the data attribute in the structure of the DOM tree, the HTML tag class for semantic identification of the data attribute, as well as the content of the HTML tag [2]. This model is most often used to analyze a single site in order to streamline its structure and content. However, for analyzing pages from different web resources, it is not effective.

A second model depicts a web page as a link set that includes a path through HTML tags [3]. The disadvantage of such a model is ignoring meaningful information that may be contained in the style classes of a particular web page but the model is effective for analyzing the structure of a web page.

A third group of models includes the model of a web resource in the form of a tree [4] where each element is indexed in the order of its "appearance" in the DOM. This model makes it possible to compare web resources with each other as tree structures. Constructed trees can be "aligned" by sequentially changing the elements. This model is the most successful for cluster analysis of web pages since it makes it possible to compare their structure using metrics.

The concept of stylistic similarity of web resources is widely used in the development of search engines to provide search results relevant to user requests. Most often, web documents are depicted in a vector space as vectors with weights of individual keywords [5]. However, the application of this approach requires the construction of a general dictionary and is not effective.

The analysis of approaches to modeling web resources revealed that for a qualitative analysis of their structure, the model must take into account information not only about the structure of the page but also about the style of its design.

To determine the similarity of web pages with each other, different methods are used. The easiest way to determine the distance between two Web documents is based on the analysis of statistics on the appearance of relevant tags in documents [6]. It is based on the hypothesis that the frequency of appearance of tags reflects some characteristics of web documents and correlates with its structure. To compare page similarities, one calculates the mean square variance in the frequency of occurrence of each of the tags in the document. To improve the results of comparisons, it is proposed to assign each of the tags a corresponding weighting factor. A significant disadvantage of this approach is that it does not take into account the sequence of tags appearing in documents.

A second approach is based on determining the similarities of web resources based on the analysis of their keywords. Using a categorical matrix [7], matches are searched.

Most often, a combination of several methods is used to analyze the similarity of web resources [1]. Clustering methods are widely used to analyze the structure since they make it possible, according to a certain criterion, to distribute pages into clusters for further analysis of their content. Among the clustering methods, the method of k-means, k-medians, and others are most often used but, when solving individual problems, the question arises of choosing cluster centers and metrics that determine the proximity of objects to cluster centers.

To assess the quality and accuracy of clustering, metrics are used that assess the degree of correspondence between expert breakdown and partitioning obtained using calculations, and the homogeneity and completeness of clusters are analyzed. Principles such as Rag Bag and Size vs Quantity also apply.

A review of literary sources [1–7] revealed that to analyze the structure of a web resource, it is expedient to build its mathematical model, which takes into account information about the structure and style of page design. It is also necessary to devise a methodology for comparing web pages using clustering and analyzing the content of the resulting clusters. To assess the quality of partitioning, it is necessary to apply metrics that are able to take into account the characteristics of the objects under study. The approach to be devised will make it possible to get groups of pages while further analysis of their composition will help formulate recommendations for improving the structure of the web resource.

3. The aim and objectives of the study

The aim of this study is to devise a methodology for analyzing the structure of websites using clustering methods and using information about the structure and content of pages. This will make it possible to assess the clarity of the structure of the resource to search for the necessary information, improve the indexing of the resource by search engines, formulate recommendations for reengineering.

To accomplish the aim, the following tasks have been set:

– to build a model of a web resource in the form of an object model and suggest a way to take into account the style of its pages in the selected information model;

– to devise a methodology for determining the structural and stylistic similarity of web pages by applying appropriate metrics for comparison, to carry out clustering of web pages, to build an algorithm and software implementation of the proposed methodology;

– to select the appropriate metrics to assess the quality of the resulting partitioning into clusters;

– to apply the proposed approach to existing web resources and analyze the results of the computational experiment and the effectiveness of the proposed methodology.

