УДК 519.5

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

# ОПТИМІЗАЦІЯ СТРУКТУРИ САЙТА В УМОВАХ НЕПОВНОЇ ІНФОРМАЦІЇ

*Розглянуто проблеми семантичної структури гіпертекстової моделі, при організації інтернет-ресурсів. Зазначено проблему необхідності її оптимізації. Запропоновано метод та модель оптимізації семантичної структури гіпертексту в умовах принципової неповноти інформації про структуру мережі.*

***Ключові слова:*** *Гіпертекст, оптимізація, мультиагентна система, індекс інформаційної компактності, індекс стратифікації, граф, цикломатичне число графа, база графа, досяжність, неповнота інформації, оптимальність, web-ресурс.*

Завдяки масовій комп'ютеризації та поширенню нових інформаційних технологій, Інтернет є одним з найважливіших джерел інформації. Але велика кількість веб-сайтів та їх відносно низька якість ускладнює і сповільнює процес пошуку необхідних відомостей. Один із підходів ефективного подання інформації на сайті – це використання гіпертекстових посилань. Гіпертекстова інформаційна модель отримує все більше визнання як структура для ефективного представлення та передачі знань [1]. Прихована інформація, що міститься у гіперпосиланнях, має мережну структуру, яка несе додаткову інформацію, що міститься як у зв'язаних висловлюваннях, так і у структурі зв'язку. Занадто розгалужена або кругова структура гіпертексту, що безпосередньо пов'язана зі структурою веб-сайту, заважає пошуку необхідної інформації. В таких умовах набуває особливої **актуальності** проблема оптимізації семантичної структури інформації.

У науковому аспекті зазначена проблема, пов'язана з необхідністю розв'язання задачі оптимізації в умовах принципової неповноти інформації про глобальну структуру гіпертексту у мережі сайтів. Як правило, відомими є обмежена підмножина зв'язків, а щодо решти існують лише експертні та статистичні оцінки.

При створенні структури сайту розглядають лінійну, решітчасту та ієрархічну структури [3], які характеризують глибиною [2]. Зазвичай вважається оптимальною глибина навігації від одного до чотирьох рівнів (більша кількість дуже ускладнює пошук на цьому рівні інформації). Але такий підхід не враховує гіпертекстовий взаємозв'язок.

Для розв'язання проблеми запропонована концепція "*семантична павутина*" (англ. *Semantic Web*) [7] – частина глобальної концепції розвитку мережі Інтернет, метою якої є реалізація можливості машинної обробки інформації. Основний акцент концепції ставиться на роботі з метаданими, що однозначно характеризують властивості і зміст ресурсів Інтернет, замість використовуваного сьогодні текстового аналізу документів. Термін вперше введений сером Тімом Бернесом-Лі в 2001 році в журналі «Scientific American» [6]. У семантичній павутині передбачається повсюдне використання, по-перше, універсальних ідентифікаторів ресурсів (URL), по-друге онтологій і мов опису метаданих.

Ця концепція була прийнята і впроваджується Консорціумом W3 [6]. Для її впровадження передбачено створення мережі документів, що містять метадані про ресурси Всесвітньої павутини.

У теорії гіпертексту для формалізації його функціонально значущих параметрів була розроблена спеціальна гіпертекстова метрика, що містить два базових параметри: ступінь інформаційної компактності й індекс стратифікації [1]. Високий рівень компактності характеризує такі гіпертекстові структури, в яких на будь-який з інформаційних блоків можна легко потрапити із будь-якого іншого блока (зазвичай це забезпечується численними перехресними посиланнями). Надмірно висока компактність може призвести до повної дезорієнтації користувача, що звернувся до даного гіпертексту, а також надзвичайно ускладнює процес відстеження спадковості понять. Низька інформаційна компактність чревата випаданням із поля зору читача гіпертексту окремих вузлів, які можуть нести важливу для формування певних понять інформацію, або взагалі робити окремі вузли в багатьох випадках недоступними. Індекс стратифікації дозволяє оцінити допустимий ступінь свободи вибору послідовності читання гіпертекстового документа. Але формальної моделі для оцінювання семантичної структури гіпертексту та загальновизнаного алгоритму не існує.

