УДК 519.7:378.147

Н.Р. Пасічник, аспірант,
М.П. Дивак, д-р. техн. наук, проф.,
Тернопільський національний економічний університет
natalia.pasichnyk@gmail.com, mdy@tneu.edu.ua

Метод та алгоритм побудови структур тематичних веб-сайтів на основі онтологічного підходу

У статті запропоновано метод синтезу структури тематичних Веб-сайтів, який уможливлює зменшити витрати на їхню розробку. Створено алгоритм автоматичного синтезу структури сайтів у вигляді онтологічних дерев.

Ключові слова: тематичний Веб-сайт, онтологічне дерево, аналіз HTML коду.

Вступ

Одним із видів програмних додатків є тематичні Веб-сайти. Відомо, що основні витрати на розробку сайтів пов'язані із заробітною платою розробників, яка сьогодні є достатньо високою [6]. Вказаний програмний продукт також відзначається суттєвими витратами на його підтримку та розвиток. Аналогічно як і у процесі розробки, підтримка тематичного сайту потребує його постійного моніторингу не менш кваліфікованим персоналом як його розробники. Власники сайту прагнуть забезпечити максимальну віддачу сайту, тобто стимулювання бажаної поведінки цільової аудиторії. Цільова аудиторія зазнає впливу конкуруючих сайтів. За цих умов, служба підтримки сайту повинна не тільки динамічно оновлювати вміст сайту, а й розвивати його структуру відповідно до найнагальніших потреб цільової аудиторії. Щоб швидко та цілеспрямовано здійснювати цю діяльність та адекватно оцінювати активність конкурентів служба підтримки сайту повинна оцінювати потреби цільової аудиторії, фактори впливу власного сайту та його основних конкурентів. Враховуючи вищезазначене, актуальною є проблема зниження витрат на розробку та підтримку функціонування тематичних Веб-сайтів за умов забезпечення нагальних потреб цільової аудиторії.

Одним із можливих шляхів, зниження витрат на розробку та підтримку тематичних Веб-сайтів є створення методів, алгоритмів та засобів для автоматизованого генерування чи реінженірінгу його структури. Саме цю актуальну задачу розглянуто в даній праці.

Постановка задачі

Для забезпечення привабливості Веб-сайту, його сторінки повинні містити актуальну інформацію з точки зору цільової аудиторії. Представники цієї аудиторії мають різну поінформованість про функціонування об'єкту, представленого Веб-сайтом та різні актуальні інформаційні потреби. Тому інформація сайту повинна бути достатньо різноманітною. Для спрощення доступу до неї сайт повинен мати певну структуру. Вона є базою для побудови інформаційного наповнення. Елементи інформаційного наповнення відрізняються як тематикою, так і ступенем та способом впливу на цільову аудиторію. Їх аналіз уможливив виділити такі основні функції сайту: представницька, надання інформаційних послуг, надання он-лайн послуг. Представницька функція забезпечується звичайним вмістом сторінок сайту і представляє особливості функціонування об'єкту. Інформаційні послуги надають за рахунок тематичних колекцій інформаційних сторінок або інформаційних ресурсів. Он-лайн послуги надають за допомогою спеціальних он-лайн сервісів. Якщо представницька функція як правило не може збільшити аудиторію сайту, то дві останніх можуть суттєво її розширити. Кожен із засобів реалізації функцій сайту має певну вартість створення та ефективність дії. Вони можуть доповнювати або виключати одне одного і по різному діяти на різних представників цільової аудиторії.

