This study explores websites such as online stores, which are considered to be a set of interconnected web pages. The task addressed relates to the high computational complexity of manual analysis of the topology of modern websites, as well as the lack of formalized mechanisms that could make it possible to integrate the semantic features of web pages into the process of automated hyperlink reconstruction.

Within the framework of this study, a website is crawled in order to obtain complete HTML documents, from which the structural features of pages are extracted (the number of headings, depth of embedding, presence of <article>, number of incoming links, etc.). The resulting vectors make it possible to construct cosine similarity matrices to assess the mutual proximity of pages. An approach has been proposed to rebuilding the link structure of the website taking into account this similarity; a comparison of the initial and transformed website was carried out using the metric characteristics of modularity, clustering, diameter, and similarity distribution.

The results demonstrate that taking into account the DOM structure allows for the formation of a logical, reasonable distribution of pages between clusters. And the subsequent automatic procedure for setting hyperlinks makes it possible to improve structural integrity by establishing effective relationships between thematically close pages.

The practical significance of this work involves the possibility of using the proposed approach for automated optimization of internal links of static websites. As a result, the architecture of the web resource is improved, website navigation becomes transparent, and website indexing by search engines is increased

Keywords: DOM model, web graph clustering, page similarity, structure optimization, relinking, cosine distance

UDC 004.912

DOI: 10.15587/1729-4061.2026.357377

DEVISING AN APPROACH TO ANALYZE AND AUTOMATICALLY RECONFIGURE THE STRUCTURE OF WEBSITES

Ivan Dolotov
Corresponding author
PhD Student*
E-mail: vanyadolotov@gmail.com
ORCID: https://orcid.org/0000-0002-4643-3464
Natalia Guk
Doctor of Physical and Mathematical Sciences, Professor,
Vice-Rector for Academic and Pedagogical Affairs*
ORCID: https://orcid.org/0000-0001-7937-1039
*Department of Computer Technologies
Oles Honchar Dnipro National University
Nauky ave., 72, Dnipro, Ukraine, 49000

Received 27.01.2026
Received in revised form 06.04.2026
Accepted date 13.04.2026
Published date 30.04.2026

How to Cite: *Dolotov, I., Guk, N. (2026). Devising an approach to analyze and automatically reconfigure the structure of websites. Eastern-European Journal of Enterprise Technologies, 2 (2 (140)), 43–51. https://doi.org/10.15587/1729-4061.2026.357377*

1. Introduction

In today's rapidly growing Internet environment, there is a constant need to automate the process of transforming web pages into structured knowledge. In the modern web space, the increasing complexity, dynamism, and semantic richness of websites require devising effective methods for their formal analysis, structuring, and automated processing.

The basis of any website is HTML documents that define the content and structure of web pages and are used to represent information in the web space. Each HTML document contains a set of elements to define the logical hierarchy of content and the similarity of web pages.

The stylistic similarity of web resources is an important aspect of their structure and design, which affects the perception of content by users, as well as navigation and the effectiveness of interaction with the website. Similarity can manifest itself at different levels, in particular in visual design, the location of interface elements, the principles of hyperlink organization, as well as the use of the same templates or stylistic components. Investigating such characteristics is important for analyzing the relationships among websites. At the same time, improving the user experience and the efficiency of website indexing by search engines through optimization and improvement of the hyperlink structure is becoming a pressing task.

2. Literature review and problem statement

One of the models of website representation is the web graph reported in [1, 2]. In them, the structure of the website is considered as a directed graph, in which the available information is represented in a hierarchy. In the case of a website representation in the form of a directed graph, web pages are considered as vertices of the graph, and hyperlinks as edges. This model provides mathematical tools for analyzing relationships within websites, but it lacks information about the content of web pages, so it is impossible to assess the thematic similarity of pages without manual analysis. The authors of [1] focused on ways to represent web resources and assess their structure, as well as describing the advantages of graph models. In [2], a crawling procedure was devised to collect information about a website and structurally compare websites of different types. However, the authors did not propose methods for optimizing the structure.

Another means of representing a web page is the DOM (Document Object Model), which is reported in [3, 4].

This model formalizes the content of HTML documents in the form of a hierarchical tree of objects. Unlike web graphs, which do not store content information, the DOM allows for detailed thematic analysis of content at the level of individual elements. The use of DOM is particularly important for analyzing, processing, and manipulating web pages. It allows one to read and modify HTML tags, interact with the structure of web documents, and apply clustering algorithms or pattern search in web resources. In this model, each HTML element is represented as a node, element attributes as attribute nodes, and text fragments inside tags as text nodes. In [3], the authors, based on these features, tried to extract relevant data from the resulting pages using their structural markup. Based on the data, changes were made to the interface and structure of the website. However, the algorithm is very sensitive to any changes in logical blocks, the addition of new tags, and page extensions, which leads to its looping and the need to adapt to each case.

In [4], an algorithm was developed that makes it possible to group pages into clusters for studying user navigation or reengineering web resources. The page is depicted as a DOM tree, and then the structural nodes are supplemented with information about cascading CSS style sheets. The method works well on small information resources but for large web resources and online stores, it is not suitable due to the presence of identical page templates and the need to analyze large DOM trees using similarity methods based on brute force.