**Метою статті** є формулювання підходів формалізації процесу оптимізації семантичної структури сайтів.

Для розв'язання поставленої задачі структуру сайта представимо у вигляді графа. Охарактеризуємо граф показниками, що дозволяють визначити ефективність його структури.

Індекс інформаційної компактності обчислюється за формулою [1]:

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

де $Max$ – максимально можливе число кроків, які необхідно пройти по посиланням, що зв'язують усі вузли гіпертексту; $Min$ – мінімальне можливе число кроків, які зв'язують усі вузли гіпертексту (у тому разі, коли усі вузли гіпертексту зв'язані з усіма).

Максимальне і мінімальне числа кроків знаходяться для всіх базових вершин (поняття базової вершини розглянуте нижче).

Реально спостережуване число кроків може бути розраховане з урахуванням імовірності вибору шляху між вершинами, вважаючи у першому наближенні ймовірності переходів по кожному з гіперпосилань сторінки рівними.

Індекс стратифікації тісно пов'язаний із цикломатичним числом графа. Дійсно, якщо граф є деревом, то існує лише один шлях між кожною парою вершин. Цикломатичне число характеризує відмінність структури графа від деревоподібної.

Остовним деревом зв'язного графа $G$ називається будь-який його підграф, що містить усі вершини графа $G$ і є деревом. Якщо $G$ – зв'язний граф, що містить $n(G)$ вершин і $m(G)$ ребер, то остовне дерево графа $G$ (якщо воно існує) повинно містити $n(G)-1$ ребер.

Отже, будь-яке остовне дерево графа $G$ є результатом видалення із графу $m(G)-(n(G)-1)=m(G)-n(G)+1$ ребер. Число $\nu(G)=m(G)-n(G)+1$ називається цикломатичним числом зв'язного графа $G$ [5].

В основу створення системи оптимізації структури сайта покладена гіпотеза: існує оптимальна складність структури [8] гіпертексту (приведене до числа вершин цикломатичне число $C_n/m$, де $m$ – кількість вершин; $C_n$ – індекс інформаційної компактності).

Система оптимізації повинна задовольняти вимоги:

– збереження досяжності фрагментів гіпертексту;

– функціонування в умовах неповної інформації про структуру мережі;

– оптимальність у середньому;

– адаптація до інтелектуально-психологічних особливостей користувача.

*Збереження досяжності.*

Вершина $w$ орграфа $D$ називається *досяжною* із вершини $v$, якщо $w=v$, або існує маршрут, що з'єднує $v$ і $w$.

Досяжність вершин описується матрицею $A_G(v,w):\{a_{vw}=1$ тоді і тільки тоді, коли існує маршрут із $v$ в $w\}$.

Граф (орграф) називається зв'язним, якщо для будь-яких його вершин існує маршрут (шлях), який їх зв'язує. Орграф називається *односторонньо зв'язним*, якщо для будь-яких двох його вершин принаймні одна досяжна з іншої.

*Функціонування в умовах неповної інформації.*

Система оптимізації повинна працювати в умовах, коли відсутня повна інформація про семантичну структуру гіпертексту. Це пов'язане з великою розмірністю мережі сайтів та її постійним збільшенням і модифікацією, що унеможливлює збір повної інформації. Можна сподіватися лише на відомості щодо структури самого сайту, що оптимізується, а також, можливо, щодо структури суміжних сайтів.

В основі алгоритму оптимізації структури графа в умовах неповної інформації лежить пошук бази графа і встановлення рівня важливості зв'язків.

Для пошуку бази графа визначимо граф сильної досяжності. *Граф сильної досяжності* $G_*^* = (V, E_*^*)$ для $G$ має таку ж множину вершин $V$ і наступну множину ребер $E_*^* = \{(u,v) \mid v$ і $u$ взаємо досяжні$\}$ [5].

