УДК 519.5

**О. В. Глонь, к. т. н.; В. М. Дубовой, д. т. н., проф.; А. М. Москвин, студ.**

# ОПТИМИЗАЦИЯ СТРУКТУРЫ САЙТА В УСЛОВИЯХ НЕПОЛНОЙ ИНФОРМАЦИИ

*В статье рассмотрены проблемы семантической структуры гипертекстовой модели при организации интернет-ресурсов. Отмечена проблема необходимости её оптимизации. Предложены метод и модель оптимизации семантической структуры гипертекста в условиях принципиальной неполноты информации о структуре сети.*

***Ключевые слова:*** *гипертекст, оптимизация, мультиагентная система, индекс информационной компактности, индекс стратификации, граф, цикломатическое число графа, база графа, досягаемость, неполнота информации, оптимальность, web-ресурс.*

Благодаря массовой компьютеризации и распространению новых информационных технологий, Интернет является одним из важнейших источников информации. Но большое количество веб-сайтов и их относительно низкое качество осложняет и замедляет процесс поиска необходимых сведений. Один из подходов эффективного представления информации на сайте – использование гипертекстовых ссылок. Гипертекстовая информационная модель получает все большее признание как структура для эффективного представления и передачи знаний [1]. Скрытая информация, содержащаяся в гиперссылках, имеет сетевую структуру. Такая структура несет дополнительную информацию, которая содержится как в связанных высказываниях, так и в структуре связи. Слишком разветвленная или круговая структура гипертекста, непосредственно связанная со структурой веб-сайта, мешает поиску необходимой информации. В таких условиях особенную **актуальность** приобретает проблема оптимизации семантической структуры информации.

В научном аспекте отмечена проблема, связанная с необходимостью решения задачи оптимизации в условиях принципиальной неполноты информации о глобальной структуре гипертекста в сети сайтов. Как правило, известным является ограниченное подмножество связей, а относительно остальных существуют лишь экспертные и статистические оценки.

При создании структуры сайта рассматривают линейную, решетчатую и иерархическую структуры [3], которые характеризуются глубиной [2]. Обычно, считается оптимальной глубина навигации от одного до четырех уровней (большее количество очень усложняет поиск на этом уровне информации). Но такой подход не учитывает гипертекстовую взаимосвязь.

Для решения проблемы предложена концепция "*семантическая паутины*" (англ. Semantic Web) [7] – часть глобальной концепции развития сети Интернет, целью которой является реализация возможности машинной обработки информации. Основной акцент концепции поставлен на работе с метаданными, которые однозначно характеризуют свойства и содержание ресурсов Интернет, вместо используемого в настоящее время текстового анализа документов. Понятие впервые введено сэром Тимом Бернесом-Ли в 2001 году в журнале "Scientific American" [6]. В семантической паутине предусматривается повсеместное использование, во-первых, универсальных идентификаторов ресурсов (URL), во-вторых, онтологии и языков описания метаданных.

Данная концепция была принята и продвигается Консорциумом W3 [6]. Для ее внедрения предусматривается создание сети документов, содержащих метаданные о ресурсах Всемирной паутины.

В теории гипертекста для формализации его функционально значимых параметров была разработана специальная гипертекстовая метрика, которая включает два базовых параметра: степень информационной компактности и индекс стратификации [1]. Высокий уровень компактности характеризует такие гипертекстовые структуры, в которых на любой из

информационных блоков можно с легкостью попасть из любого другого блока (обычно это обеспечивается многочисленными перекрестными ссылками). Излишне высокая компактность может привести к полной дезориентации пользователя, обратившегося к данному гипертексту, а также чрезвычайно осложняет процесс отслеживания наследственности понятий. Низкая информационная компактность чревата выпадением из поля зрения читателя гипертекста отдельных узлов, которые могут нести важную для формирования определенных понятий информацию, или вообще делать отдельные узлы во многих случаях недоступными. Индекс стратификации позволяет оценить допустимую степень свободы выбора последовательности чтения гипертекстового документа. Но формальной модели для оценивания семантической структуры гипертекста и общепризнанного алгоритма не существует.

Целью статьи является формулировка подходов формализации процесса оптимизации семантической структуры сайтов.

Для решения поставленной задачи структуру сайта представим в виде графа. Охарактеризуем граф показателями, которые позволяют определить эффективность его структуры.

Индекс информационной компактности вычисляется по формуле [1]:

$$Cp = \frac{Max}{Max - Min}, \tag{1}$$

где $Max$ – максимально возможное число шагов, которые необходимо пройти по ссылкам, которые связывают все узлы гипертекста.