4. The study materials and methods

The object of our research is a web resource, which consists of web pages. It is believed that web pages have a certain structure, style, content, and can be divided into certain groups taking into account this information. The task of analysis of the structure of the web resource is stated as follows: it is required to cluster the pages of the web resource, taking into account the structural model of the pages, the content of the information located on the pages of the web resource, and the style of page design.

A DOM model is used to depict the structure of a web resource, which converts the structure of the corresponding HTML document and its contents into an object model – the DOM tree. Information about the styling of web pages is added to the nodes of the DOM tree.

To determine the similarity of web pages, a generalized criterion has been applied that takes into account the structural and stylistic similarity of pages with the corresponding coefficients. Structural similarity is calculated using the metric of distance between trees. To determine the distance between trees, trees are aligned using the operations of renaming, removing, and adding tree nodes.

The similarity of web page design by styles is determined using the Jaccard metric. To determine the pages close to each other, the k-means method of clustering with a cosine measure of distance is used; the number of clusters is determined by the Elbow method; the analysis of elements in the middle of the cluster is carried out using a modification of the Zhang-Shasha algorithm. To assess the quality of the constructed partitioning into clusters, the metrics of accuracy, completeness were used, the comparison took place using expert partitioning. The proposed approach was implemented in the form of an algorithm and software using the Python programming language and the corresponding libraries. A computational experiment was performed to analyze the structure of individual websites that exist on the Internet, as well as for groups of pages from different web resources. The structure of the formed clusters was analyzed, the metric characteristics, the standard similarity of elements in the middle of clusters were calculated, the influence of algorithm parameters on the result of clustering was investigated.

5. Results of devising a procedure for analyzing the similarity of web resources

5. 1. Building a mathematical model of a web resource

Each web page of a modern website is an HTML document that is depicted by a set of HTML tags using a special language. Tags separate the structural elements of the page content, for example, title, text, images. To set the structure, the markup language of the HTML document is used. In addition to setting the structural elements of web resource pages, modern sites use different styles to design their pages. CSS is used to describe styles. CSS styles are rendered in a separate file and linked to the HTML page through the corresponding <link> element. Given that this information is known for each site, it is proposed to use it to analyze the similarity of web resources.

To depict the structure of a web resource, we use the DOM model , which converts the structure of the corresponding HTML document and its contents into an object model [8]. The DOM model is depicted as a tree: the root <*html*> is the title of the HTML document (link to the site); the left subtree <*head*> stores meta tags for browsers and search engines, document name, scripts, and styles; right subtree <body> stores the content of the web page (text, images, media files), that is, the information that is displayed in the browser window. The <body> tag and its child elements can be handled by different styles, which are specified by the *id* and *class* attributes and specified for each tag within the <*body*>. Using these attributes or the name of the tag itself, one can assign individual styles to elements using cascading style sheets. Note that the text in the middle of the tag does not contain child tags, it forms text nodes, and is located on the lower level of the tree. The DOM supports object-oriented representation of a web page.

The use of such a model of representation of a web resource makes it possible to compare web resources with each other as tree structures. To do this, the procedure for "aligning" trees will be used, which involves changing their elements in such a way as to bring them to the same appearance, and counting the number of operations required for this. To carry out the "alignment" procedure, the elements of the tree are indexed. Indexing begins with the elements at the lowest level and goes to the root element <*html*> from left to right. When indexing, all meta tags that are irrelevant to the comparison of two pages are omitted. Hereafter, our work uses the DOM in the form of a tree with indexed elements, the use of which makes it possible to determine the similarity of web resources and web resource pages with each other.

In addition to the structural elements of a web page, it is proposed to take into account design styles as its characteristics. The classes used to style web pages are located in CSS files (Cascading Style Sheets). This data is also important information for determining the similarity of web pages. One can add style information to DOM elements using the *style* attribute or *class* attribute.

The constructed mathematical model will make it possible to determine the similarity of web pages in structure and style.

