TASHKENT UNIVERSITY OF
INFORMATION TECHNOLOGIES
NAMED AFTER MUHAMMAD AL-KHWARIZMI

MUHAMMAD AL-XORAZMIY NOMIDAGI
TOSHKENT AXBOROT TEXNOLOGIYALARI
UNIVERSITETI

BULLETIN OF TUIT:

MANAGEMENT AND COMMUNICATION TECHNOLOGIES



ISSN:2181-1083 DOI: Date:2025-11-10
B. Mo'minov, Sherzod Abdimannonovich Husanov № 3(29)2025

A Mathematical Model and Algorithm for Evaluation Based on the Similarity of Web Document Objects

B. Mo'minov
Doctor of Technical Sciences, Professor
Tashkent State University of Economics.

Sherzod Abdimannonovich Husanov
Senior Lecturer, Department of Multimedia Technologies, Tashkent University of Information Technologies named after Muhammad al-Khwarizmi.

Abstract: This article develops a mathematical model and an algorithm for identifying and evaluating the similarity of web document objects. The proposed approach enables the assessment of the functionality of objects on web pages, the degree of their interrelatedness, and their effectiveness based on user interactions. In addition, algorithmic solutions based on a graph model are proposed for analyzing object similarity and their mutual interactions.

Keywords : Web document, object similarity , graph model, clustering, algorithm, user interaction.

Introduction: In modern web systems, analyzing user activity and determining the interrelationships of objects on pages is of great importance. By determining the similarity of objects, the efficiency of web pages is increased, their adaptability to the user is ensured. Therefore, in this article, a mathematical model and algorithm for evaluating web document objects based on their similarity are developed.

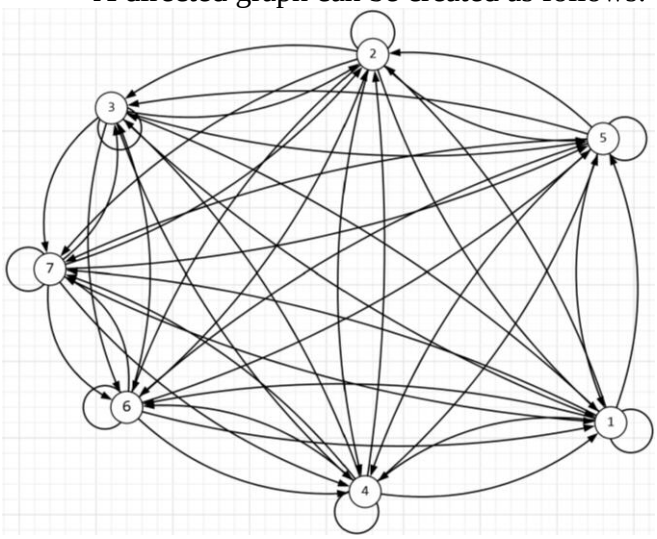
As mentioned in the above paragraph on calculating the similarity of web document objects, algorithms for evaluating models based on the similarity of objects are proposed in two parts.

the similarity of objects in a web document be given in the form of a matrix $Q = (q_{i,j})$Q , where Q is the similarity of objects i and j of the web page, respectively . If each property of an object is viewed as a node in a directed graph, a graph model can be constructed for the similarity of objects in a web document.

Directed graph$D = (V,\ E)$ If we express it as, then V is a vertex of the graph and in this case its objects are vertices of the graph. E is is an ordered set, $e_i \in V, i = 1,..,n,n -$the number of objects.

The edges entering the graph vertex$e_i^I \in V,$ Number of objects with outgoing arcs$e_i^O \in V,\ i = 1,..,n,n -$ Let the ends of these arcs be the input or output weights w_{ij} , and the similarity of the i -object to the j -object is taken.