$Min$ – минимальное возможное число шагов, которое связывает все узлы гипертекста (в том случае, когда все узлы гипертекста связаны со всеми).

Максимальное и минимальное числа шагов находятся для всех базовых вершин (понятие базовой вершины рассмотрено ниже).

Реально наблюдаемое число шагов может быть рассчитано с учетом вероятности выбора пути между вершинами, считая в первом приближении вероятности переходов по каждому из гиперссылок страницы равными.

Индекс стратификации тесно связан с цикломатическим числом графа. Действительно, если граф является деревом, то существует лишь один путь между каждой парой вершин. Цикломатическое число характеризует отличие структуры графа от древовидной.

Остовным деревом связанного графа $G$ называется любой его подграф, который содержит все вершины графа $G$ и является деревом. Если $G$ – связанный граф, содержащий $n(G)$ вершин и $m(G)$ ребер, то остовное дерево графа $G$ (если оно существует) должно иметь $n(G)-1$ ребер.

Таким образом, любое остовное дерево графа $G$ является результатом удаления из графа $m(G)-(n(G)-1)=m(G)-n(G)+1$ ребер. Число $\nu(G)=m(G)-n(G)+1$ называется цикломатическим числом связанного графа $G$ [5].

В основу создания системы оптимизации структуры сайта положена гипотеза: существует оптимальная сложность структуры гипертекста (приведено к числу вершин цикломатическое число $C_n/m$, где m – количество вершин; индекс информационной компактности $Cp$).

Система оптимизации должна удовлетворять требованиям:

– сохранение достижимости фрагментов гипертекста;

– функционирование в условиях неполной информации о структуре сети;

– оптимум в среднем;

– адаптация к интеллектуально-психологическим особенностям пользователя.

*Сохранение достижимости.*

Вершина $w$ орграфа $D$ называется *достижимой* из вершины $\nu$, если $w=\nu$, или существует маршрут, соединяющий $\nu$ и $w$.

Достижимость вершин описывается матрицей $A_G(v,w):\{a_{vw}=1$ тогда и только тогда, когда существует маршрут из $v$ в $w\}$.

Граф (орграф) называется связным, если для любых его вершин существует маршрут (путь), который их связывает. Орграф называется *односторонне связным*, если для любых двух его вершин по крайней мере одна достижима из другой.

*Функционирование в условиях неполной информации.*

Система оптимизации должна работать в условиях, когда отсутствует полная информация о семантической структуре гипертекста. Это связано с большой размерностью сети сайтов и ее постоянным увеличением и модификацией, которая делает невозможным сбор полной информации. Можно надеяться лишь на информацию относительно структуры самого сайта, который оптимизируется, а также, возможно, относительно структуры смежных сайтов.

В основе алгоритма оптимизации структуры графа в условиях неполной информации лежит поиск базы графа и установления уровня важности связей.

Для поиска базы графа определим граф сильной достижимости. *Граф сильной достижимости* $G_*^* = (V, E_*^*)$ для $G$ имеет множество вершин $V$ и множество ребер $E_*^* = \{(u,v) |\ v$ и $u$ взаимнодостижимые$\}$ [5].

Из определения достижимости и сильной достижимости непосредственно следует, что для всех пар $(i,j)$, $1 \le i,j \le n$, значение элемента матрицы сильной достижимости $A_{G_*^*}(i,j)$ равно 1 тогда и только тогда, когда оба элемента $A_G(i,j)$ и $A_G(j,i)$ равны 1, т.е.:

$$A_{G_*^*}(i,j) = A_G(i,j) \wedge A_G(j,i). \qquad (2)$$

По матрице $A_{G_*^*}$ можно выделить компоненты сильной связанности графа $G$ следующим образом:

1. Поместим в компоненту $K_1$ вершину $v_1$ и все такие вершины $v_i$, для которых $A_{G_*^*}(1,j)=1$.

2. Пусть уже построенные компоненты $K_1,\ldots,K_i$ и $v_k$ – вершина с минимальным номером, которая еще не попала в компоненты. Тогда поместим в компоненту $K_{i+1}$ вершину $v_k$ и все такие вершины $v_i$, для которых $A_{G_*^*}(k,j)=1$.

3. Повторяем шаг (2) до тех пор, пока все вершины не будут распределены по компонентам.

Пусть $K$ и $K'$ – компоненты сильной связности графа $G$. Компонента $K$ достижима из компоненты $K'$, если $K=K'$, или существуют такие две вершины $u \in K$ и $v \in K'$, для которых $u$ достижима из $v$. $K$ строго достижима из $K'$, если $K \ne K'$, и $K$ достижима из $K'$. Компонента $K$ называется минимальной, если она не является строго достижимой ни из какой компоненты. Подмножество вершин $W \subseteq V$ называется порождающим, если из вершин $W$ можно достичь любой вершины графа. Подмножество вершин $W \subseteq V$ называется базой графа, если оно является порождающей, но никакое его собственное подмножество не является порождающим.