5. 2. Procedure for determining the similarity of web pages. Algorithm and software implementation

To determine the similarity of web pages, it is proposed to apply an approach that takes into account indicators of structural and stylistic similarity [9]. The general similarity of the two web pages is determined by combining the results of the analysis of structural S_{struct} and stylistic S_{style}similarities:

$$S = p \cdot S_{struct} + (1 - p) \cdot S_{style}, \tag{1}$$

where p is a constant that indicates the weight of structural similarity in a combination of criteria.

The selected mathematical model for representing web resources in the form of DOM trees makes it possible to compare their structures with each other. To compare the structures of two trees, the method of tree alignment using the concept of tree editing distance (Tree Edit Distance) as a metric will be used [10].

The distance between trees T_1 and T_2 is equal to the number of operations that must be performed to convert tree T_1 to tree T_2. The set of operations that can be performed on a tree is denoted by $A(T)$. Node operations Remove, Remote, and Update are used.

If the operations are assigned a cost, then the distance is calculated as the sum of the costs of performing operations, taking into account the cost of individual operations. The cost of the operation $\alpha \in A(T)$ is denoted by $cost(\alpha)$, then the distance between the trees T_1 and T_2 is calculated as follows:

$$TED(T_1, T_2) = \begin{cases} 0, T_1 = T_2 \\ \min_{\alpha \in A(T)} \left\{ cost(a) + TED(a(T_1), T_2) \right\}. \end{cases} \tag{2}$$

Function (2) returns the value of "cost" as a real positive number. Since the axioms of the metric are executed for the *TED* function, a metric space of trees is formed that can be compared with each other.

It is considered that the cost of each of the operations is constant, in addition, only one operation is performed during one conversion.

The value of "cost" is associated with the sequence of operations performed to change the tree, necessary to convert it from its initial state to "aligned", that is, from T_1 to T_2. Since the number of tree edits is not limited, the "cost" is normalized by setting the maximum criterion, and the structural similarity of the two DOM trees is calculated as follows:

$$S_{struct}=1-\frac{TED(T_1,T_2)}{y\_\max(|T_1|+|T_2|)}, \tag{3}$$

where y_max is the maximum number of y_remote, y_remove, and y_update operations.

In addition to the structural similarity (3) between web resources, there is also a stylistic similarity. Classes used to style web pages are present in CSS files (Cascading Style Sheets). This data is also important information for determining the similarity of web pages. Styles can be added to DOM elements using style or class attributes, so the proposed web resource representation model makes it possible to analyze the similarity of styles. A set of style class names can be obtained using the DOM API and xpath() function.

Since the set of styles used is limited, the Jaccard metric can be used to determine the stylistic similarity. Jaccard's similarity coefficient based on styles is calculated by defining some of the style names common to web pages D_1 and D_2.

$$S_{style}=\frac{|A\cap B|}{|A|+|B|-|A\cap B|}, \tag{4}$$

where $A=classes(D_1)$, B=classes(D_2), *A, B* are sets of D_1 and D_2 page style classes, respectively.

The disadvantage of the selected metric for determining the measure of stylistic similarity (4) is that certain templates can be used when developing pages of one web resource. If this is true, then the similarity indicators may approach unity.

To cluster web pages using information about their structural and stylistic image, the following combination of methods is proposed. In the beginning, one needs to get the DOM of web pages in the form of text strings, each of which was converted into a vector using the Tf-Idf metric. This metric is a statistical measure used to assess the importance of each element in the context of a document. To analyze strings, the weight of a certain element is considered to be proportional to the frequency of occurrence of this element in a string and inversely proportional to the frequency of use of this element in all strings being considered.

To cluster the resulting DOM web pages, the k-means method with a cosine distance measure will be used, and the Elbow method is chosen to determine the number of clusters. The choice of the k-means method is due to the speed of clustering and the possibilities for processing significant amounts of data. Given that the data to be clustered has no emissions, the disadvantage of the method associated with emission sensitivity does not limit, in this case, the possibility of application.