In [5], the concept of DOM trees was introduced. A universal mathematical algorithm WISDOM was proposed. Before its implementation, it was necessary to create manual templates for each website separately. The essence of the algorithm is to select and analyze only meaningful content, while ignoring «noise». The logic of the algorithm is based on traversing the DOM tree, each node is checked using heuristic rules. That work became fundamental in the field of research specifically on DOM trees. However, the algorithm loses accuracy when used on modern large HTML documents.

The DOM tree is constructed separately for each HTML file. This property provides high structural detail and makes it possible to accurately localize content related to a specific topic or function of the page. Based on these properties, in [6], the DOM is used for the task of semantic information extraction for further classification of pages. The authors emphasize that in the process of scaling or expanding web resources, the primary logic of building a hypertext structure is often violated. In [6], an analysis of web resources based on DOM structures was carried out, after which the website structure is built from scratch. However, in this case, the website structure is formed without taking into account the principles of logical navigation.

One of the key approaches to web page comparison is the structural analysis of the DOM, which is based on the representation of HTML documents as hierarchical trees. In [7], methods of segmentation and construction of DOM blocks are considered, taking into account structural cohesion. A method of selecting subtrees with a similar tag structure is proposed to improve web page segmentation. However, the approach requires a thorough analysis of the visual and attributive characteristics of elements. The issue of scaling and limited performance when applied to large web resources is not solved.

An option to overcome these difficulties may be the tree similarity metrics proposed in [8]. The generalized editing distance method estimates the number of basic operations required to transform one tree into another. In fact, the similarity of DOM trees is estimated based on the distance between them. The algorithm is much faster than segmentation methods. However, it is too slow for modern websites due to the growing volume of modern web resources.

In [9, 10], methods that focus on the textual content of pages are proposed. In [9], the Jaccard similarity method is described, in which the similarity between pages is determined based on the intersection of sets of words. In [10], the authors described the Levenstein distance method, which calculates the minimum number of changes required to convert one text to another. These approaches demonstrate good efficiency in detecting pages similar in content but completely ignore the nested DOM hierarchy, which significantly reduces the accuracy in thematic classification tasks.

That gives grounds to argue about the feasibility of studying the vectorization of DOM structures, which makes it possible to effectively represent a tree in the form of a numerical vector. Work [11] demonstrates the possibilities of such a representation. The work uses the TF-IDF approach to HTML tags, which considers the DOM as a text document. The authors use the TF-IDF method for clustering and assigning a rank to pages, highlighting the most important ones, by further using this data to rebuild the structure of the website. This approach solves the problem of processing speed but there may be a loss of the context of the nesting of elements. Therefore, existing approaches either demonstrate a high load when processing large amounts of data, or are inferior in accuracy to semantic analysis, which justifies the need for further improvement of combined structural-vector models for analyzing web resources.

Our review of related literature [1–11] revealed unresolved problems related to the need to build a mathematical model of a web resource that takes into account the properties of its structure and the functional purpose of web pages. In the studies of graph models, there are no methods for taking into account the thematic affiliation of pages and the image of their vector representation taking into account semantics. In the case of studies related to DOM structures, the structural features of pages are not taken into account, which could lead to a violation of the logic of the website. There is also a need to algorithmize the procedure for rebuilding hyperlinks in accordance with the detected thematic and structural proximity of pages. Existing metrics do not take into account the semantic features of pages, so they need to be improved to ensure the ability to measure the level of similarity of web pages and assess the cluster structure of a website.

3. The aim and objectives of the study

The aim of our research is to devise an approach for analyzing and automatically rebuilding the structure of websites using information about the structure and content of pages. This will make it possible to assess the quality of the current structure and provide recommendations for rebuilding internal links based on the thematic and structural similarity of pages.

To achieve the goal, the following tasks were set:

– to formalize the representation of a web resource in the form of an object model taking into account the style and thematic properties of pages;

– to devise an approach for determining the structural and style similarity of web pages and perform automatic rebuilding of a set of hyperlinks between pages – page relinking;

– to select appropriate metrics for assessing the quality of rebuilding the website structure;

– to apply the proposed approach to real web resources, as well as analyze the results of a computational experiment in order to assess the effectiveness of its use.

4. The study materials and methods

The object of our study is websites such as online stores, which are considered as a set of interconnected web pages. The subject of the study is the structure of the website, represented in the form of a hypertext and DOM model, the methodology for assessing the similarity of pages, and the procedure for automatic setting of hyperlinks (the relinking procedure).

The hypothesis of the study assumed that by formalizing the internal structure of web pages through a multidimensional vector representation of DOM models and further changing the structure of the web graph using cosine similarity, it is possible to achieve an increase in the level of structural and thematic homogeneity of the web resource.

It is assumed that for the tasks of automatic analysis or setting of hyperlinks, it is advisable to use a graph model recreated on the basis of DOM, where nodes correspond to pages or sections of the website, and edges – hyperlinks between them. The cosine similarity measure between the vectors of page features is a valid indicator for determining the feasibility of establishing a hyperlink between them. It is also assumed that the resulting DOM structure is sufficient to form a stable vector of page features, and that links to and from pages with a large number of links have unconditional structural value, regardless of their semantic similarity.

Accepted simplifications include ignoring the weight of the hyperlink, that is, all hyperlinks are considered to be of equal value, regardless of their location. Also, the work considers only internal links, and external ones are completely excluded from the analysis.