Із визначення досяжності і сильної досяжності безпосередньо випливає, що для всіх пар $(i,j)$, $1 \le i,j \le n$ значення елемента матриці сильної досяжності $A_{G_*^*}(i,j)$ дорівнює 1 тоді і тільки тоді, коли обидва елементи $A_G(i,j)$ і $A_G(j,i)$ рівні 1, тобто:

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

За матрицею $A_{G_*^*}$ можна виділити компоненти сильної зв'язності графа $G$ таким чином:

1. Помістимо в компоненту $K_1$ вершину $v_1$ і всі такі вершини $v_i$, що $A_{G_*^*}(1,j) = 1$.

2. Нехай уже побудовані компоненти $K_1, \ldots, K_i$ і $v_k$ – вершина з мінімальним номером, яка ще не потрапила до компонентів. Тоді помістимо в компоненту $K_{i+1}$ вершину $v_k$ і всі такі вершини $v_i$, що $A_{G_*^*}(k,j) = 1$.

3. Повторюємо крок (2) доти, доки всі вершини не будуть розподілені по компонентах.

Нехай $K$ і $K'$ – компоненти сильної зв'язності графа $G$. Компонента $K$ досяжна із компоненти $K'$, якщо $K = K'$, або існують такі дві вершини $u \in K$ і $v \in K'$, що $u$ досяжна із $v$. $K$ строго досяжна із $K'$, якщо $K \ne K'$, і $K$ досяжна із $K'$. Компонента $K$ називається мінімальною, якщо вона не є строго досяжною ні із якої компоненти. Підмножина вершин $W \subseteq V$ називається породжуючою, якщо із вершин $W$ можна досягти до будь якої вершини графа. Підмножина вершин $W \subseteq V$ називається базою графа, якщо вона є породжуючою, але ніяка його власна підмножина не є породжуючою.

Підмножина вершин $W \subseteq V$ є базою $G$ тоді і тільки тоді, коли містить по одній вершині із кожної мінімальної компоненти сильної зв'язності $G$ і не містить ніяких інших вершин.

Звідси витікає така процедура побудови всіх баз графа $G$:

1. Знайти всі компоненти зв'язності $G$.
2. Визначити порядок на них і виділити мінімальні відносно цього порядку компоненти.
3. Породити одну або всі бази графа, вибираючи по одній вершині із кожної мінімальної компоненти.

Після вибору бази графа розмічаються ребра. Вага ребра $(v,u)$ визначається за виразом:

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

де $l_i$ – відстані від вершин $i$ до найближчої породжуючої вершини; $P_{ij}$ – статистична оцінка ймовірності відвідування вершини $i$ через гіперпосилання з вершини $j$.

Під час оптимізації зв'язків графа за критеріями інформаційної компактності та індексу стратифікації вилучатимемо посилання (ребра графа), які мають найменшу вагу.

*Оптимальність у середньому* зумовлена статистичним підходом до визначення важливості (ваги (3)) гіперпосилань, а також поступовим уточненням оцінок імовірностей переходів при розрахунку індексу компактності (1).