Подмножество вершин $W \subseteq V$ является базой $G$ только тогда, когда содержит по одной вершине из каждой минимальной компоненты сильной связности $G$ и не содержит никаких других вершин.

Отсюда вытекает следующая процедура построения всех баз графа $G$:

1. Найти все компоненты связности $G$.

2. Определить порядок на них и выделить минимальные относительно этого порядка компоненты.

3. Породить одну или все базы графа, выбирая по одной вершине из каждой

минимальной компоненты.

После выбора базы графа размечаются ребра. Вес ребра $(v,u)$ определяется выражением:

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

где $l_i$ – расстояния от вершин $i$ до ближайшей порождающей вершины;
$P_{ij}$ – статистическая оценка вероятности посещения вершины $i$ через гиперссылки из вершины $j$.

При оптимизации связей графа по критериям информационной компактности и индекса стратификации будем изымать ссылки (ребра графа), имеющие наименьший вес.

*Оптимум в среднем* предопределен статистическим подходом к определению важности (веса (3)) гиперссылок, а также постепенным уточнением оценок вероятностей переходов при расчете индекса компактности (1).

*Адаптация к интеллектуально-психологическим особенностям пользователя* осуществляется путем установления индивидуального оптимума показателя информационной компактности и индекса стратификации, которые определяются на основе статистического анализа сеансов работы пользователя в Интернет.

Предлагается использовать мультиагентную технологию [7, 9] и разработать соответствующий агент, который бы мог анализировать и оптимизировать структуру сайта.

Программы-агенты размещаются на web-серверах. При переходах пользователя из сайта на сайт агенты обмениваются информацией относительно структуры сайтов. На основе этой информации осуществляется оптимизация структуры сайтов и временная деактивация связей с наименьшим весом.

Обработка входного запроса (рис. 1) предусматривает анализ его параметров, которые содержщих служебную информацию о клиенте, в частности, адрес источника, из которого был осуществлен переход (Referrer). В случае наличия такого адреса,