A directed graph can be created as follows:



1. A graph model based on the similarity of web document objects.

In the graph model based on the similarity of web document objects, similarity weights are not provided. These weights are expressed in matrix form as follows

$$Q = \begin{pmatrix} q_{1,1} & q_{1,2} & q_{1,3} & \dots & q_{1,n} \\ q_{2,1} & q_{2,2} & q_{2,2} & \dots & q_{2,n} \\ q_{3,1} & q_{3,2} & q_{3,3} & \dots & q_{3,n} \\ \dots & \dots & \dots & \dots & \dots \\ q_{n,1} & q_{n,2} & q_{n,3} & \dots & q_{n,n} \end{pmatrix} \quad (1)$$

Web document objects The following definitions are given for evaluating the similarity ratio in a graph model:

Definition 1. When i=j , the similarity of web page objects $q_{i,j} = 0$ i and j is considered to be exact.

And from this, from the metric of $q_{i,j}$ incoming and e_j^O outgoing arcs for ,e_i^I $q_{i,j} = e_i^I - e_j^O = 0$ It turns out that.

The similarity matrix of web document objects is not a completely symmetric matrix, i.e. the values are not the same. Based on the values of the matrix, the e_i^I input Qrow, i.e.

$(q_{i,1}, q_{i,2}, \dots, q_{i,n})$and e_j^O the output row $Q$the column of, i.e. $(q_{1,i}, q_{2,i}, \dots, q_{n,i})$for arcs values are accepted.

When assessing the activity of an object on a web page, it is proposed to draw conclusions from its similarity matrix based on the following expression.

$$W_j = \sum_{i=1}^{n} q_{i,j} - \sum_{i=1}^{n} q_{j,i} \quad , j = 1..m \qquad (2)$$

For the value of expression (2) above, the following property must be introduced:

1. If $W_j > 0$ so, $q_i$it is a transition from one object to another, and this object is considered extremely important for the web page.

2. If $W_j = 0$ there $q_i$is an equal number of visits from an object to another object and from another object to this object, it is considered a good (satisfactory) object for a web page.

3. If $W_j < 0$ so, $q_i$it is a transition from one object to another, and this object is considered extremely unsatisfactory for the web page. This object should be improved or removed as soon as possible.

It is proposed to estimate both the most active and the most passive objects from the values of the matrix defined by expression (1) above.

The most active object is evaluated as follows:

$$d_{max}(q_i, q_j) = \max_j \left(\min_i q_{i,j}\right)$$

The most passive object is evaluated as follows:

$$d_{min}(q_i, q_j) = \min_j \left(\min_i q_{i,j}\right)$$

Good objects, between the most active and the most passive objects, are evaluated as follows:

$$d_{x_i}(q_i, q_j) = \max\left(d_{x_i}(q_i, q_j), W_i + d_{x_i}(q_i, q_j)\right)$$

Evaluating the similarity of objects in a web document allows for object sorting.

To determine the rating of objects in a web document, it is necessary to determine the number of user accesses to this object, that is, the number of arrivals and departures. This is done by controlling the movement of the cursor on the web page or by controlling the movement of the mouse and counting its increments for the object. For this, you can rely on the following 5 main mouse events:

1. MouseDown – movement of the mouse within the object's area of influence according to the x coordinate

2. MouseUp - movement of the mouse within the object's range of influence according to the y coordinate

3. MouseMove – moves the mouse within the object's range of influence exit movement

4. Click (mouseClick) – selection action by clicking the mouse button once within the object's range of influence

5. DbClick (mouseDbClick) - double-clicking the mouse button within the scope of an object, the selection action, if this event occurs, gives the action of moving to another object.

All of the above mouse actions are available and designed in all programming languages and scripting languages. These events can be handled in objects written for each page of a web document. The following approaches are proposed for this.

1. the MouseDown and MouseUp events, a unique identifier i (i=0) is declared for each object . This i++ operation is executed once per event on the object in real time.

2. MouseMove event, a unique identifier j (j=0) is declared for each object . This j++ operation is performed once per event on the object in real time.

3. Click event, a unique identifier c (c=0) is declared for each object . This C++ operation is executed once per event on the object in real time.

4. DbClick event, a unique identifier dc (dc=0) is declared for each object . This dc++ performs the action once per event on the object in real time.

