V. M. Dubovoy, Dc. Sc. (Eng); O. M. Moskvin

DEVELOPMENT OF THE HYPERTEXT STRUCTURES FUZZY CLASSIFICATION SYSTEM

In this paper fuzzy system for the evaluation of information hypertext structures optimality is suggested. The existing metrics are investigated for the formalization of their functionally important parameters on the basis of data obtained while studying the influence of hypertext structure types on the values of their evaluation indices.

***Keywords**: Hypertext, hypertext metric, compactness index, stratification index, fuzzy classification, structure optimality.*

Introduction

Hypertext information systems are extremely common because they are the basis of Internet resources. Hypertext fragments network is characterized by a complicated, chaotic and non-optimized structure, which complicates considerably the search for required information. The problem of hypertext characteristics objective evaluation with the aim of its optimization is gaining ever more significance with the current development of Internet technologies..

In hypertext theory for the formalization of its functionally important parameters, a special hypertext metric exists [1]. It includes two evaluation indices – information compactness index and stratification index. As a model of hypertext information environment the oriented graph has been chosen with corresponding fragments being its nodes and connections between them being its edges.

Information compactness index characterizes the degree of the hypertext structure intersection by links [2]:

$$Cp = \frac{CD_{\max} - CD}{CD_{\max} - CD_{\min}}, \quad (1)$$

where $CD_{\max}$ is maximally possible number of steps to be taken through the links connecting all hypertext nodes; $CD_{\min}$ – minimally possible number of steps connecting all hypertext nodes; CD – indicator of the paths in the graph. To determine it, the calculation of a transformed matrix of distances is required.

The value of information compactness index changes in the range of [0; 1], which allows for the mutual comparison of hypertext documents. Absolutely link-free hypertext is characterized by zero value of information compactness index – *Cp=0*, and, vice versa, absolutely linked one – by the value of *Cp=1*. High compactness level is characteristic of such hypertext structures where each information block can be accessed from any other information block, which is usually provided by a large number of cross-references. It should be noted that extremely high compactness can lead to total disorientation of a hypertext structure user. At the same time, if information compactness is too low, it causes the situation when hypertext fragments remain out of the user's sight or separate fragments become inaccessible.

Stratification index is considered in detail in [1] and introduced for a hypertext linearity characterization [3]:

$$St = \frac{AP}{LAP}, \quad (2)$$

where AP is absolute stratification, LAP – linear absolute stratification of the hypertext with n nodes identical to absolute stratification of linear hypertext of analogous dimensionality. It is calculated by the formula:

$$LAP = \begin{cases} \dfrac{n^3}{4}, & \text{if } n \text{ even} \\ \dfrac{n^3 - n}{4}, & \text{if } n \text{ odd} \end{cases} . \quad (3)$$

In the case of an absolutely stratified hypertext, stratification index takes the value of *St=1* and, if vice versa, *St=0.* In fact, stratification index makes it possible to evaluate connectivity degree of the elements that are at different levels of the hierarchy.

The fraction of missing paths is semantically connected with information compactness index:

$$K_m = \frac{Q_m}{n^2 - n}, \quad (4)$$

where Q_m is a number of missing paths in the graph [3]. In order to calculate the coefficient of the missing paths, first it is necessary to define the graph distances matrix.

The maximal number of missing paths is $n^2 - n$, minimal – 0. The fraction of missing paths changes in the range of [0; 1] and allows for the mutual comparison of hypertext document systems.

Cyclomatic number is characterized by graph structure being different from dendritic structure and is calculated by the formula:

$$Cn = m(G) - n(G) + p, \quad (5)$$

where m(G) is the number of edges, n(G) – number of nodes, p – number of the graph linked components [3].

Cyclomatic number shows the smallest number of edges to be eliminated so that the graph would become a tree. For a strongly linked graph *p*=1.

The analysis of hypertext structure evaluation criteria has shown that their separate usage is not universal and impossible for obtaining adequate characteristic of a structure due to their functional limitations.