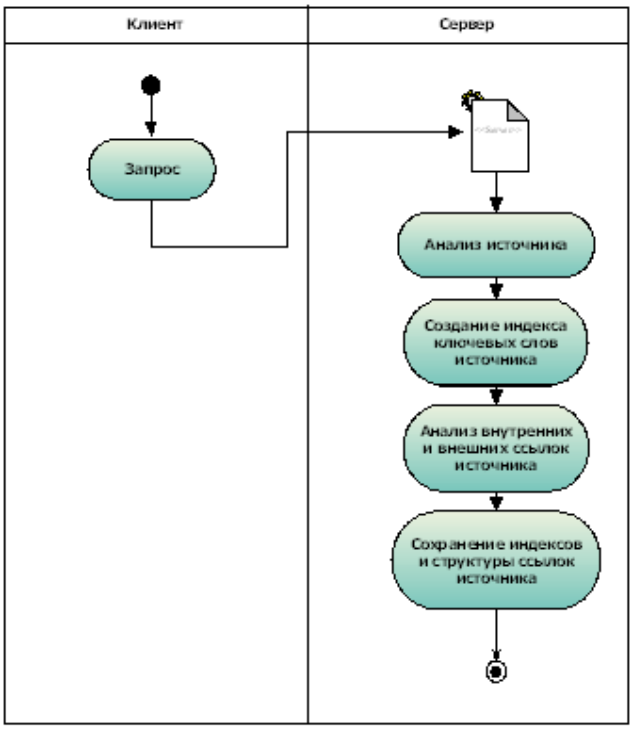


Рис. 1. UML диаграмма обработки входящих запросов

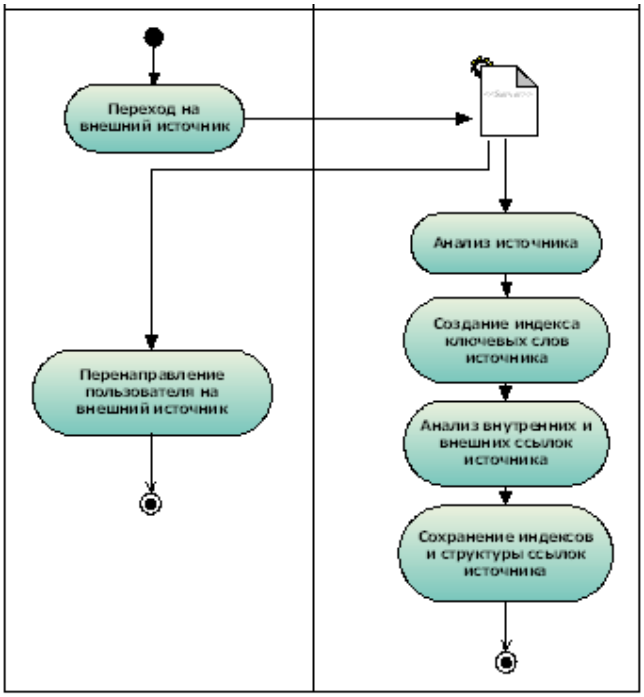


Рис. 2. UML диаграмма обработки исходящих запросов

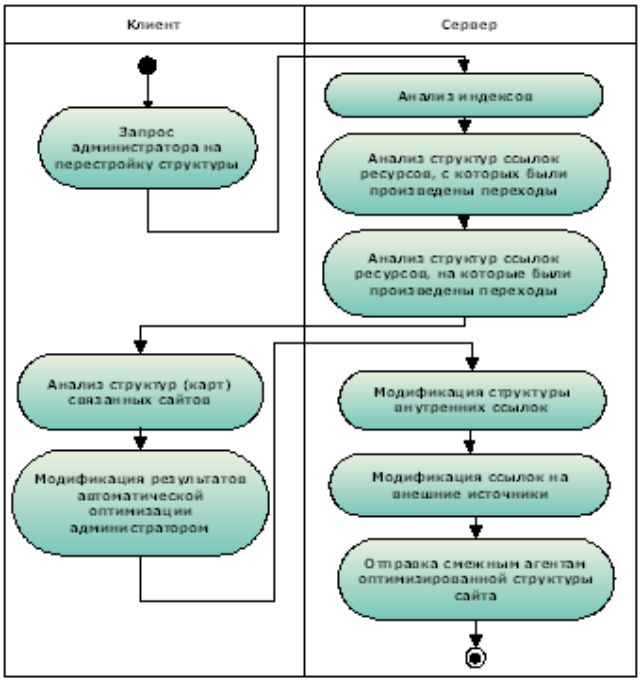


Рис. 3. UML диаграмма обработки собранных данных и модификации структуры ресурса

автоматически происходит анализ источника, размещенного по данному адресу, его индексация, анализ структуры, анализ внешних и внутренних ссылок. Исключения в данном случае составляют поисковые системы, тематические каталоги, системы оценки рейтинга. Кроме того возможно проведение анализа ресурса в глубину, к определенному установленному администратором уровню.

Результатом обработки и анализа входного запроса, при условии наличия Referrer адреса ресурса, является обработанная и сохраненная в базе данных информация о данном ресурсе, который будет использован для оптимизации структуры сайта и ссылок на внешние источники.

Исходный запрос, схема которого изображена на рис. 2, состоит из процедуры анализа конечного ресурса и перенаправления на него пользователя. Анализ конечного ресурса предусматривает его индексацию или реиндексацию в случае изменения данных, анализ структуры, внешних и внутренних ссылок.

Обработка собранных данных (рис. 3) инициируется администратором и включает в себя: анализ индексов, анализ структур внешних ссылок проанализированных ресурсов. На основе этого принимается решение о модификации структуры сайта.

Результатом модификации структуры ресурса является обновление структуры сайта и передача ее агентам смежных сайтов.

## Выводы

Предложен подход к оптимизации структуры информационного ресурса, который в условиях неполной информации о структуре сети обеспечивает сохранение оптимальной достижимости фрагментов гипертекста.

Сущность подхода заключается в представлении структуры информационного ресурса в виде графа, ребрами которого являются ссылки, связывающие гипертекст. Для структуры ресурса принимается гипотеза об оптимальной сложности, которая численно выражается приведенным к числу вершин цикломатическим числом. Оптимизация структуры в условиях неполной информации осуществляется путем поиска базы графа и установкой уровней важности связей, на основе чего модифицируется структура и как следствие происходит управление уровнем информационной компактности.

Использование интеллектуальной мультиагентной технологии позволяет автоматизировать процесс оптимизации, которая будет способствовать повышению эффективности использования Интернет-ресурсов.

## СПИСОК ЛИТЕРАТУРЫ

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.

***Дубовой Владимир Михайлович*** – заведующий кафедрой компьютерных систем управления;

***Глонь Ольга Витальевна*** – доцент кафедры компьютерных систем управления;

***Москвин Алексей Михайлович*** – студент кафедры компьютерных систем управления.
Винницький национальный технический университет.