UDC 519.5

O.V. Glon, Cand. Sc. (Eng.); V.M. Dubuvoy, Dr. Sc. (Eng.), Prof.; O.M. Moskvin, Student

SITE STRUCTURE OPTIMIZATION IN CONDITIONS OF INCOMPLETE INFORMATION

The problem of semantic structure of hypertext model for organization of internet-resources are considered method and model intended for optimization of hypertext semantic structure in conditions of basic non-completeness of information regarding network structure are suggested.

Keywords: *hypertext, optimization, multiagent system, index of information density, index of stratification, graph, graph cyclomatic number, graph base, non-completeness of information, optimality, web-resource.*

Due to mass computerization and introduction of new information technologies, Internet is nowadays one of the most important information sources. Large websites amount and their comparatively low quality, make the process of the information search rather slow and difficult. One of the efficient approaches for information representation on website is hypertext links usage. Hypertext information model is determined as structure for efficient presentation and knowledge transfer [1]. Hidden information included in hyperlinks has net structure and brings additional information contained both in related views and link structure. The distributed or circular hypertext structure related to site structure hinders information search. Thus the problem of semantic information structure optimization is a topic one.

In the scientific aspect one should point out the problem of the necessity of the task solution optimization on conditions of principle information imperfection about the global hypertext structure in sites network. As a rule, only a subset of restricted ties is known, for the rest ones only the expert and statistical estimations exist.

Creating the site structure there are investigated linear, latticed and hierarchical structures [3], which are characterized by depth [2]. Usually, the navigation depth from one to four levels is considered to be optimal levels (greater quantity complicates information search). But this approach does not take into account the hypertext interrelation.

To solve a problem, there has been offered Semantic Web concept [7] – a part of global concept of internet network evolution, the goal of which is opportunity implementation of machine information processing. The metadata, that characterize the properties and content of Internet resources instead of currently used document text analyses are emphasized.

The term was first introduced by Tim Berners-Lee in «Scientific American», 2001 [6]. The semantic web is characterized by the usage of universal resource identifier (URI) and metadata description languages.

This concept was accepted and promoted by W3 consortium [6]. For the realization of this concept the creation of document network that contains global web metadata was suggested.

In the hypertext theory for the formalization of its functionally meaningful parameters exists a special hypertext metrics certificate which includes two basic parameters: degree of informative compactness and index of stratification [1]. A high level of compactness characterizes such hypertext structures in which it is easy to get from one informative block into the other (usually it is provided by numerous cross references). Very high compactness can result in complete disorientation of the hypertext user appealed, and also complicates the process of watching of heredity of concepts. A low informative compactness fails to retain in readers sight separate hypertexts knots that carry important information for forming the defined notions, or in many cases to make separate knots accessible.

Stratification index allows to estimate possible choice freedom level of hypertext document reading sequence. But the formal model for estimating semantic hypertext structure and universally

recognized algorithm actually does not exist.

The article is aimed at the formation of approaches to semantic sites structure optimization process. To solve this task, let's show a site structure as a graph. We describe the graph by indexes which allow to define the efficiency of its structure.

The index of informative compactness is calculated using the following formula [1]:

$$Cp = \frac{Max}{Max - Min}, \quad (1)$$

where Max - maximal possible steps amount, to be made between links that unite all hypertext nodes; Min - minimal possible steps amount, that unites all hypertext nodes (in case, all hypertext nodes are united)

Maximal and minimal steps amount is calculated for all base vertexes (the notion of a base vertex is considered below). Actually observed steps number can be evaluated considering probability of path choice between vertexes, taking into account the equality of switching probabilities for each hyperlink.

Stratification index is closely connected with the graph cyclomatic number. Indeed, if the graph is a tree, only one track exists between each vertex pairs. The cyclomatic number characterizes graph structure difference from tree-like structure. Any subgraph that contains all vertexes of graph G and is a tree can be referred to the spanned tree of the constrained graph G. If G - is a constrained graph with $n(G)$ vertexes and $m(G)$ ribs, then spanned tree of G graph (if it exists) should contain $n(G)-1$ ribs. Consequently, any spanned tree of the G graph is a result of removing $m(G)-(n(G)-1)=m(G)-n(G)+1$ ribs from the graph. The number $\nu(G)=m(G)-n(G)+1$ is a cyclomatic number of the constrained graph G [5].