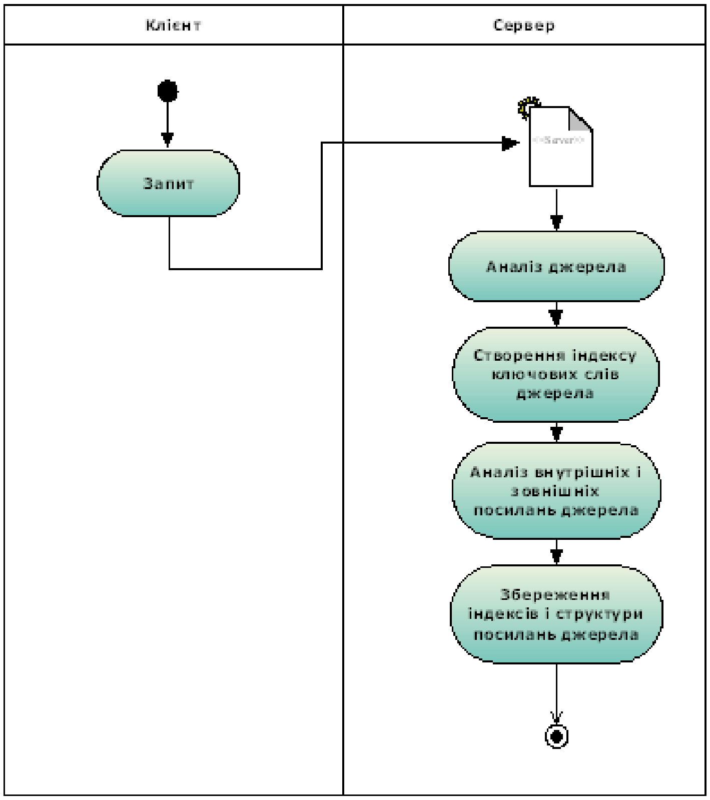


Рис. 1. UML діаграма обробки

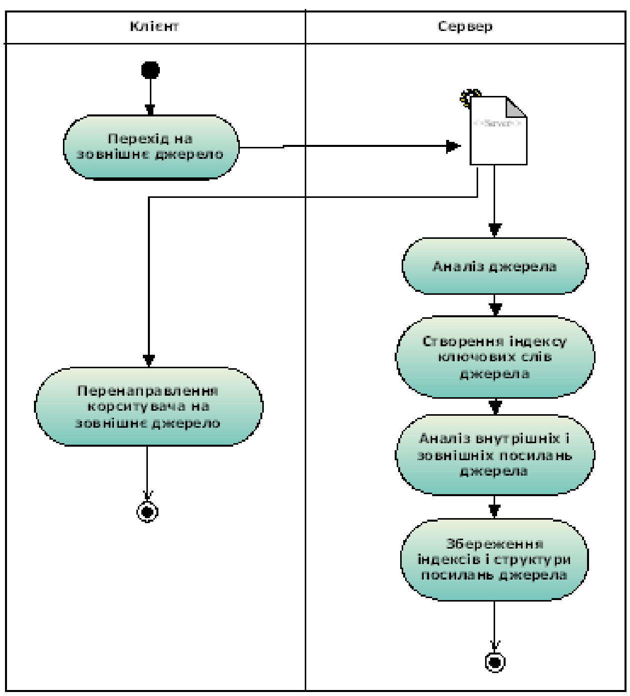


Рис. 2. UML діаграма обробки вихідних запитів

*Адаптація до інтелектуально-психологічних особливостей користувача* здійснюється шляхом встановлення індивідуального оптимуму показника інформаційної компактності та індексу стратифікації, які визначаються на основі статистичного аналізу сеансів роботи користувача в Інтернет.

Пропонується використати мультиагентну технологію [7, 9] та розробити відповідний агент, який міг би аналізувати й оптимізувати структуру сайту.

Програми-агенти розміщуються на web-серверах. При переходах користувача з сайта на сайт агенти обмінюються інформацією щодо структури сайтів. На основі цієї інформації здійснюється оптимізація структури сайтів і тимчасова деактивація зв’язків із найменшою вагою.

Обробка вхідного запиту (рис. 1) передбачає аналіз його параметрів, що містять службову інформацію про клієнта, зокрема, адресу джерела, з якого був здійснений перехід (Referrer). У разі наявності такої адреси, автоматично відбувається аналіз джерела, розміщеного за даною адресою, його індексація, аналіз структури, аналіз зовнішніх і внутрішніх посилань. Винятки в цьому випадку становлять пошукові системи, тематичні каталоги, системи оцінки рейтингу. Крім того можливе проведення аналізу ресурсу в глибину, до певного встановленого адміністратором рівня.