The internal structure of each page is described using the DOM model in the form of a node tree, which reflects the hierarchy of HTML elements. At the first stage, the web resource is crawled, during which the HTML codes of web pages are loaded. The work considered only internal links, and also eliminated unnecessary parameters, fragments and other URL variations that do not affect the content.

A vector representation of the page was formed based on the DOM structure. A modernized approach based on the content mining method was used for vectorization. The vector representation contains structural and semantic features, in particular, the number of incoming/outgoing links, tree depth, number of images, sections, headings, the presence of <article> and <price> tags, page type, and a set of thematic attributes (data-title). Each page is represented as a multidimensional feature vector. Not only the information directly encoded in HTML tags was used but also parameters that describe the type of pages and their thematic affiliation. This enables the preservation of structural information about the page in numerical form.

To determine the degree of similarity between pages, a cosine similarity measure was used. Based on the results of the calculations, a similarity matrix is formed, which is used to analyze the structural proximity of pages and further optimize the set of hyperlinks. The structure is rebuilt by replacing weakly relevant links with links between pages with greater vector similarity, taking into account the threshold value and preserving service pages with a large number of incoming links.

The Python programming language was used for the software implementation of the methodology. Information about the internal structure of the website was collected using the crawling procedure and using the requests, BeautifulSoup, lxml, and urllib.parse libraries. The results of the algorithm were stored in the form of similarity matrices and lists of changed hyperlinks.

To assess the effectiveness of structure reconstruction, a system of metrics was used: average, median, and variance of cosine similarity, modularity coefficient, and clustering coefficient. Statistical indicators characterize the degree of structural homogeneity of pages, while graph metrics make it possible to assess the strength of intra-topical connections and the level of clustering of the web graph before and after relinking.

The computational experiment was conducted for real web resources with different structural organization. Comparative analysis was performed by comparing the values of metrics before and after automatic optimization of internal hyperlinks.

5. Results of devising the methodology for analyzing the similarity of web resources

5. 1. Construction of the formalization of the web resource representation

A website is a set of web pages that are interconnected and united under one domain name. Pages look like separate documents and can be text documents, images, videos, interactive elements, and contain the main content of the website that the user sees when visiting. Each page has its own unique URL and hyperlinks to some other pages of the website, which makes it possible to establish connections between them and provide the user with website navigation.

The hypertext model of the website H is defined as a set consisting of two sets

$$H = \{P, L\}$$

where $P = \{p_1, p_2,...,p_n\}$ is the set of pages of the website; $L = \left\{1 \mid \exists p_1, p_2 \in l\left((p_1, p_2)\right)\right\}$ is the set of hyperlinks between pages.

The structure of the hypertext model of the website corresponds to the model in the form of a directed unweighted graph $G = (N, E)$, in which $N = P$, $E = L$. In the constructed graph, N is the set of vertices, the elements of which describe the pages of the website, E is the set of edges of the graph, the elements of which correspond to hyperlinks between pages.

If there is a constructed web graph, it is possible to analyze whether the pages of the website are correctly connected to each other by hyperlinks and whether the navigation is understandable for the user.

An addition to the hypertext model H is the document model, which describes the internal structure of each web page. In web technologies, such a structure is defined by the DOM – a hierarchical tree of nodes, where each node corresponds to an HTML element or a text fragment of the page.

For the page $p_i \in P$, we construct a DOM-tree, which can be formally represented as follows

$$D_i = \left(W_i, F\right)_i,$$

where $W_i = (w_1, w_2,...,w_i)$ is the set of nodes of the DOM tree, and $F_i \subseteq W_i \times W_i$ is the set of parent-child relationships between nodes that define the hierarchy of elements on the page.

Thus, the hypertext model H is extended by the set of DOM trees

$$H_{DOM} = \left\{\left(p_i, D_i\right) | \, p_i \in P\right\}.$$

In such an extended model, the vertex $p_i \in P$ of the web graph G contains additional information about its internal DOM structure, which make it possible to take into account not only the connections between pages but also the structural characteristics of each page.

Next, the DOM vectors of pages were formed. Each page is depicted as a set of features, in particular, the number of headings, the depth of nesting of structural elements, the presence of <article> tags, the number of , the average number of child elements, etc. are taken into account. These vectors act as input data for calculating the similarity of pages, determining clusters, and building an optimized website model. Thus, the DOM model has become the basis for interpreting the functional structure of a web resource, and the graph model makes it possible to establish and rebuild logical connections between elements of this structure.

The DOM tree transformation vector D is depicted in the following way

$$V_i \begin{pmatrix} L_i^{out}, L_i^{in}, Dep_i, Im_i, Div_i, Dep_i, \\ Pr_i, H_i, Ar_i, Nd_i, P^*, T^* \end{pmatrix},$$

where L_i^{out}, L_i^{in} is the number of outgoing/incoming links from/to page i, Dep_i –maximum depth of the tree, Im_i – number of images, Div_i – number of div sections, Pr_i – presence of the price tag price, H_i – number of headers header, Ar_i – presence of the article title article, Nd_i – number of child nodes of the tree, P^* – page type description vector, T^* – page tag description vector data–title.