Further, in the middle of each cluster, the analysis of elements using the Zhang-Shasha algorithm will be applied [10], the elements of each are compared with each without repetitions. The specified algorithm works with indexed trees, which are DOM trees in structure. The application of the principles of dynamic programming makes it possible to effectively calculate the distances between the root nodes of trees. To group such web pages according to the indicator of structural similarity, a modification of the AP-TED algorithm was chosen, the use of which reduces the computational complexity of the algorithm to the linear one. The number of tree editing distances is recursive and involves calculating the distance between trees according to the editing strategy "along all routes", starting with the lower elements of the tree, followed by the rejection of unnecessary strategies. This approach ensures effective utilization of RAM.

The proposed methodology is implemented by the following algorithm:

0. Acquire input information about web resources: build a list of DOM web pages using links to each of them, carry out vectorization of the obtained DOM using the TF-IDF method, determine the set of styles of web pages.

1. Determine the number of clusters using the Elbow method.

2. Cluster the elements of the DOM list of web pages using the k-means method.

3. Apply the Zhang-Shasha algorithm to analyze the content of each of the clusters by comparing the cluster elements of each with each without repetition.

For the practical implementation of the proposed approach, the Python programming language, special libraries and methods were used. Lxml and urllib libraries were used to process HTML files, access site URLs, pricing web pages, the Pandas library – to implement data cleaning methods, the difflib module was used to find and process differences between sequences, the sklearn library – to implement methods clustering and calculating the TF-IDF metric.

The software consists of a module for reading the DOM of a web page by a link to it and a module for implementing algorithms for comparing two web pages by its DOM and style classes. The module for reading the DOM of a web page is implemented using the Urllib library, which makes it possible to get the DOM in the form of strings. To implement the Zhang-Shasha structural similarity algorithm, the parse() method was used, with the help of which only a set of HTML tags of the compared pages was obtained. Tag sequences are compared using the difflib library, the ratio() function makes it possible to determine the percentage of convergence between two sequences. To calculate the Jaccard metric and obtain a set of style classes, each of the compared web pages uses the get_classes() method, which is implemented using the Parsel library using the regular expression doc.xpath.

5. 3. Selecting a metric to assess the quality of partitioning

Since the proposed combination of methods make it possible to obtain an approximate solution to the clustering problem, in this work a comparison of the obtained results of the breakdown with expert breakdown was used to assess the quality of the result. Expert partitioning was built on the basis of the structure of sites, clusters corresponded to certain categories of pages depending on the thematic focus of the site. Standard metrics for this class of problems for assessing accuracy and completeness were used:

$$P=\frac{T}{T+F},\quad R=\frac{T}{T+FN},$$

where *T* is the number of expert partitioning clusters separated by the algorithm; *F* – the number of clusters that are absent in the expert breakdown but separated by the algorithm; *FN* – the number of expert partitioning clusters that are absent in the partition, which is obtained using the algorithm. We assume that the cluster C_k of expert partitioning is found by the algorithm if it consists of more than half of the vertices of the C_k cluster and less than half of the vertices of any of the other expert partitioning clusters.

According to similar formulas, accuracy and completeness in the middle of the cluster are calculated. The value T is equal to the maximum number of vertices of the expert cluster, which are allocated by the algorithm into a separate cluster during partitioning; F and FN are equal to the number of vertices added and excluded by the algorithm from expert partitioning, respectively.

5. 4. Results of computational experiment

Analysis of the similarity of web pages using the proposed approach was carried out for two types of tasks:

– definition of the subject to which the cluster with web pages belongs;

– analysis of the structure of a separate website using clustering of its web pages by type – product page, categories, news, blog, etc.