Результатом обробки і аналізу вхідного запиту, за умови наявності Referrer адреси ресурсу, є оброблена і збережена у базі даних (БД) інформація про даний ресурс, яка буде використана для оптимізації структури сайта і посилань на зовнішні джерела.

Вихідний запит, схема якого зображена на рисунку 2, складається із процедури аналізу кінцевого ресурсу, та перенаправлення на нього користувача. Аналіз кінцевого ресурсу передбачає його індексацію або реіндексацію у разі зміни даних, аналіз структури, зовнішніх і внутрішніх посилань.

Обробка зібраних даних (рис. 3) ініціюється адміністратором і містить: аналіз індексів, аналіз структур зовнішніх посилань проаналізованих ресурсів. На основі цього приймається рішення про модифікацію структури сайту.

Результатом модифікації структури ресурсу є оновлення структури сайту та передача її агентам суміжних сайтів.

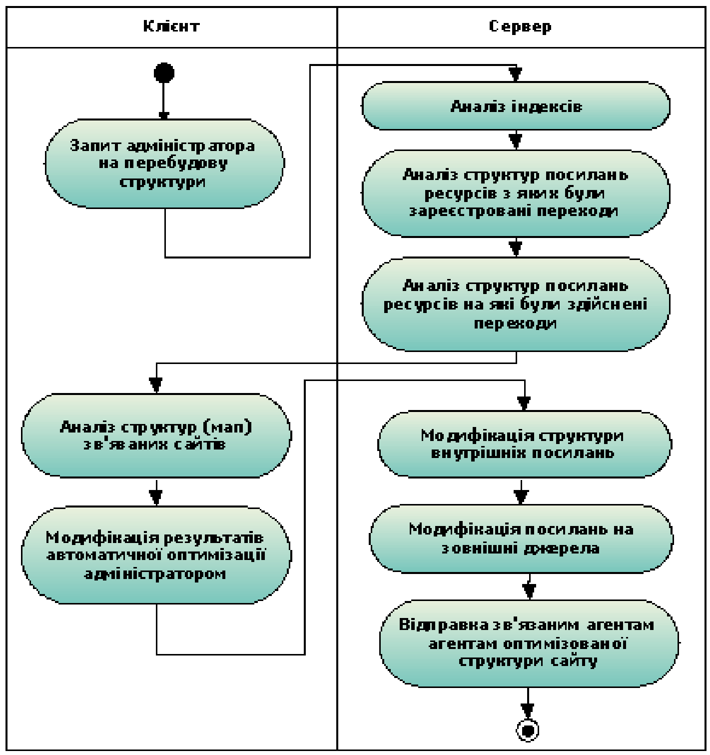


Рис. 3. UML діаграма обробки
зібраних даних і модифікації структури ресурсу

## Висновки

Запропоновано підхід до оптимізації структури інформаційного ресурсу, що в умовах неповної інформації про структуру мережі забезпечує збереження оптимальної досяжності фрагментів гіпертексту.

Сутність підходу полягає у представленні структури  інформаційного ресурсу у вигляді графа, ребрами якого є посилання, що зв'язують гіпертекст. Для структури ресурсу приймається гіпотеза про оптимальну складність, що чисельно виражається приведеним до числа вершин цикломатичним числом. Оптимізація структури в умовах неповної інформації здійснюється  шляхом пошуку бази графу і встановлення рівня важливості зв'язків, на основі чого проводиться модифікація їх структури і як наслідок відбувається керування рівнем інформаційної компактності.

Використання інтелектуальної мультиагентної технології дозволяє автоматизувати процес оптимізації, що сприятиме підвищенню ефективності використання Інтернет-ресурсів.

## СПИСОК ЛІТЕРАТУРИ

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 Володимир Михайлович*** – завідувач кафедри комп’ютерних систем управління;

***Глонь Ольга Віталіївна*** – доцент кафедри комп’ютерних систем управління;

***Москвін Олексій Михайлович*** – студент кафедри комп’ютерних систем управління.

Вінницький національний технічний університет