The following hypothesis forms the basis of site structure optimization system creation: there exists an optimal hypertext structure complexity [8] (reduced to vertex number cyclomatic number C_n/m, where m - is a vertexes number; Cp - compactness informative index).

The optimization system should meet the following requirements:

- the storage of hypertext fragments attainability;
- functioning in conditions of incomplete information about network structure;
- optimality on average;
- adaptation to intellectual and psychological characteristics of the user.

Attainability storage.

The vertex w of directed graph D, is called attainable from vertex ν, if $w=\nu$ or there exists a route that joins ν and w. The vertex attainability is described by matrix $A_G(v,w):\{a_{vw}=1$ only if there exists a route from v to $w\}$.

A graph (directed graph) is called connected, if there exists a route that joins any of its parts. A directed graph is called one-side connected, if for each of its two vertexes at least one is attainable from another.

Functioning in conditions of incomplete information

The optimization system should work in conditions of full information lack about semantic hypertext structure. It is explained by the high dimension of sites network, its constant growing and modification, that makes impossible full information collection.

One can rely only on the information about site structure and structure of adjacent sites.

The search of graph base and setting the links importance level form the basis of optimization algorithm in conditions of incomplete information.

Let's determine a strong attainability graph for the graph base search. A strong attainability graph $G_*^*=(V,E_*^*)$ for G has the same V vertex multitude and ribs multitude $E_*^*=\{(u,v)\mid$ v and u are reciprocally attainable$\}$ [5].

It follows from the definition of attainability and strong attainability that for all (i,j),

$1 \le i, j \le n$, pairs, a value of strong attainable matrix element $A_{G*}(i, j)$ equals 1 only when both $A_G(i, j)$ and $A_G(j, i)$ are equal to 1, i.e.

$$A_{G*}(i, j) = A_G(i, j) \wedge A_G(j, i), \tag{2}$$

Strong attainable components of graph G can be defined using A_{G*} in the following way:

1. Place a vertex v_1 and vertexes v_i for which $A_{G*}(1, j) = 1$ in K_1 multiplier.

$K_1, \ldots, K_i$ are already build multipliers, v_k - vertex with minimal number, that hasn't got to multipliers. Place v_k and vertexes v_i for which $A_{G*}(1, j) = 1$ in K_{i+1}.

Repeat step (2) until all vertexes become allocated to multipliers.

Let K and K' be multipliers of strong attainability of graph G. K multiplier is attainable from K', if $K = K'$ or there exist two vertexes $u \in K$ and $v \in K'$ where u is attainable from v.

K is strictly attainable from K', if $K \neq K'$ and K is attainable from K'. K multiplier is called minimal, if it is not strictly attainable from any of the multipliers.

A vertexes subset $W \subseteq V$ is called generative in case all graph vertexes are attainable from W vertexes. A vertexes subset $W \subseteq V$ is called a base of the graph, if it's generative, but none of its own subsets is generative.

Hence it follows the procedure of G graph bases construction.

1. Find all tie-up components of G.
2. Define an order on them and select minimum components with regard this order.
3. Generate one or all graph bases, choosing one vertex from every minimum component.

Ribs are being marked after selection of graph base. Rib weight (v,u) is determined by expression

$$\rho_{vu} = \min[l_v, l_u] \cdot \max[P_{vu}, P_{uv}], \tag{3}$$

where l_i – distance from i vertexes to the nearest generative vertex; P_{ij} – statistical estimation probability of i vertex visit through hyperlink from vertex j.

During graph relations optimization according to information compactness criteria and index of stratification, we remove references (graph ribs) that have the least weight.

Optimality on average is caused by statistical approach for determining the importance (weight (3)) of hyperlinks, and clarifying probability estimations during compactness index (1) calculation.