To implement the task of the first type, several different web resources were selected so that they intuitively differ in subject matter but some of them formed groups (clusters) with similar information. Web resources were selected that belong to:

1) banking operations:

– https://www.privatebanking.hsbc.com/investment-services/alternative-investments/hedge-funds/;

– https://bank.gov.ua/;

– https://my.ukrsibbank.com/ru/personal/;

– https://www.privatebanking.hsbc.com/investment-services/alternative-investments/hedge-funds/;

2) educational activities:

– https://beetroot.academy/en/courses/online/front-end;

– https://www.classcentral.com/course/html-css-javascript-for-web-developers-4270;

3) sports:

– https://football.ua/;

– https://www.ua-football.com/;

– https://sport.ua/football;

4) mathematical information:

– https://en.wikipedia.org/wiki/Machine_learning;

– https://en.wikipedia.org/wiki/Mathematical_optimization;

– http://en.wikipedia.org/wiki/Training,_validation,_and_test_sets.

To solve the problem of the second type, the site for the sale of seeds http://semena-dnepr.org.ua and the website of the faculty at a higher education establishment http://fpm.dnu.dp.ua were considered. The website of the online store consists of the main page, pages of categories and subcategories of goods, product cards, blog pages. For the site of a higher education institution, pages are selected that display general information about the faculty, news, information about departments, information for applicants.

For the problem of the first type, Table 1 shows the results of the division into clusters: the composition of the cluster, the number of elements and the result of calculating the similarity of web resources within the cluster. Tables 2, 3 give the results of partitioning for the pages of the site http://fpm.dnu.dp.ua, obtained for different predetermined number of clusters; Table 2 – for 3 clusters, Table 3 – for 4 clusters.

Fig. 1 shows the dependence of the sum of intracluster distances on the number of clusters for the problem of the first type a and the second type b.

A rectangle indicates the area in which a characteristic curve bend is formed (elbow point), which determines the optimal number of clusters in the breakdown.

Table 1

Results of splitting groups of pages from different websites into clusters

Cluster	Cluster composition	Number of elements	Similarity within the cluster
1	https://www.privatebanking.hsbc.com/investment-services/alternative-investments/hedge-funds/; https://bank.gov.ua/; https://my.ukrsibbank.com/ru/personal/; https://www.privatebanking.hsbc.com/investment-services/alternative-investments/hedge-funds/	4	0.0163
2	https://beetroot.academy/en/courses/online/front-end; https://www.classcentral.com/course/html-css-javascript-for-web-developers-4270	2	0.164
3	https://football.ua/; https://www.ua-football.com/; https://sport.ua/football	3	0.017

4	https://en.wikipedia.org/wiki/Machine_learning;
	https://en.wikipedia.org/wiki/Mathematical_optimization;
	http://en.wikipedia.org/wiki/Training,_validation,_and_test_sets
		3	0.3216

Table 2

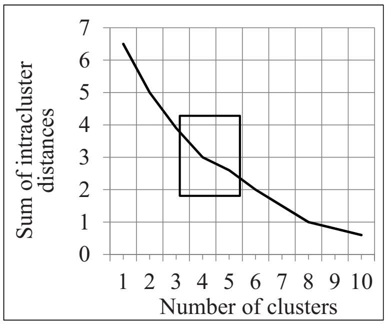
Results of splitting pages of website http://fpm.dnu.dp.ua/ into 3 clusters

Cluster	Cluster composition	Number of elements	Similarity within the cluster
1	http://fpm.dnu.dp.ua/; http://fpm.dnu.dp.ua/category/news/; http://fpm.dnu.dp.ua/category/analytics/	3	0.6819
2	http://fpm.dnu.dp.ua/fakultet/kafedri/kafedra-kompyuternix-texnologij/; http://fpm.dnu.dp.ua/fakultet/kafedri/kafedra-obchislyuvalnoi-matematiki-ta-matematichnoi-kibernetiki/; http://fpm.dnu.dp.ua/fakultet/kafedri/kafedra-matematichnogo-zabezpechennya-eom/	3	0. 6885
3	http://fpm.dnu.dp.ua/abiturientam/vstup-do-bakalavratu/; http://fpm.dnu.dp.ua/abiturientam/vstup-do-aspirantury/; http://fpm.dnu.dp.ua/abiturientam/vstup-do-magistraturi/	3	0.8177