To construct a vector representation of a website page based on the DOM structure, a set of features that automatically came from the HTML code was used. The data source is the HTML markup of the page, which is read using crawling. That made it possible to obtain both structural and semantic information about the page elements. In particular, the number of outgoing $\left(L_i^{out}\right)$ and incoming $\left(L_i^{in}\right)$ links is determined, which is calculated based on the adjacency matrix of the constructed web graph. Quantitative tags (Im_i or Div_i), as well as tree characteristic description tags (Dep_i or Nd_i) are calculated based on the available HTML tags in the DOM tree. Additionally, a page type description vector (P^*) is formed based on its functional purpose – category, product card, article, etc. Thus, a page refers to an article if there is an <article> tag, < h1 > + text, a publication date, a block with paragraphs (<p>). V_i also contains a data-title attribute vector (data-val, data-type, etc.), which determines the subject of the page according to the corresponding attributes in HTML. Thus, the constructed V_i vector for the page summarizes both structural and semantic characteristics that can be used for further analysis, clustering, or optimization of the website structure.

5. 2. An approach to determining structural and stylistic similarity of web pages and automatic reconstruction of the set of hyperlinks between pages

After converting the DOM structures of web pages into a vector representation, it becomes possible to formally estimate the similarity between pages. The resulting vectors make it possible to apply classical distance or similarity metrics, in particular cosine similarity. Let V_i and V_j be vector representations of pages p_i and p_j, respectively. Then cosine similarity is defined as follows

$$sim\left(p_i, p_j\right) = \frac{V_i \cdot V_j}{\left\|V_i\right\| \left\|V_j\right\|}. \quad (1)$$

The calculated similarity value for each pair of pages was interpreted as the potential "connection strength" between them in the website structure. Based on the results of the similarity calculation, a similarity matrix $S \in R^{n \times n}$ is formed, where $S_{ij} = sim(p_i, p_j)$. The obtained similarity matrix allowed us to identify pages with a similar structure, even if they do not have a direct hyperlink to each other.

In the case of restructuring an existing website, there is a possibility of deterioration of the link logic compared to the original structure. Developers most often create a website with a certain idea of its structure. In addition, an actively maintained website is constantly updated: pages are added or deleted, links within the website are changed, and existing pages are changed. Therefore, the most effective way to improve the structure of a website is precisely the relinking of an existing website taking into account its original structure. The methodology "Semantically controlled optimization of internal relinking based on vector similarity of DOM structures" was used to restructure the website. Therefore, when changing automatically the structure L_i^{out} will remain unchanged.

Due to the specificity of building modern websites, some pages that have a large number of incoming links (home page, search page, rules page) can become almost isolated in the cluster structure. They are most often elements of the universal interface, and these pages can be accessed from almost any part of the website. Sometimes pages do have special tags, but if they are absent, then one can use the *Main* indicator, which reflects the number of incoming links. If $Main_i > 0.5|P|$, then the page will be considered the main one, and when the set of hyperlinks L of the website is automatically rebuilt, links to such pages will not be removed.

The key stage of automating the procedure for rebuilding the structure is changing the set of hyperlinks L taking into account the structural similarity of the pages. The following algorithm was used as the basis for building the approach:

Input data:
P – homepage of the website (initial URL).
$\theta_{min} \in [0,1]$ – threshold value of cosine similarity.
Output data:
Optimized graph $G`=(N,E`)$.
Algorithm steps:
1. Initialization:
– $P \leftarrow \emptyset$ – set of pages;
– $E \leftarrow \emptyset$ – set of hyperlinks;
– $List \leftarrow \{P\}$ – queue of pages to be processed;
– $Visited \leftarrow \emptyset$ – set of already processed pages.

2. Traversal and construction of graph with DOM vectorization:

While $List \neq \emptyset$:
2. 1. Take the first element p_i from *List*.
2. 2. Perform HTTP request to p_i and build DOM tree D_i.
2. 3. Save $V \leftarrow V \cup \{ p_i \}$.
2. 4. Perform DOM vectorization:

$$V_i = \varphi(D_i) \in R^m.$$

2. 5. Extract all hyperlinks from page p_i:

$L_i = \{ p_j | p_j$ mentioned in DOM as hyperlink$\}$.

2. 6. For each $p_j \in L_i$:
Add edge $E \leftarrow E \cup \{(p_i, p_j)\}$.
If $p_j \notin Visited \cup List$, add it to *List*.
2. 7. Add p_i to *Visited*.
3. Compute similarity matrix:
For all pairs $(p_i, p_j) \in V \times V$ compute

$$S_{ij} = \frac{V_i \cdot V_j}{\|V_i\| \, \|V_j\|}.$$

Form a matrix $S = \left[S_{ij} \right]_{|V| \times |V|}$.
4. Hyperlink optimization:
For each page $p_i \in P$:
4. 1. Determine the set of outbound links

$L_i = \{p_j \mid (p_i, p_j) \in E\}$.

4. 2. Find the page with the least similarity:

$$p_{j^*} = \arg\min_{p_j \in L_i} S_{ij}.$$

4. 3. If $S_{ij^*} < \theta_{\min}$ and $Main_{p_j^*} \leq 0.5|P|$:
Delete (p_i, p_{j^*}) from E.
Find page $p_h = \arg\max_{p_k \notin L_i} S_{ik}$.
Add (p_i, p_h) to E.
5. Completion:
Return modified graph $G`=(N,E`)$.