Adaptation to intellectual and psychological characteristics of the user is carried out by setting individual optimum of informative compactness index and index of stratification, that are evaluated using statistical analysis of user sessions in the Internet.

It is suggested to use multiagent technology [7,9] and develop appropriate agent, to provide analysis and site structure optimization.

Program-agents should be located on web servers. During user navigation between resources agents exchange information about site structures. On the basis of this information optimization and temporary links deactivation with the least weight are carried out.

Handling of incoming query (figure 1) foresees the analysis of its parameters that contain service information about a client, including source address from where the user has come (Referrer resource). The analysis of internal structure, external links and information indexing are provided automatically in case of referrer address presence in query service headers. Information search systems, catalogs, rating systems are excluded from the analyzed ones. Besides, it's possible to provide in-depth analysis of resource up to administrator specified level. A result of the incoming query processing, in case of referrer address presence, is processed and saved into database information, to be used further for site structure and external links optimization.

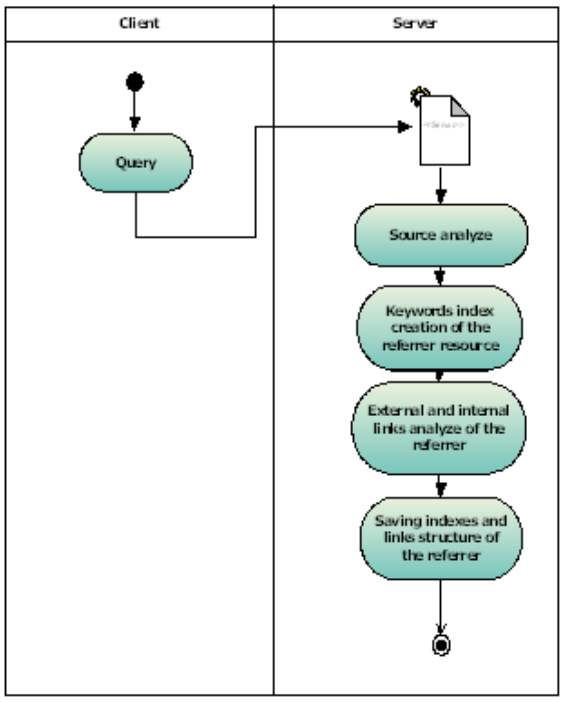

Figure 1 – UML diagram of handling incoming queries

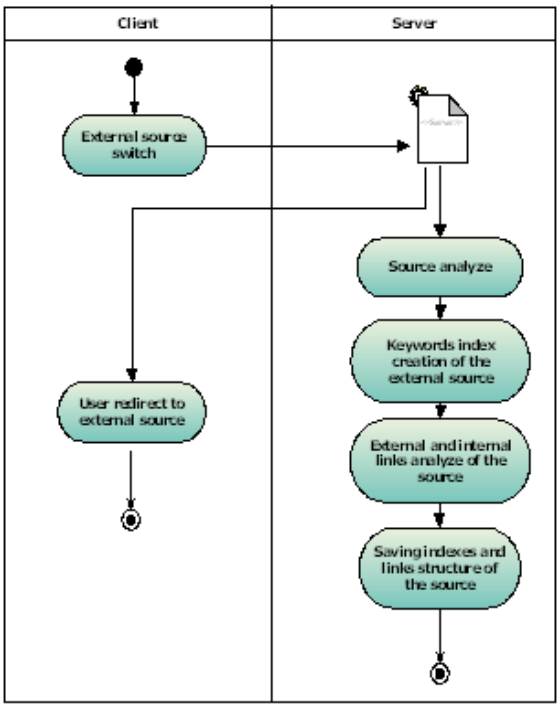

Figure 2 – UML diagram of handling outgoing requests

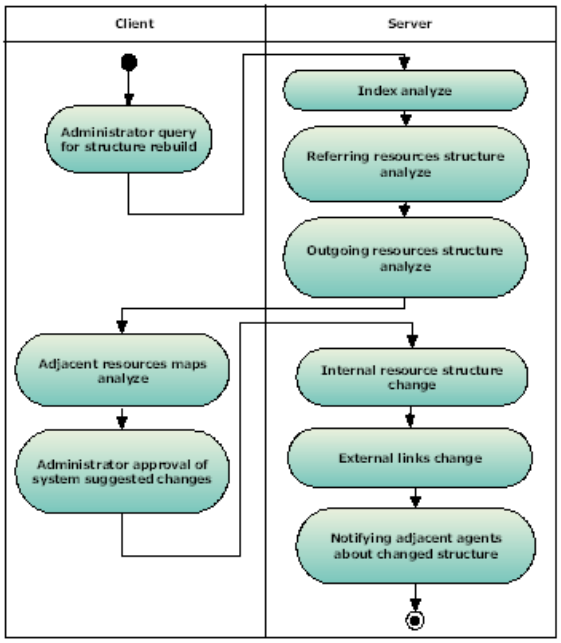

Figure 3 – UML diagram of handling collected data and resource structure modification

Outgoing request, the layout of which is displayed on figure 2, consists of final resource analyzing and user redirection to this resource. Final resource analyzing foresees it's indexation or re-indexation., external and internal links analyzing. Processing of collected data (figure 3) initiated by the administrator includes: processing and analyzing indexes, internal and external links thus to make a decision about site structure modification. Modification results in updating resource structure and its transferring to adjacent site agents.

Conclusion

Suggested approach to informative resource structure optimization provides saving of optimal attainability of hypertext fragments in conditions of incomplete information.