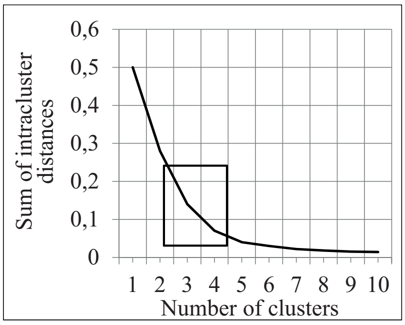
Table 3

Table name Results of splitting pages of website http://fpm.dnu.dp.ua/ into 4 clusters

Cluster	Cluster composition	Number of elements	Similarity within the cluster
1	http://fpm.dnu.dp.ua/category/news/; http://fpm.dnu.dp.ua/category/analytics/	3	0.9472
2	http://fpm.dnu.dp.ua/fakultet/kafedri/kafedra-kompyuternix-texnologij/; http://fpm.dnu.dp.ua/fakultet/kafedri/kafedra-obchislyuvalnoi-matematiki-ta-matematichnoi-kibernetiki/; http://fpm.dnu.dp.ua/fakultet/kafedri/kafedra-matematichnogo-zabezpechennya-eom/	3	0.6885
3	http://fpm.dnu.dp.ua/abiturientam/vstup-do-bakalavratu/; http://fpm.dnu.dp.ua/abiturientam/vstup-do-aspirantury/; http://fpm.dnu.dp.ua/abiturientam/vstup-do-magistraturi/	3	0.8177
4	http://fpm.dnu.dp.ua/ - main page		1



a



b

Fig. 1. Selection of the optimal number of clusters using the Elbow method: *a* — for pages from different web resources; *b* — for pages from a separate web resource

6. Discussion of results of clustering the pages of a web resource

According to the results of the computational experiment (Table 1), 4 clusters were formed for the problem of the first type, which group web pages by topic.

The arithmetic mean similarity between the elements within the cluster for clusters 1–3 is quite low since the selected pages belong to different sites and cannot have a similar structure and style. The value of this metric for cluster No. 4 is higher because all web pages from cluster 4 belong to the same https://en.wikipedia.org/ website.

Using the Elbow method, the number of clusters for the k-means method is determined. The characteristic inflection of the dependence curve, after the formation of which the values of the objective function begin to decrease more slowly, determines the number of clusters in the partition. The dependence of the objective function on the number of clusters (Fig. 1, *a*), which is built for the problem of the first type, indicates the impossibility of determining the inflection point of the curve and accurately determining the number of clusters. This behavior of dependence is caused by the low similarity of the structure of the pages of web resources in question.

The problem of the second type is considered on the example of the website of the faculty at a higher education establishment http://fpm.dnu.dp.ua. The result of applying the Elbow method in this case is shown in Fig. 1, *b*. The dependence has a characteristic inflection point and defines 3 clusters in the partition. From the analysis of Table 2, one can see that the arithmetic mean similarity between the elements within the clusters is high. It acquires the greatest importance for a group of pages, regulating admission to the specialties of the faculty and, indeed, are similar in structure, content, and style of design. Lower page similarity rates are observed for a cluster with pages about structural units of the faculty, which is explained by the presence in their DOM of HTML tags for individual design elements, for example, additional panels that are not used on all pages.

When splitting the same set of pages into 4 clusters (Table 3), the best results of the similarity of elements within the clusters were obtained. Compared to the previous partition, cluster No. 4 was separated, which contains only the main page of the website. Due to this, cluster No. 1 became more homogeneous, the similarity rate between the elements of the cluster approached 95 %.