5. 3. Selecting metrics to assess the quality of partition

After the relinking process, it is necessary to assess the changes in the connections between thematically related pages. To assess the quality of the relinking, the following metrics were used in our work: mean, variance, median, modularity (Q) and clustering (C). The mean reflects the overall structural similarity between all pages. The higher the mean value, the more similar the pages are to each other in terms of DOM structure, type, number of sections, etc. It is calculated as follows

$$\mu = \frac{1}{N} \sum_{i<j} S_{ij}. \quad (2)$$

To display the spread of cosine similarity values, the concept of dispersion was used, in particular, the dispersion value makes it possible to determine how homogeneous the website structure is. It is calculated as follows

$$\sigma^2 = \frac{1}{N} \sum_{i<j} \left(S_{ij} - \mu \right)^2. \quad (3)$$

To determine the typical (average in the overall value) similarity of pages, the median calculation was used. In the case of increasing the values of this metric, on average, the pages become more similar to each other. Using the vector center of mass shift, it is estimated how much the website pages have changed in vector space on average after the automatic change in the website structure, and the calculation of the metric is carried out as follows

$$\Delta = \left\| \mu_{before} - \mu_{after} \right\|. \quad (4)$$

To determine the connectivity of a web graph, the modularity coefficient is used, and to determine the strength of the connectivity of vertices, the clustering coefficient is used. These coefficients are calculated as follows:

$$Q = \frac{1}{|E|} \sum_{ij} \left(A_{ij} - \frac{e_i^{out} e_j^{in}}{|E|} \right) \delta(p_i, p_j), \quad (5)$$

$$C = \frac{1}{|V|} \sum_i \frac{2N_{tv_i}}{\max(e_i^{out}, e_j^{in}) * (\max(e_i^{out}, e_j^{in}) - 1)}. \quad (6)$$

If the modularity coefficient takes on large values, the web graph consists of several strongly connected internally and weakly connected groups of vertices (clusters). The clustering coefficient estimates how strongly pages belonging to the same topic are connected to each other.

5. 4. Results of the computational experiment

The proposed website representation, the approach to determining the similarity of web pages was applied to analyze the structure of websites, to perform relinking of pages of the websites of online stores «Seeds of the Country» (http://semena-dnepr.org.ua/), «Sportano» (https://sportano.ua/), and «DICH» (https://dich.com.ua/).

Table 1 gives characteristics (number of pages, number of links between them) of the websites that were subject to analysis.

Table 1

Basic website characteristics

Website name Basic characteristics	«Seeds of the Country»	«DICH»	Sportano.ua
Number of pages	2518	12416	19621
Number of hyperlinks	365329	618425	1366512

Table 2 gives metric characteristics for the websites of online stores «Seeds of the Country» and «DICH» before and after restructuring the set of hyperlinks of the selected sites.

Table 2

Website metrics

Website name Metric specifications	Website «Seeds of the Country»		Website «DICH»	
	Before restructuring	After restructuring	Before restructuring	After restructuring
Mean (μ)	0.8229	0.8435	0.7645	0.7761
Variance (σ^2)	0.1193	0.1102	0.2194	0.2113
Median (M)	0.8895	0.9001	0.8856	0.8897
Mass center shift	0.0206		0.0116	
Number of modified hyperlinks	22335($\approx$ 6,11%)		18741($\approx$ 3,03%)	

Fig. 1, 2 show fragments of the website "Seeds of the Country" before and after the restructuring, respectively. The pages of two different categories of goods correspond to the vertices of the web graph, which are marked in red and blue, respectively. The links between the pages of the website within each category are marked in green and blue lines, respectively, and the connections between the categories are marked in purple lines.