Results shown in fig. 1 illustrate insufficiency of each index for hypertext quality evaluation. Stratification index remains the same for both dendritic structures without internal hierarchical links and with them. On the other hand, the compactness index value is almost the same for both linear closed structure and for hierarchical structure.

The goal of the work is to build the system for fuzzy evaluation of hypertext structure quality as a means of complex application of the existing indices.

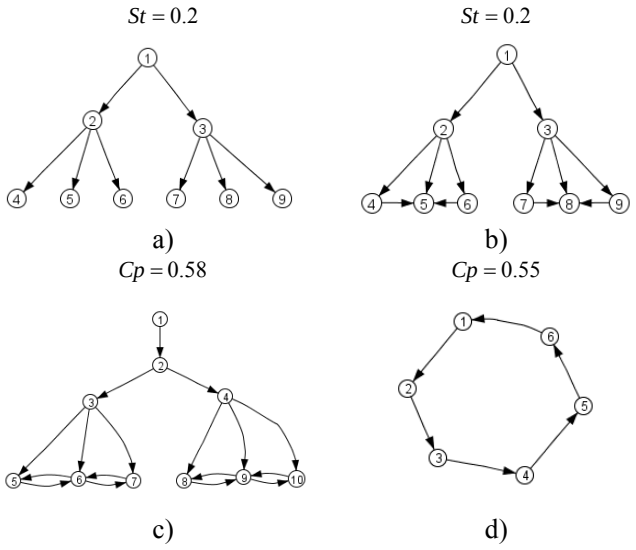


Fig. 1. Investigation of the structures form dependence on their indices values: a, b – stratification index; c, d – information compactness.

For evaluation of the hypertext structure influence on the criteria values, research was conducted on the modification of hypertext hierarchical structures and calculation of the values of the above-discussed indices for them.

Basic hierarchical structure that was subjected to modifications is a tree having only unidirectional links. This structure is characterized by distinct stratification and has the form presented in fig. 2. The investigations were conducted by iteration method: at each step, by means of the developed hierarchical structures generator, one edge was added to the graph, and

computations of the information compactness index, stratification index were performed as well as computations of missing paths fraction and cyclomatic number of the graph.

Investigation was carried out on the influence of different connection types on the values of the hypertext structure evaluation criteria:

- Bilateral links between the neighboring levels;
- unidirectional links from lower levels to the levels of higher hierarchy;
- unidirectional links from higher levels to the levels of lower hierarchy;
- horizontal links between the elements of individual subhierarchies;
- horizontal links between the elements of different subhierarchies.

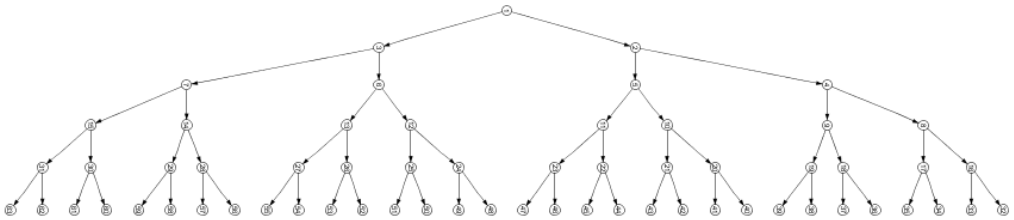


Fig. 2. The form of the basic hierarchical structure

Graphic principle of graph structure modification, used in this work, is shown in fig. 3. From lower levels links are added to all higher nodes of each hierarchy. These links are shown with dash-dot lines in fig. 3.

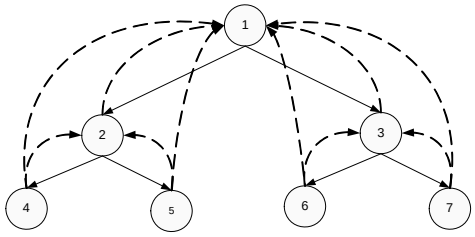


Fig. 3. Graphical interpretation of the structure modification principle

As it is evident from the research results (fig.4), variations of the stratification index value have linear character. Initial increase of stratification index is caused by the appearance of links between the end nodes (those not having outgoing connections) and those occupying higher position in the hierarchy. Slow falling of stratification index results from the structure being complicated by cross-references.