Не залежно від функціонального призначення тематичного Веб-сайту першим етапом його розробки є створення узагальненої моделі структури цільового сегменту як основи для формування його інформаційного наповнення, а також побудови процедури ідентифікації інтенсивності представницької функції сайту на певній цільовій аудиторії. Фактично вказана модель є онтологічним деревом, яке відповідає вимогам цільової аудиторії. Об'єктивно для побудови такої моделі можливо застосувати два шляхи: 1) аналіз вмісту численних Веб-сторінок - результатів запитів до пошукових серверів, тобто за аналогією до найпоширеніших Веб-сайтів, які є привабливими для даної цільової аудиторії; 2) у спосіб опитування експертів. Обидва шляхи вимагають розробки у певній мірі формального методу синтезу онтологічного дерева. В основу методу синтезу структури Веб-сайту покладемо багатокроковий пошук та упорядкування концептів

онтологічного дерева. Перейдемо до розгляду даного методу.

Метод синтезу структури тематичного Веб-сайту

Представники аудиторії цільового Веб-сегменту аналізують численні сайти, відзначаючи подібність та відмінність їхньої структури. Користувачі шукають потрібну для них інформацію, але прагнуть щоб шлях цього пошуку був максимально простим та звичним. Тому доцільно основу структури сайту будувати із найпоширеніших структурних елементів сайтів цільового сегмента. Формалізуємо метод побудови такої основи. При інтуїтивній специфікації структури розроблюваного сегмента дослідник переглядає структури кількох конкуруючих сайтів. Недолік такого аналізу полягає у великих витратах та можливості упущення за масивами повторюваної інформації унікальних структурних особливостей. Використання інформаційних технологій дозволяє забезпечити повноту аналізу великої множини структур сайтів з уникненням перегляду великого обсягу однотипних сторінок.

Структура поєднання сторінок Веб-сайту може бути багаторівневою, хоча з позицій зручності навігації та повноти індексації рівнів пошуковими серверами, використання рівнів вище третього не рекомендується. У будь-якому випадку задачу виділення типової структури Веб-сайту можна декомпонувати на подібні між собою задачі встановлення елементів типового головного меню та задачі аналізу типового наповнення цих елементів.

Формалізуємо метод побудови типового головного меню тематичного Веб-сайту. Нехай тематику таких сайтів задано загальним текстовим маркером SAM предметної області та маркером SS її специфікатора. За допомогою пошукового сервера та операції конкатенації символьних стрічок, за тематичним запитом $QS = SAM \ \& \ SS$ отримаємо множину SP Веб-сторінок, представлених своїми *HTML* кодами

$$SP(QS, P) = \{HP_i\}_{i=1}^{P}, \quad (1)$$

- де P - потужність множини SP; HP_i - елемент множини, що визначає $HTML$ - код i - тої сторінки.

Для подальшого аналізу доцільно використовувати лише сторінки підмножини SPR, в яких предметна область є об'єктом аналізу, а не елементом ширшого контексту. Критерієм виконання цієї вимоги є наявність маркера предметної області в заголовку сторінки

$$SPR = \left\{HP_{i^*}^* \mid HP_{i^*} \in SP,\ HP_{i^*}.title \supset SAM\right\}_{i=1}^{P^*}. \quad (2)$$

Із коду $HTML$ сторінки необхідно вибрати інформацію, які характеризують її структуру. У першу чергу інформація про структуру представлено в меню даної сторінки. Зміст пунктів меню встановлюємо на основі анкорів тегів у вигляді "$< a \ldots href \ldots >$", які групуються в ієрархічні списки. Вибрані елементи верхнього рівня утворюють впорядкований список $LRA(HP_i^*)$ анкорів сторінки

$$LRA(HP_i^*) =< A_{ij}\left(HP_i^*\right) >_{j=1}^{APi} \quad (3)$$

Якщо список анкорів відсутній, то інформацію про структуру вибираємо з виділених елементів сторінки, якщо вони містять відносно небагато елементів, а їх кількість не надмірно велика. Інформація такого роду є впорядкованим списком $LRB(HP_i^*)$ виділених елементів сторінки :

$$LRB(HP_i^*) =< B_{ik}\left(HP_i^*\right) >_{k=1}^{BPi} \quad (4)$$

Звичайно такі списки можуть містити і випадкову інформацію. Однак елементи таких списків, які повторюються, вже дають інформацію про засоби структурування інформації цільового сегменту. Тому на основі списків анкорів та виділених елементів сформуємо звичайну SCF та узагальнену GCF множини частот концептів, які визначаємо у такий спосіб:

$$SCF = \left\{(C_l, NC_l) \mid C_l \in \bigcup_i LRA(HP_i^*)\right\} \quad (5)$$

$$GCF = \left\{(C_m, NC_m) \mid C_m \in \left(\bigcup_i LRA(HP_i^*)\right) \bigcup \left(\bigcup_i LRB(HP_i^*)\right)\right\} \quad (6)$$

Для виявлення концептів, релевантних до предметної області впорядкуємо елементи множини *SCF* у порядку спадання частот елементів. При цьому частину понять одразу включаємо в концептуальну множину SCN

$$SCN = \left\{(C_l, NC_l) \mid FCL_l = \frac{NC_l}{P^*} > F_0\right\}, \quad (7)$$

де величина F_0 як правило належить інтервалові [0.2; 0.5], а конкретне значення вибирають виходячи зі специфіки дослідження. Якщо таким чином поповнити множину концептів не вдалося, то здійснюємо відбір серед концептів із низькою частотою. Розглянемо впорядковану вибірку частот *SSF*

$$SSF = \left\{NC_l \mid (C_l, NC_l) \in SCF\right\} \quad (8)$$

Найвищі частоти цієї вибірки перевіряємо на аномальність за критерієм 4σ [2]. Концепти, що відповідають виявленим аномальним значенням включаємо в множину *SCN*. Якщо і

далі ця множина залишається пустою, то збільшуємо потужність множини сторінок для аналізу, виданих пошуковим сервером. Коли при цьому множина *SCN* не поповниться, то робимо висновок, що базові концепти даної предметної області не розпізнано.

Якщо вдається поповнити концептуальну множину SCN, то її концепти C_l включаємо в такий запит до пошукового сервера:

$$QS = SAM \ \&\ \{C_l \mid (C_l, NC_l) \in SCN\} \ \&\ SS \quad (9)$$

Після цього повторюємо спроби поповнення множини концептів на основі нового уточненого запиту та вище описаної процедури.

При остаточному уточненні набору *K* концептів, що специфікують дану предметну область, необхідно їх впорядкувати. Для цього, для кожної сторінки, список анкорів якої містить концепти предметної області, проводимо їхнє ранжування. При цьому рангові номери виставляємо лише концептам. Так R_{ik} позначає номер (ранг) *k*-го концепту на *i*-тій сторінці. Необхідно також враховувати нерівносильність впливу $HTML$ сторінок на аудиторію предметної області, оскільки багато користувачів переглядають перші 10 – 20 елементів видачі пошукових серверів, і лише в дуже нечисленних випадках - сторінки із другої сотні списку. Для врахування цього факту потрібно використати деяку монотонно спадну вагову функцію $w(i)$.

Оскільки для перших 10-20 сторінок їх важливість плавно спадає, а далі це спадання значно прискорюється, в якості такої функції зручно використовувати кубічний сплайн. Для її однозначного визначення необхідно накласти хоча б 4 умови. Зокрема покладаємо важливість першої сторінки рівною 1 при нульовій похідній для цього аргументу, деякі характерні ваги наприклад для 20-ї та 100-ї сторінок, які вибираються експертним шляхом. На основі описаного підходу вводимо нормовану систему ваг $G(i) = \dfrac{g(i)}{\sum_i g(i)}$. Після цього усереднений ранг *k*-го концепту по предметній області обчислюємо за допомогою наступного співвідношення:

$$RA_k = \sum_i R_{ik} G(i) \quad . \quad (10)$$

На основі усереднених рангів і відбувається ранжування концептів. Узгодженість рангів оцінюється за коефіцієнтом конкордації Кендела [3,4]:

$$W = \frac{12 \sum_{k=1}^{K} \left(\sum_{i=1}^{I} R_{ik} G(i) - \overline{R} \right)}{I^2 (K^3 - K)} \quad (11)$$

де $\overline{R} = \dfrac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{I} R_{ik} G(i)$.