These variables perform their function from the moment the web page is launched to the moment it receives statistical data. These 4 events are constantly running for each object and are calculated using the following expression.

$$E_I = i + c, \quad E_O = j + dc$$

$E_I -$ These are incoming calls to the object and $E_O -$outgoing calls from the object. The events are divided into 2 types, taking into account the characteristics of the events.

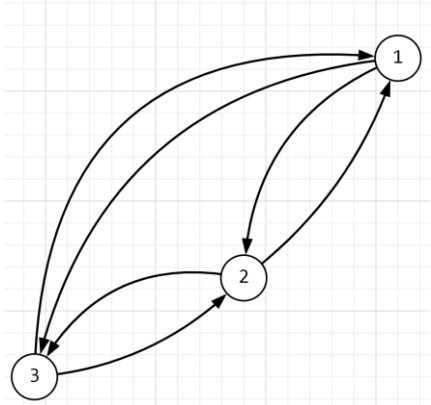
An information model can be constructed by representing event-based input and output requests in the form of a graph model. This allows the current object to be considered as the vertices and edges of the graph, and the incoming and outgoing requests to be taken as the vertices and edges of the graph.

Directed graph$G = (O, E)$ If we express it as, then O is a vertex of the graph and in this case the objects are vertices of the graph. E is is an ordered collection , with incoming and outgoing

calls, and the number of objects, respectively $e_i \in V, i = 1,..,n,n -$.

The reference arcs that enter the graph vertices$e_i^I \in V$, Number of objects in outgoing address arcs$e_i^O \in V$, $i = 1,..,n,n -$ Let the ends of these arcs w_{ij} be the input or output weights, and the number of input or output references of the i -object to the j -object is taken.

A directed graph can be created as follows:



2. An example of a graph model based on the architecture of web document objects.

In general, let there be n web pages in a web document and m objects $N = nm$on each of these pages . The total number of objects The input and output references of a web document can be represented in the form of a matrix M. For this, let the following be true:

1. Let each corresponding row of the matrix M give corresponding output references.

2. Let each corresponding column of the matrix M give corresponding input references.

3. this case, the diagonal elements of the matrix M are value$E_D = (E_I + E_O)/2$ is calculated by the expression.

The general form of the matrix M is expressed as follows:

$$M = \begin{pmatrix} E_{D_{1,1}} & E_{O_{1,2}} & E_{O_{1,3}} & \cdots & E_{O_{1,n}} \\ E_{I_{2,1}} & E_{D_{2,2}} & E_{O_{2,3}} & \cdots & E_{O_{2,n}} \\ E_{I_{3,1}} & E_{I_{3,1}} & E_{D_{3,3}} & \cdots & E_{O_{3,n}} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ E_{I_{n,1}} & E_{I_{n,2}} & E_{I_{n,3}} & \cdots & E_{D_{n,n}} \end{pmatrix}$$

M each from the matrix E_I^k , E_O^k The following expression is appropriate for the object

$$E_I^k = \sum_{i=1, i\neq k}^{N} E_{I_{k,i}}, E_O^k = \sum_{i=1, i\neq k}^{N} E_{O_{k,i}} \; k = 1 \ldots N \qquad (2)$$

(2) we can draw the following conclusions:

for $E_I^k - E_O^k < 0$each k If so, this object needs to be updated or modified.

for $E_I^k - E_O^k = 0$each k If so, this object is considered well-satisfied.

for $E_I^k - E_O^k > 0$each k If so, this object is very good and indicates that more such objects should be created.

Based on the above proposition, a modified version of the matrix M for a web document, M *, is derived.

$$M^* = \begin{cases} 1, & E_I^k - E_O^k > 0 \\ 0, & E_I^k - E_O^k = 0 \\ -1, & E_I^k - E_O^k < 0 \end{cases}$$

M^* Based on the values of the vector, a visual representation of the objects of the web document that require modification is created.