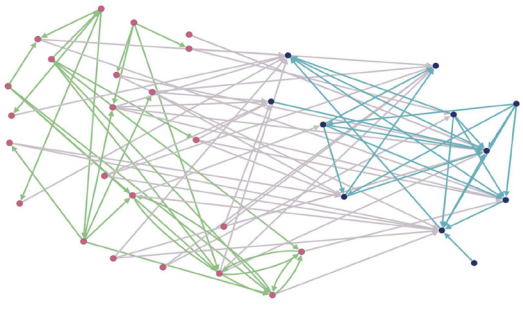


Fig. 1. Website fragment before automatic structure change

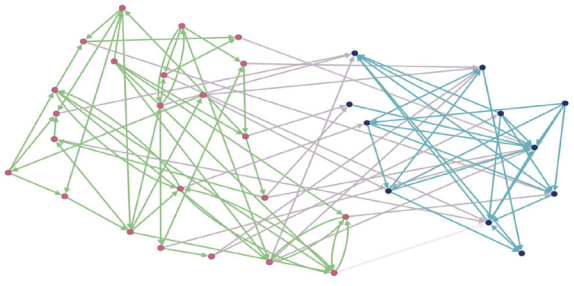
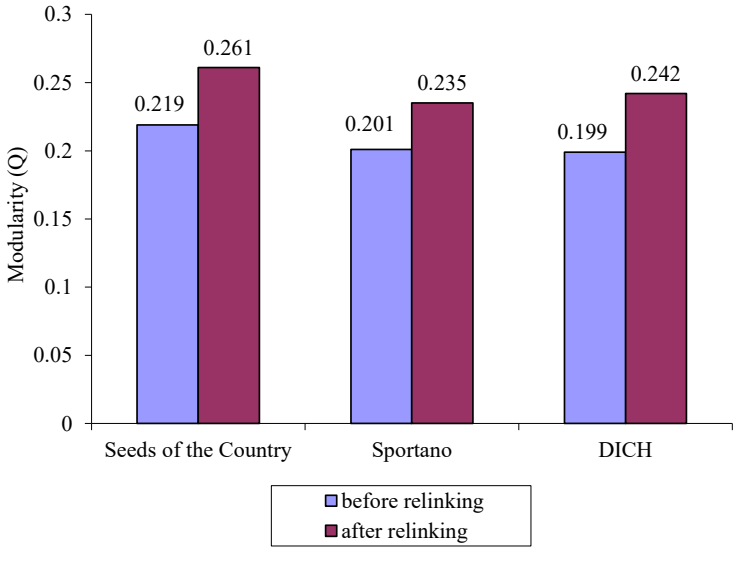


Fig. 2. Website fragment after automatic structure change

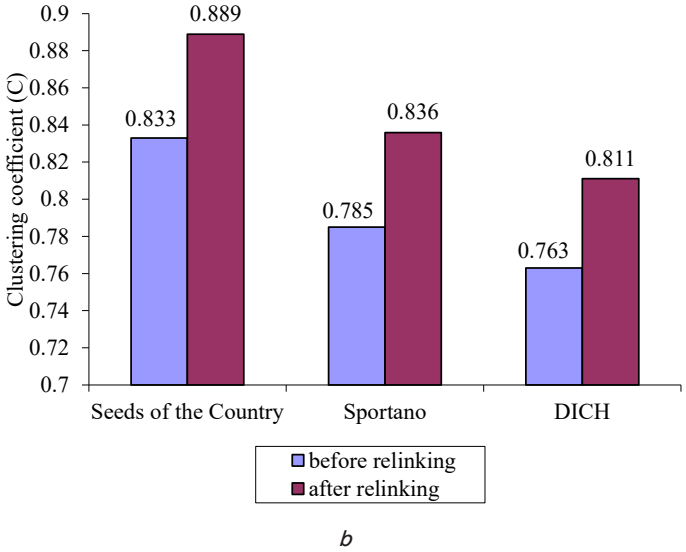
Table 3 describes changes in the characteristics of the clusters selected for consideration.

Fig. 3 shows the values of modularity and clustering metrics before and after relinking of the websites of online stores «Seeds of the Country», «Sportano», and «DICH».

The bar charts in Fig. 3 display the values of the Q and C metrics for the studied websites before relinking (in purple) and after relinking (in red). For the websites under consideration, an improvement in the metrics values is observed after relinking.



a



b

Fig. 3. Charts of website evaluation before and after relinking, using *a* – modularity; *b* – clustering coefficient

Table 3

Description of the content of clusters at the «Seeds of the Country» website

Cluster name Cluster occupancy	Product cluster «Beet» (https://semena.in.ua/buriak-...)		Product cluster «Melon» (https://semena.in.ua/dinya-...)	
	Before restructuring	After restructuring	Before restructuring	After restructuring
Number of vertices in a cluster	23		11	
Number of internal links	28	46	22	25
Number of links to a neighboring cluster	34	16	5	2

6. Discussion of results based on automated reconstruction of internal hyperlinks of websites

The use of an extended hypertext model, supplemented by a vector representation of DOM trees, allowed us to successfully formalize the structure of the studied web resources. This approach has made it possible to transform HTML documents into a multidimensional feature space, which became a reliable basis for further analysis of hidden logical connections in the internal architecture of sites without the need for manual expert intervention.

The implementation of the method of semantically guided optimization based on the calculation of the cosine similarity measure proved its effectiveness during the practical reconstruction of test web graphs. Analysis of the formed similarity matrices revealed the ability of the algorithm to accurately identify structurally and thematically close pages that did not have direct connections in the original structure. As a result of the implementation of the proposed algorithm, weakly relevant hyperlinks were automatically replaced with optimal ones, while the introduced criterion for preserving service pages guaranteed the absence of violations in global navigation.

Our study analyzed websites of online stores characterized by a different number of pages and unequal density of internal hyperlinks among them (Table 1).

Comparison of metric characteristics for online store websites (Table 2) before and after relinking showed that the characteristics of online store websites differ primarily in the level of structural homogeneity of pages and the nature of internal links. This causes different values of average indicators and variability of the corresponding metrics.

Online store websites are usually constructed on designer platforms, in particular WordPress, Shopify, or OpenCart. Such systems provide ready-made templates and tools for customizing the design, adding products, organizing the cart, and paying without the need to write programming code. As a result, the pages of such websites have close structural vectors, which determines a relatively high value of the average $\mu \approx 0.8229$ and a small variance $\sigma^2 \approx 0.1193$. In addition, the structure of such resources is formed taking into account a clear thematic and hierarchical organization (categories, subcategories, product cards), so the variability of the variance indicators σ^2 is usually small.