За умови, що

$$I(K-1)W > \chi^2_{K-1,\alpha} \quad , \quad (12)$$

ранжування вважається значущим [5].

Якщо ранжування по повному списку концептів не можна вважати значущим, то усуваємо по одному із списку в порядку зростання їхніх ваг, тобто сумарних ваг сторінок, де вони зустрічаються. Усунення починаємо від концепта із найменшою вагою і продовжуємо аж до отримання значущого ранжування. У такому разі відібрану множину концептів розбиваємо на дві частини: множину *RG* із погодженими рангами та *NR*, рангування по якій встановити не вдалося.

Алгоритм синтезу структури тематичного Веб-сайту

На основі наведених теоретичних положень сформуємо алгоритм синтезу структури сайту:

1. Встановлюємо лічильник ітерацій $CI = 1$ та формуємо підмножини SPR множини SP $HTML$ кодів веб-сторінок на основі запиту $QS = SAM$ & SS для яких маркер SAM предметної області міститься в назві коду

2. Формуємо множину SCF частот концептів на основі анкорів тегів, а якщо вона пуста то узагальнену GCF множину частот концептів на основі виділених елементів Веб-сторінок.

3. Формуємо концептуальну множину SCN предметної області за критерієм перевищення мінімальної частоти F_0 та оцінюємо її потужність PSC.

4. Якщо концептуальна множина порожня, поповнюємо її елементами із аномальними частотами.

5. Якщо $PSC = 0$ та $CI = 1$, то завершення алгоритму. Якщо потужність концептуальної множини зменшилась в порівнянні з попереднім кроком, то перехід на пункт 6, інакше збільшуємо лічильник ітерацій $CI = 1$ та формуємо підмножини SPR множини SP $HTML$ кодів Веб-сторінок на основі початкового запиту QS,

поповненого елементами концептуальної множини даного кроку і переходимо на пункт 2.

Ранжуємо елементи найпотужнішої із побудованих множин SCN та аналізуємо значущість коефіцієнту конкордації цього ранжування. При не значимості коефіцієнта конкордації послідовно вилучаємо із множини концепти із найменшими вагами аж до отримання значущості згаданого коефіцієнта.

Чисельні експерименти

На основі запропонованого методу досліджено процес побудови типової структури сайту факультету українського вузу. В цьому випадку покладаємо *SAM*="факультет", SS="ua". Експериментальним шляхом встановлено прийнятне значення F_0 на рівні 0,3, оскільки допустимий рівень 0,4 призводив до надто звуженого набору концептів при розгляді лише 100 перших сторінок запиту. Результати першого етапу досліджень наведені в таблиці 1.

На четвертому кроці кількість відібраних концептів зменшилася, тому процес їх набору припинено, а за базову множину взято сукупність відібраних концептів на третьому кроці. У подальшому реалізовувалася процедура ранжування відібраних концептів. У таблиці 1 концепти наводилися в порядку спадання їхніх частот. У таблиці 2 наведені адреси сторінок, за допомогою аналізу яких на 3 кроці відбиралися концепти, а на рисунку 1 – фрагмент видачі пошукового сервера, що ілюструє принцип відбору згаданих сторінок. Рисунок 1 наочно ілюструє принцип відбору релевантних сторінок для аналізу за критерієм наявності в заголовку сторінки ключового терміну *"факультет"*. У таблиці 3 наведено результати ранжування відібраних концептів, а таблиця 4 ілюструє процес побудови згаданих рангів.

Концепти множини SCN впорядковуються по частоті згадування на відібраних сторінках. Текстові описи концептів наведені в таблиці 3. В стрічках таблиці 4 наводяться ранги концептів по порядку їх появи на відповідній сторінці. Будуються середні значення рангів по всіх аналізованих сторінках як основа для їх впорядкування.