The essence of the approach lies in presentation of informative resource as a graph, where the ribs are links that join hypertext. For the resource structure we suggest hypothesis about optimal complexity, which is expressed by the cyclomatic number of the vertex number. Structure optimization in conditions of incomplete information is provided by searching graph base and setting links importance levels, on which the modification of their structure and handling of informative compactness level is based on.

Intellectual multiagent technology usage allows to make optimization process automatized, that will result in increasing efficiency utilization of Internet resources.

REFERENCES

1. Методы оптимизации компьютерной обучающей среды по лингвистике для систем дистанционного обучения в Интернете [Електронний ресурс] / Кедрова Г. Е. // Материалы научно-практической конференции "Эффективность использования новых информационных технологий в учебном процессе" (ЭНИТ-2000)". - Ульяновск, 2000 – Режим доступу: http://www.philol.msu.ru/~kedr/kedr-ulj.htm

2. Иллюстрация понятия "глубина сайта" [Електронний ресурс] // Профессиональная студия веб-дизайна "Антула". – Москва. – Режим доступу: http://www.antula.ru/deep-sait.htm.

3. Оценка надежности сайта. критерии надежности сайта [Електронний ресурс] // Профессиональная студия веб-дизайна "Антула". – Москва. – Режим доступу.: http://www.antula.ru/web-design_safe.htm.

4. Семантическая паутина [Електронний ресурс] // Википедия. – Режим доступу: http://ru.wikipedia.org/wiki/Семантическая паутина

5. Основы дискретной математики [Електронний ресурс] / Дехтярь М. И. // Интернет Университет Информационных Технологий. – 08.2007. – Режим доступу: http://www.intuit.ru/ department/ds/discrmath/9/.

6. 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.

7. Новиков Д. А. Сетевые структуры и организационные системы. – М.: ИПУ РАН, 2003. –102 с.

8. Губко М. В. Математические модели оптимизации иерархических структур. – М.: ИПУ РАН, 2006. – 264 с.

9. Jabadie A., Lin J., Morse A. Coordination of groups of autonomous agents using nearest neighbor rules // IEEE Trans. – 2003. – Vol. AC-48, № 6. – P. 988-1001.

Dubovoy Volodimir Michailovich – head of the department computer controlled systems;

Glon Olga Vitaliivna – docent of the department computer controlled systems;

Moskvin Oleksiy Michailovich – student of the department computer controlled systems.
Vinnytsia national technical university.