In particular, the website “Seeds of the Country” has 365,329 links between pages, of which 22,335 were changed by the algorithm, which is 6.11% of the total. Most of the changes are related to links between product pages, and the recorded increase in μ and decrease in σ^2 indicates a more even distribution of incoming links after relinking. This, on the one hand, simplifies navigation to specific pages, and on the other hand, increases the template nature of the internal structure of the website.

Fig. 1, 2 show a fragment of the website «Seeds of the Country» before and after the restructuring of the internal structure due to the relinking of its pages, respectively. The figures show two different categories of goods, the pages of which are displayed by red and blue vertices of the graph. Links in the middle of the first and second categories are marked with green and blue lines, respectively, and links between categories are marked with purple lines.

From the analysis of Fig. 1, 2, it can be seen that the number of links between pages within a category after relinking increased from 28 to 46 links for the first category of goods, from 22 to 25 links for the second category of goods. Before relinking, it was observed that the connections between the two categories under consideration are dense; however, after relinking, a clearer division of the website structure into clusters is observed. This is due to the increase in the number of links between thematically related pages within each of the categories.

Table 3 describes changes in the cluster structure of the website “Seeds of the Country”. Before the relinking, there were a large number of links, namely 34, from product pages in the “Beet” category to product pages in the “Melon” category. Before the relinking process, the number of links to products in other categories was greater than the number of internal links. After the relinking process, the clusters became denser with a greater number of internal links.

Analysis revealed that most of the links between the pages of the website that were changed during the relinking were related to the product pages. However, not all changes in the structure are positive. Sometimes, during the operation of the algorithm, a link between the product pages of one category was replaced by a link to the product page of another category. This shows that even pages of the same topic can have many differences in structure. Such differences have a greater impact when calculating cosine similarity than the thematic identity of the pages.

Before the relinking of the DICH website (https://dich.com.ua), there were 618425 links among pages in its structure. As a result of applying the relinking algorithm, a smaller number of changed links is observed, namely 18741, which is 3.03% of the total number of links. This may indicate a high quality of the logical structure of the web resource, which makes it more difficult to find weakly connected pages of identical topics, or the inability to more clearly distinguish the topic of the pages and assign it to one specific cluster. However, a higher value of $\sigma^2 \approx 0.2194$ rather indicates a wider thematic spread of the website pages than that of the «Seeds of the Country» website. This is due, among other things, to the fact that product pages are more difficult to assign to only one category.

Fig. 3 shows values of the modularity metrics and clustering coefficient before and after relinking of the websites of online stores «Seeds of the Country», «Sportano», and «DICH».

For online shopping websites, these indicators are higher due to the possibility of a clear thematic division of pages into clusters by product categories. After applying the proposed approach to hyperlink relinking, the values of modularity and clustering coefficient increased, which indicates the formation of clearly expressed thematic clusters in the structure of websites.

In Fig. 3, the largest increase in modularity is observed for the website «Seeds of the Country» and it is 0.042 (≈19.4%). The average initial value of modularity for the analyzed websites is 0.192, the increase in modularity is 0.024 (≈12.6%). In work [12], the author uses structural analysis of the web graph and the Louvain clustering method to devise a recommender system for changing the link structure. The initial average modularity is 0.176, and the increase is 0.015 (≈8.6%). This indicates a greater effectiveness of methods for semantic analysis of a web resource in the case of online shopping websites.

Thus, the proposed approach makes it possible not only to recreate the current structure of the website but also optimize the structure using internal linking between pages of the web resource. This is achieved by identifying potentially relevant pages between which new navigational links can be created. The devised approach provides a quantitatively measurable improvement in the structural organization of websites, makes it possible to form more pronounced thematic clusters and optimize navigational links, while maintaining the basic logic of the resource.

The limitation of our study is that the DOM model analyzes only the HTML structure, that is static websites. Websites that have a dynamic structural structure of pages cannot be analyzed by these methods. The disadvantage of this study is the analysis of only the semantic features of web pages. The DOM model contains information about the structure, but it is in an unstructured form. In the future, it is possible to use a combined model based on the graph and DOM model.

Further studies may involve construction of a combined DOM-graph model, which is simultaneously able to take into account both the structural characteristics of pages and their thematic features. However, the study of such a model is associated with methodological difficulties, in particular, the lack of established approaches and known methods for analyzing such a structure. Therefore, there is a need to adapt existing analysis methods that can work with heterogeneous sets of characteristics, or to devise new methodological approaches for such models.

7. Conclusions

1. A website is formally depicted through a combination of a directed graph of hyperlinks with a vector representation of DOM-structures of pages, which allowed us to form a holistic multi-level system for analyzing the structure of a web resource. Unlike traditional approaches that take into account only the topology of links, the extended model integrates structural and semantic features of pages, which provides a deeper interpretation of the functional organization of the website. Formalization of the page in the form of a DOM-vector of features allowed us to quantitatively assess the similarity of pages, perform their clustering, and provide substantiated recommendations for restructuring the internal structure. The combination of graph and vector components creates a theoretically grounded toolkit for analyzing connectivity, identifying structural imbalances, and optimizing the navigation model of the website.

2. An approach to determining the similarity of web pages based on the cosine similarity of their DOM-vectors has been devised, which makes it possible to formalize the structural and semantic proximity of pages in numerical form. Building a similarity matrix forms the basis for informed optimization of internal linking by replacing the least relevant links with those that are structurally consistent, while preserving key navigation nodes (in particular, pages with a high Main score).