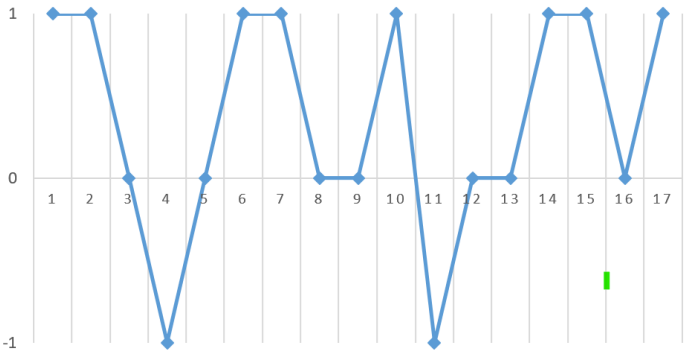


Figure 3. Graph of object values based on references for a web document consisting of 17 objects.

The input and output references of web documents can be changed based on them, but there are also objects that are the best complements of objects, that is, belong to the same class. To identify such objects, the clustering problem can be used to identify objects belonging to the same class for a web document, since the analysis of web documents does not specify classes in advance . For this, the following metric can be used:

$$d_{x_i}(x_i, x_j) = \max\left(d_{x_i}(x_i), |x_i - x_j| d_{x_i}(x_j)\right)$$

d_{x_i} Using the metric, a cluster-based clustering design was performed (see Figure 4).

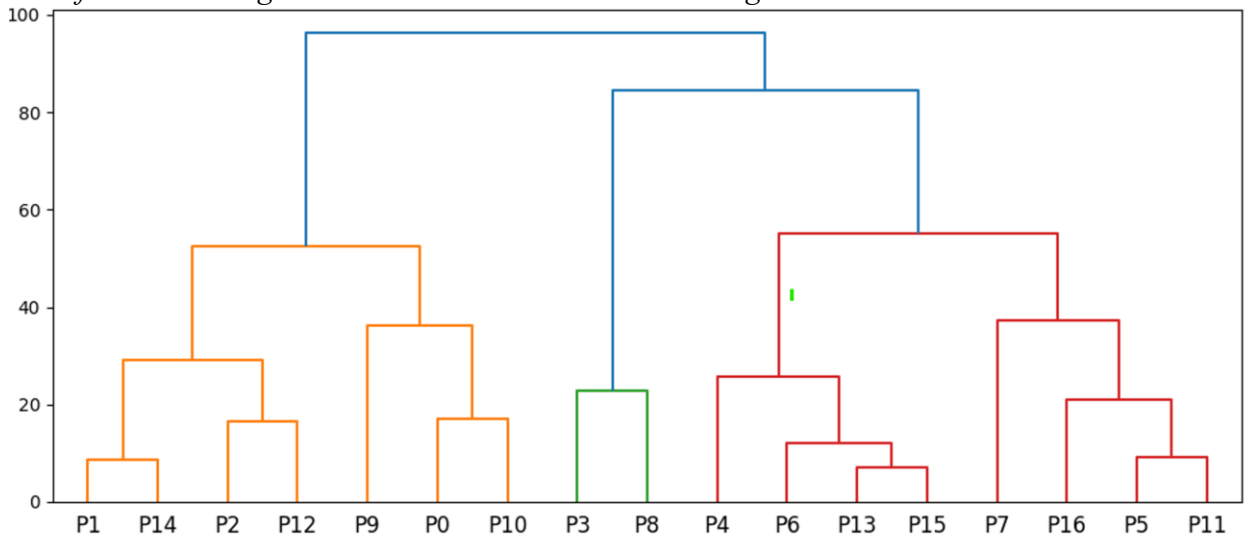


Figure 4. Example of clustering a web document

consisting of 17 objects.