Таблиця 1. **Результати процесу відбору базових концептів**

Номер кроку	**Кількість відібраних сторінок**	**Кількість відібраних концептів**	**Перелік концептів**
1	15	3	головна, абітурієнт, новини
2	17	8	факультет, головна, абітурієнт, новини, контакти, кафедри, студенти, фотогалерея
3	13	11	головна, абітурієнт, кафедри, факультет, студент, наука, бібліотека, контакти, форум, фотогалерея, новини
4	8	8	головна, абітурієнт, новини, форум, наука, бібліотека, контакти, фотогалерея

Таблиця 2. **Характеристики Веб-сторінок остаточно відібраних концептів**

Номер сторінки	**Позиція сторінки**	**Адреса сторінки**
1	3	http://istfak.org.ua/
2	12	http://www.history.univ.kiev.ua/
3	14	http:// istorikznu. at. ua/index/nash_fakultet/0-6
4	18	http://fknit.com.ua/
5	28	http://www-psyhology.univer.kharkov.ua/
6	31	http://www.university.kherson.ua/About/Faculty/Faculty_of_Law.aspx
7	43	http://istfak.lg.ua/
8	56	http://fizmat.chnpu.edu.ua/
9	57	http://fizmatsspu.sumy.ua/
10	58	http://www.znu.edu.ua/ukr/university/departments/fizvosp
11	61	http:// istorikznu. at. ua/index/nash_fakultet/0-6
12	70	http://www.fitis.ck.ua/
13	75	http://forlan.org.ua/index.php?option=com_content&task=section&id=15&Itemid=115

Таблиця 3. **Результати встановлення рангів відібраних концептів**

Матспо-дівання рангів	1	3	3.22	3.63	3.8	4.55	4.71	4.8	5	6	6.4
Порядкові ранги	1	2	3	4	5	6	7	8	9	10	11
Ранги частот	1	3	4	6	11	2	5	9	8	10	7
Кон-цепти	голов-на	кафед-ри	факуль-тет	нау-ка	нови-ни	абітурі-єнту	сту-денту	фо-рум	кон-такти	фото-галерея	бібліо-тека

Рисунок 1 - Фрагмент другої сторінки списку посилань пошукового сервера для кроку 3

Таблиця 4. **Середні значення рангів відібраних концептів**

Номер сторінки	**Вага сторінки**	**Номери концептів, впорядкованих по частоті**										
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1	1	1					2	4		5		3
2	0.8	1	6	3	2	5	4		7			
3	0.78	1	5	2	4	6		8			7	
4	0.73	1	8	4	3				7	5	6	2
5	0.5	1	5	2		4	3	6				
6	0.49		1	6	5	2			4	3		7
7	0.48	1	3		2		5		4			
8	0.3	1	6	3		5	4			8	7	2
9	0.29	1		2			3					
10	0.28	1	3		2		4					
11	0.27	1	4	2	3	5		7			6	
12	0.2	1	7		5	6			2	3	4	
13	0.19	1	2		3		4	7	6			5
Середнє значення рангів		1	4.55	3	3.22	4.71	3.63	6.4	5	4.8	6	3.8
Вага концепту		5.82	5.02	4.16	4.22	3.34	3.84	2.74	2.89	2.72	2.28	2.71

Для зважування значущості сторінок використані наступні параметри $g(20) = 0.8,\ g(100) = 0.1$. Вага концепту визначається як сума ваг сторінок на яких він зустрічається.