The application of the proposed approach to the website of the online store http://semena-dnepr.org.ua made it possible to obtain satisfactory results of partitioning: clusters consisting of pages of product categories, pages with individual products, news pages, the main page of the site are separated. The arithmetic mean similarity between the elements within each of the clusters is high and is equal to 95–98 %, which means the same identical structure of pages that end up in the same cluster. During the development of websites of online stores, most often used are templates that dynamically form the same structure pages. The content of the pages is stored in the database and changes in different HTML tags when you go to a specific link that points to a product or category.

It should be noted that to demonstrate the proposed approach, examples provide a small number of web resource pages. However, the computational experiment was conducted for all pages of the http://semena-dnepr.org.ua site, the number of which was about 500, and the site http://fpm.dnu.dp.ua with a total number of pages of 830.

The result of clustering was compared to expert breakdown, which was formed on the basis of the structure of the online store. The expert breakdown consisted of clusters with product category pages, product card pages, blog pages, and news pages. The result of clustering using the proposed approach was evaluated in relation to the proposed expert breakdown using accuracy and completeness metrics. According to the results of the algorithm, the value of clustering accuracy was achieved – 0.89, the completeness value – 0.83. Analysis of the obtained estimates showed that lower accuracy and completeness values were obtained for a large cluster, which consists of product cards. This is due to the getting into it of some news pages. It should be noted that some product cards, on the contrary, ended up in a cluster with blog pages.

The devised mathematical model and procedure for comparing web pages by structure and style made it possible to obtain acceptable results of partitioning the pages of a web resource. Isolated clusters overwhelmingly contain similar pages. The presence of pages that differ from most elements of the cluster indicates a violation of the structure of the web resource. The found discrepancy should be eliminated to improve the structure of the site. The proposed approach makes it possible to find the inconsistency of pages with clusters, analyze the causes of its appearance, and get recommendations for eliminating inconsistencies during the reengineering of a web resource.

The limitations of the approach include the fact that not all web pages grant permission to read their structure, some of them have closed access to their DOM. This situation should be treated as an exception, followed by warning the user about the lack of access to reading the DOM page at its link.

The disadvantages of the proposed approach include the fact that the results of clustering significantly depend on the use of templates when building site pages. In the case when the development is carried out in this way, the indicator of stylistic similarity is high and has an impact on the determination of the overall similarity of pages. To eliminate this problem in the future, it is necessary to develop a procedure for determining the coefficients for combining indicators of structural and stylistic similarity. Further research should also address the development of algorithms for calculating the distance between DOM trees, which would have less computational complexity and be more efficient. It is interesting to use new computing technologies – neural networks to compare the structure of web resources.

7. Conclusions

1. The web resource model has built in the form of a DOM tree that stores information about the structure of a web page and consists of a root with a link to the site, a left subtree with meta tags for browsers and search engines, and a right subtree with web page content. To the elements of the tree, we added information about the design styles of individual elements. This model has made it possible to depict web pages as discrete structures – trees, and compare them with each other by alignment.

2. To determine the similarity of web resource pages, a generalized criterion has been applied that takes into account the structural and stylistic similarity of pages with the corresponding weighting factors. To compare the structures of two DOM trees, a tree alignment method will be used, which is based on the transformation of one tree to the appearance of another tree. The editing distance is used as a metric, and the operations of renaming a node, deleting a node, and adding a tree node are used as editing operations. To determine the similarity of web pages by style, the Jaccard metric is used. To cluster the resulting DOM web pages, the k-means method with a cosine distance measure is used, and the Elbow method is selected to determine the number of clusters. Intracluster analysis was carried out using a modification of the Zhang-Shasha algorithm. The application of the generalized criterion and the combination of a number of methods have made it possible to cluster web pages and analyze the content of the formed clusters. The result of the cluster analysis makes it possible to identify individual pages of the web resource that need to be configured to maintain the logical structure of the web resource. The proposed approach is implemented in the form of an algorithm and software using the Python programming language and the corresponding libraries, which makes it possible to automatically carry out the procedure for analyzing the structure of a web resource.