A general algorithm for evaluating the similarity of Web document objects has been developed based on mathematical models of evaluation. It includes :

1. Similarity of objects
2. Number of object references
3. Similarity in number of object references

Separate subprograms are created for these 3 mathematical models and they return values in the form of matrices. To bring them to the same type, the following criteria are introduced:

For the similarity of objects, the interval [a, b] is introduced and the following expression is used in it.

$$O_1(x) == \begin{cases} 1, & x > b \\ 0, & x \in [a,b] \\ -1, & x < a \end{cases}$$

Number of object references for$x = E_I^k - E_O^k$ is entered and the following expression is used in it.

$$O_2(x) == \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}$$

Similarity in number of object references forC^+, C^- A class family is introduced and the following expression is used in it.

$$O_3(x) == \begin{cases} 1, & x \in C^+ \\ 0, & x \notin C^+ va\ x \notin C^- \\ -1, & x \in C^- \end{cases}$$

The general expression for the similarity of objects is defined as follows:

$$O = \frac{O_1(x) + O_2(y) + O_3(z)}{3}$$

The algorithm based on the object similarity assessment model proposed above is as follows.

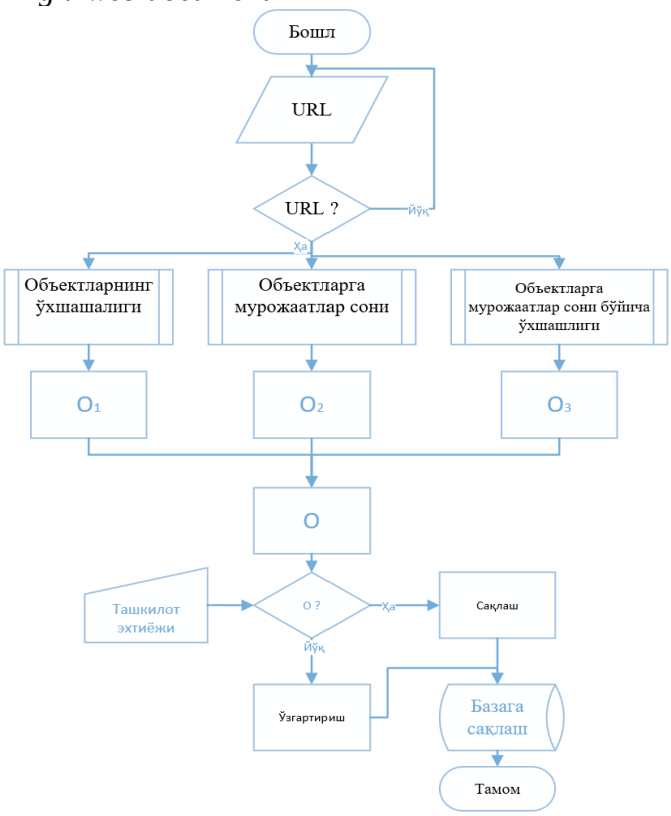


Figure 5. Web document object similarity-based scoring algorithm.

Based on this algorithm, the problems of updating a web document and attracting customers by changing the objects of a web document are solved. Currently, each organization has created its own calculation mechanisms, mainly based on traditional, well-known methods. This proposed assessment of the similarity of web document objects is aimed at educational and legal sites. In this assessment, there are events that perform a mutual function, and they must be processed as noise in the document. This issue will be considered in the next paragraph.

References:

[1]. Viyyameena M. K., Kavita K. "A Survey on Similarity Measures in Text Mining" / Machine Learning and Applications: An International Journal, Vol.3, No.1, March 2016. — Metrics that measure document similarity and their role in clustering are explained.
[2]. Wu M., et al. "Modeling Text Similarity: A Survey" / 2023. — Modern applications and directions of text similarity modeling.
[3]. Cha S. H. "Comprehensive Survey on Distance/Similarity Measures" — An analysis that considers various similarity/distance metrics.
[4]. Battler D. "A Short Survey of Document Structure Similarity Algorithms" 2004. — Analysis of similarity algorithms by document structure.
[5]. Mohammed S. M., Jaksī K., Zeebary S. R. M. "A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms" — Cases in semantic similarity and document clustering.
[6]. Ahmed T. "Exploring Mathematical Models and Algorithms for Plagiarism Detection in Text Documents: A Proof of Concept" — Mathematical models and algorithms for detecting plagiarism in text documents.