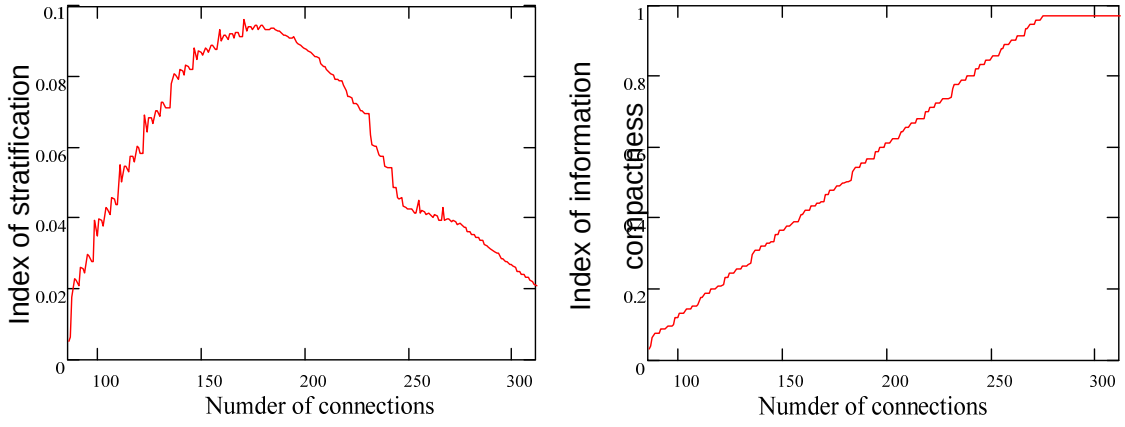


Fig. 4. Graphic dependences of the variations in the hypertext structure evaluation criteria on the number of unidirectional links of the “from bottom to top” type in the graph:
a) stratification index; b) information compactness index

The conducted research makes it possible to establish heuristic relationship between the variations of individual hypertext characteristics and its quality:

1. The change of a hypertext structure by the addition of only one higher level to feedback links allows the user to return only one step back in the navigation process. Such method gives the possibility to make the hypertext accessible from any node. Separate usage of only this approach is not good because it requires a large number of user transitions between the nodes during the information search process. A more acceptable approach to hypertext structure modification is the one where except feedback links unidirectional links are added to the levels of a higher hierarchy. The speed of accessing the fragments will be higher in such structure. The optimal complexity will be considered to be such that corresponds to the inflection point of the characteristic of stratification index dependence on the number of links. This point corresponds to the beginning of the structure degeneration process.

2. As the research has shown, introduction of unidirectional links from higher to lower hierarchic levels destroys stratification and has almost no influence on the variations of information compactness index. Therefore separate usage of this approach alone is insufficient: although the time of access from higher to lower levels increases, higher levels become overloaded with references, which prevents users from returning to them.

3. Horizontal connections between the elements of different subhierarchies have inconsiderable influence on the information compactness index. And really, the time of accessing fragments increases considerably due to the appearance of links between different hierarchies. The optimal complexity of such links will be considered to be such that corresponds to the inflexion point of the characteristic of stratification index dependence on the number of links.

The obtained results are summarized in the form of a fuzzy system of the hypertext structure quality evaluation. The application of fuzzy logic to this problem is justified because the obtained dependences of the hypertext structure type on the values of its indices have complex non-linear character. Search for analytical dependences that describe them is problematic and evaluation parameters can give, in some cases, inconsistent results.

Research and development of the fuzzy system were performed in Matlab 6 applications package using Fuzzy Logic Toolbox.

In accordance with research results, fuzzy knowledge base of the evaluation results classification was formed. For the creation of fuzzy inference system three linguistic variables were defined – information compactness index, stratification index and missing paths fraction. For fuzzy inference system, usage of the system of Mamdani type is suggested where the values of input and output variables are given by fuzzy terms [4]. The knowledge base consists of 19 fuzzy rules presented in Table 1.