Побудова коефіцієнту конкордації по повному набору параметрів показала незначущість такого ранжування $W = 0.05,\ \chi^2_{emp} = 6.31,$ $\chi^2_{10,0.05} = 18.31$. При послідовному редукуванні списку концептів значущі значення коефіцієнта конкордації отримано для сукупності із 9 елементів. При цьому отримано наступні значення $W = 0.18,\ \chi^2_{emp} = 18.60,$ $\chi^2_{8,0.05} = 15.51.$ Оскільки емпіричне значення критерію χ^2 перевищує критичне значення, то гіпотезу про значимість отриманого впорядкування концептів слід прийняти із рівнем значимості в 5%. При цьому впорядкований список концептів склав *RG = < головна, кафедри, факультет, наука, абітурієнту, студенту, форум, контакти, бібліотека >*, а невпорядкована множина отримала наступне представлення *NR ={ новини, фотогалерея }.*

Висновки

У статті розглянуто один із можливих шляхів зниження витрат на розробку та підтримку тематичних Веб-сайтів шляхом синтезу онтологічного дерева як загальної моделі структури цільового сегменту. В основу методу синтезу структури Веб-сайту покладено багатокроковий пошук та упорядкування концептів онтологічного дерева.

У результаті проведених досліджень отримано такі наукові та практичні результати. Вперше запропоновано метод частотного аналізу структурних елементів тематичних Веб-сторінок, у якому використовуємо формальний аналіз елементів структур Веб-сайтів, описаних у HTML кодах цільових сторінок. Це уможливлює формалізацію процедури побудови головного меню типового Веб-сайту певної предметної області. Ефективність запропонованого методу та алгоритму підтверджено розв'язуванням практичної задачі синтезом типової структури меню сайту факультету.

Список використаної літератури

1. Пасічник Н.Р. Формалізм в постановці задачі створення якісного сайту / Н.Р. Пасічник, М.П. Дивак // Наукові праці ДонНТУ. Серія „Інформатика, кібернетика та обчислювальна техніка. – 2011. – Вип 14 (188). – С. 325-329.

2. Закс Л. Статистическое оценивание / Л. Закс. – М.: Статистика, 1976. – 598 с.

3. Дивак М.П. Методичний посібник з дисципліни "Системний аналіз" [Електрон. ресурс] / М.П. Дивак. – Режим доступу: http://library.tneu.edu.ua/files/EVD/IV_06/POSIBN_EK.pdf

4. Петров Е.В. Видимый результат, или Система сбалансированных показателей для службы персонала [Електрон. ресурс] / Е.В. Петров, А.А. Югов, О.В. Гурина. – Режим доступу: http: //www. iteam. ru/ publications /human/ section_45/ article_2708/

5. Ромашкина Г.Ф. Коэффициент конкордации в анализе социологических данных / Г.Ф. Ромашкина, Г.Г. Татарова // Социология. – 2005. – № 20. – С. 131-158.

6. Учет расходов организации на создание интернет-сайта [Електрон. ресурс]. – Режим доступу: http:// ulet-studio. ru/ stat. php? id=2 &Uchet_ rashodov_ organizatsii _ na _ sozdanie _ internet-sayta

Надійшла до редколегії 30.01.2011

Н.Р. ПАСИЧНЫК, М.П. ДЫВАК
Тернопольский национальный экономический университет

N.R. PASICHNYK, M.P. DYVAK
Ternopil National Economic University

МЕТОД И АЛГОРИТМ ПОСТРОЕНИЯ СТРУКТУР ТЕМАТИЧЕСКИХ ВЭБ-САЙТОВ НА ОСНОВЕ ОНТОЛОГИЧЕСКОГО ПОДХОДА

В статье предложен метод синтеза структуры тематических Веб-сайтов, который позволяет сократить затраты на их разработку. Создан алгоритм автоматического синтеза структуры сайтов в виде онтологических деревьев.

Ключевые слова: тематический Веб-сайт, онтологическое дерево, частотный анализ, HTML код.

MATRIX THE METHOD AND ALGORITHM OF CONSTRUCTION OF THE CONTENT WEBSITES STRUCTURES BASED ON THE ONTOLOGICAL APPROACH

The article offers method of the thematic Website's structure synthesis. The algorithm of automated synthesis of thematic Website's structure is developed.

Keywords: thematic Website, ontology tree, frequency analysis, HTML code.