3. To assess the quality of partitioning, accuracy and completeness metrics were used, which can be applied in the presence of expert partitioning of web resource pages into clusters. High rates of accuracy and completeness are obtained for web resources, in the middle of which a structured preservation of web pages is organized in accordance with the catalog, which contributes to the accurate separation of clusters. A decrease in accuracy and completeness indicators indicates a violation in the structure of the web resource and the need to perform a site reengineering procedure.

4. The proposed approach is applied to analyze groups of pages from different sites and the structure of individual websites existing on the Internet. The result of clustering of pages of a separate web resource in relation to expert partitioning is evaluated, the values of accuracy and completeness metrics are calculated. It has been established that for pages from different sites it is possible to group them by topic since the average arithmetic similarity between the elements in the middle of the cluster is quite low and does not exceed 10–15 %. For individual websites created with templates, page splitting is obtained, which corresponds to the structure of the site, the resulting groups of pages have similar properties, style, and content. The arithmetic mean similarity between elements within clusters for such web resources is high and reaches 70–90 %. A decrease in accuracy and completeness indicators indicates the presence in the cluster of elements that differ in structure and style from the target elements and require transferring them to other sections of the web resource to maintain its logical structure.

Conflict of interests

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

References

1. Jain, A., Gupta, B. B. (2017). Phishing Detection: Analysis of Visual Similarity Based Approaches. Security and Communication Networks. doi: https://doi.org/10.1155/2017/5421046

2. Vdovin, I. V., Ovchinnikova, R. Y. (2018). Data extraction from the internet network with the use of structural-semantic clustering of web pages. Dynamics of Systems, Mechanisms and Machines (Dynamics), 6 (4), 106–113. doi: https://doi.org/10.25206/2310-9793-2018-6-4-106-113
3. Feng, J., Qiao, Y., Ye, O., Zhang, Y. (2022). Detecting phishing webpages via homology analysis of webpage structure. PeerJ Computer Science, 8, e868. doi: https://doi.org/10.7717/peerj-cs.868
4. Grigera, J., Gardey, J., Garrido, A., Rossi, G. (2021). A Scoring Map Algorithm for Automatically Detecting Structural Similarity of DOM Elements. Proceedings of the 17th International Conference on Web Information Systems and Technologies. doi: https://doi.org/10.5220/0010716300003058
5. Wu, H., Yuan, N. (2018). An Improved TF-IDF algorithm based on word frequency distribution information and category distribution information. Proceedings of the 3rd International Conference on Intelligent Information Processing. doi: https://doi.org/10.1145/3232116.3232152
6. Bozkir, A., Sezer, E. (2018). Layout-based computation of web page similarity ranks. International Journal of Human-Computer Studies, 110, 95–114. doi: https://doi.org/10.1016/j.ijhcs.2017.10.008
7. Moreno, V., Génova, G., Alejandres, M., Fraga, A. (2020). Automatic Classification of Web Images as UML Static Diagrams Using Machine Learning Techniques. Applied Sciences, 10 (7), 2406. doi: https://doi.org/10.3390/app10072406
8. Shin, K., Ishikawa, T., Liu, Y.-L., Shepard, D. L. (2021). Learning DOM Trees of Web Pages by Subpath Kernel and Detecting Fake e-Commerce Sites. Machine Learning and Knowledge Extraction, 3 (1), 95–122. doi: https://doi.org/10.3390/make3010006
9. Gowda, T., Mattmann, C. A. (2016). Clustering Web Pages Based on Structure and Style Similarity (Application Paper). IEEE 17th International Conference on Information Reuse and Integration (IRI). doi: https://doi.org/10.1109/IRI.2016.30
10. Zhang, K., Shasha, D. (1989). Simple fast algorithms for the editing distance between trees and related problems. SIAM Journal on Computing, 18 (6), 1245–1262. doi: https://doi.org/10.1137/0218082