Table 1

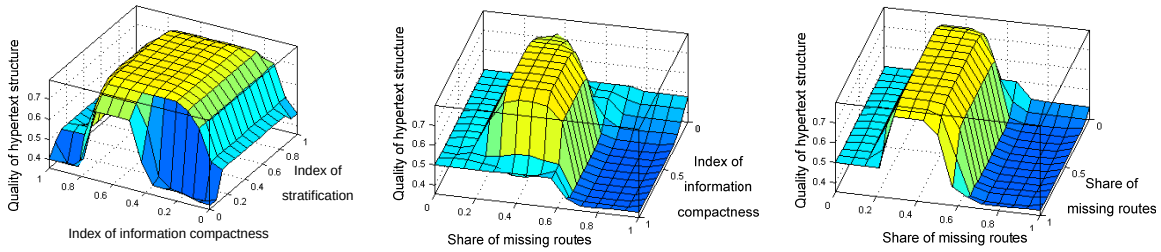
Fuzzy knowledge base rules

Stratification index	Information compactness index	Missing paths fraction	Hypertext structure quality
high	high	high	low
high	lower than average	average	average
average	high	low	low
average	high	low	high
average	high	average	low
average	higher than average	low	average
average	higher than average	average	average
average	higher than average	average	high
average	lower than average	low	low
average	lower than average	low	low
average	average	high	low
average	average	low	low
average	average	low	high
average	average	average	high
low	high	low	low
low	high	average	low
low	lower than average	average	
low	low	low	low
low	low	average	low

Visualization of the "inputs-outputs" surface is performed by means of Surface Viewer module from Matlab application package.

Fig. 5 shows "inputs-outputs" surfaces for output variable from the combination of input variables – stratification index, information compactness index and missing paths fraction. In accordance with the obtained results, a certain convex area corresponds to the term of «high» membership function of the output variable.

Visualizations shown in fig. 5 (a, b, c) confirm that the region of optimal solutions constitutes a certain set of them. The ranges of evaluation parameters variations for the optimal solutions region are as follows: [0.2; 0.85] for information compactness index, [0.1; 0.8] for stratification index and [0.25; 0.6] for missing routes fraction.



a) b) c)

Рис. 5. "Inputs-outputs" surface in SurfaceViewer: a) inputs – information compactness index, stratification index; b) inputs – missing paths fraction, stratification index; c) inputs – missing paths fraction, information compactness index

Conclusion. The developed system of fuzzy classification of Internet resources semantic structure on the basis of three indices (information compactness index, stratification index and missing paths fraction) makes it possible to consider the peculiarities of each index in structure classification. The developed rules can be used for the construction of automated system for evaluation of the semantic structure of sites using FuzzyJ toolbox that constitutes a set of libraries implementing fuzzy logic mechanisms for Java language.

REFERENCES

1. The Semantic Web [Електронний ресурс] / Tim Berners-Lee, James Hendler, Ora Lassila // Scientific American Magazine. – May, 2001. – Режим доступу до журн.: http://www.sciam.com/article.cfm?id=00048144-10D2-1C70-84A9809EC588EF21.

2. Botafogo R. A. Identifying hierarchies and useful metrics /E. Rivlin, B. Shneiderman // ACM Transactions on Information Systems (TOIS). – 1992. – №2. – P.142 – 180.

3. Harary F. Structural models. An Introduction to the Theory of Directed Graphs / Harary F., Norman R., Cartwright D. – Wiley: New York, 1965. – 415 p.

4. Ротштейн О. П. Интеллектуальные технологии идентификации: нечеткие множества, генетические алгоритмы, нейронные сети / Ротштейн О. П. – Вінниця: «УНІВЕРСУМ-Вінниця», 1999. – 320 с.

Dubovoy Vladimir – Dc. Sc. (Eng.), prof., head of Computer control systems department. phone: (0432) 598-157, E-Mail: dub@faksu.vstu.vinnica.ua.

Moskvin Olexiy – post-graduate student of Computer control systems department.
Vinnytsia National Technical University.