3. To ensure a comprehensive assessment of the quality of automatic reconstruction of the internal structure of a website both in the vector space of features and at the level of the graph structure, a system of metrics has been proposed, in particular, the average, variance, median, modularity and clustering coefficient were calculated. The proposed set of indicators formed a consistent system of criteria for objective comparison of the website structure before and after relinking and allowed us to identify both an increase in thematic coherence and possible signs of excessive templatization of the structure.

4. Analysis of the results of our experimental study confirmed the effectiveness of the proposed approach to automated reconstruction of internal hyperlinks based on the similarity of DOM structures. Statistical indicators (average, variance, median) allowed us to quantitatively assess the degree of structural homogeneity of pages and the nature of changes in their mutual similarity. Modularity and clustering metrics reflected the transformation of the web graph topology and the strength of internal thematic connections. It was found that the application of the devised approach leads to an increase in modularity and clustering coefficient after relinking. This indicates the formation of clearer thematic clusters and improved internal connectivity of product categories. It also confirms the increase in the thematic orderliness of the web resource structure.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study and the results reported in this paper.

Funding

The study was conducted without financial support.

Data availability

The data will be provided upon reasonable request.

Use of artificial intelligence

The authors declare the use of generative AI in the research and manuscript preparation process. The tasks delegated to generative AI tools are checking the grammar, spelling, and punctuation of the manuscript, and searching for literary sources. The generative AI tool used is ChatGPT-4.

The final version of the literary sources was generated by the authors themselves. The use of artificial intelligence tools did not affect the scientific results or conclusions of the study, and full responsibility for the content and integrity of the manuscript rests with the authors.

Acknowledgments

The research was supported by the Ministry of Foreign Affairs of the Czech Republic under project 25-PKVV-UM-004 «Development of Education, Research, and International Cooperation at Oles Honchar Dnipro National University» implemented by Charles University and Oles Honchar Dnipro National University.

Authors' contributions

Ivan Dolotov: Methodology, Software, Validation, Writing – review and editing; Resources, Data curation, Visualization; Project administration; **Nataliia Huk:** Conceptualization, Methodology, Formal analysis, Investigation, Supervision, Writing – original draft, Validation.

References

1. Huk, N. A., Dykhanov, S. V., Matiushchenko, O. D. (2020). Algorithm for building a website model. Bulletin of V.N. Karazin Kharkiv National University, Series «Mathematical Modeling. Information Technology. Automated Control Systems», 47, 25–34. https://doi.org/10.26565/2304-6201-2020-47-03
2. Dolotov, I. O., Guk, N. A. (2023). Clustering of a weighted webgraf with the usage of modularity. 2023: Problems of applied mathematics and mathematical modeling, 23, 25–32. https://doi.org/10.15421/322305
3. Ma, W., Chen, X., Shang, W. (2012). Advanced Deep Web Crawler Based on Dom. 2012 Fifth International Joint Conference on Computational Sciences and Optimization, 605–609. https://doi.org/10.1109/cso.2012.138
4. 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. https://doi.org/10.15587/1729-4061.2022.265961
5. Kao, H.-Y., Ho, J.-M., Chen, M.-S. (2005) WISDOM: Web Intrapage Informative Structure Mining based on Document Object Model. IEEE Transactions on Knowledge and Data Engineering, 17 (5), 614–627. https://doi.org/10.1109/tkde.2005.84
6. Ahmad Sabri, I. A., Man, M. (2018). Improving Performance of DOM in Semi-structured Data Extraction using WEIDJ Model. Indonesian Journal of Electrical Engineering and Computer Science, 9 (3), 752. https://doi.org/10.11591/ijeecs.v9.i3.pp752-763
7. Huynh, H., Le, T., Nguyen, V., Nguyen, T. (2024). A DOM-structural Cohesion Analysis Approach for Segmentation of Modern Web Pages. https://doi.org/10.21203/rs.3.rs-4392630/v1
8. Shin, K., Niiyama, T. (2018). The Mapping Distance – a Generalization of the Edit Distance – and its Application to Trees. Proceedings of the 10th International Conference on Agents and Artificial Intelligence, 266–275. https://doi.org/10.5220/0006721902660275
9. Jalal, A. A., Jasim, A. A., Mahawish, A. A. (2022). A web content mining application for detecting relevant pages using Jaccard similarity. International Journal of Electrical and Computer Engineering (IJECE), 12 (6), 6461. https://doi.org/10.11591/ijece.v12i6.pp6461-6471
10. Kumar, B. T. H., Vibha, L., Venugopal, K. R. (2016). Web page access prediction using hierarchical clustering based on modified levenshtein distance and higher order Markov model. 2016 IEEE Region 10 Symposium (TENSYMP), 1–6. https://doi.org/10.1109/tenconspring.2016.7519368
11. Roul, R. K., Devanand, O. R., Sahay, S. K. (2014). Web Document Clustering and Ranking using Tf-Idf based Apriori Approach. IJCA Proceedings on ICACEA, 2, 34. https://doi.org/10.48550/arXiv.1406.5617
12. Meleshko, Ye. (2019). Graph clustering methods in social networks for building recommendation systems. Control, Navigation and Communication Systems, 2 (54), 129–134. https://doi.org/10.26906/SUNZ.2019.2.129