УДК 004.9

Замятін Д. С., Михайлюк А. Ю., Михайлюк В. А., Петрашенко А. В.

СФОКУСОВАНИЙ КРАУЛІНГ ЯК ЗАСІБ ЗНИЖЕННЯ РЕСУРСОЄМНОСТІ ПОШУКУ У WEB

У статті досліджується проблема створення системи моніторингу тематичних Web-ресурсів для корпоративного середовища. Запропоновано класифікацію основних алгоритмів обходу ресурсів. Розроблено класифікацію метрик ранжування сайтів за ознакою об'єктів, на основі яких виконується оцінювання. Проведено попередній розрахунок оцінки придатності сфокусованого пошуку для даної задачі.

Ключові слова: пошукові системи, Web краулери, тематичний пошук.

На даний час Інтернет набув статусу найбільш вживаного середовища обміну інформацією і знайшов застосування у всіх сферах життя суспільства. При цьому, через різнорідність та збільшення обсягу інформації у Web-просторі з часом ускладнюється виділення потрібних користувачам матеріалів. Крім того, велике значення для багатьох категорій користувачів, таких як викладачі, які мають перебувати на вістрі останніх досягнень у науці, або керівники навчальних підрозділів, які прагнуть бути в курсі сучасних тенденцій у теорії та практиці педагогіки, має оперативність доступу до новин та іншого динамічного контенту. Для вирішення даної проблеми існують різноманітні програмні засоби, зокрема пошукові системи, які спираються на зв'язаність сторінок гіперпосиланнями та моніторингові системи, які оперативно відслідковують цільові сайти. При цьому найдоступнішим на сьогодні засобом виявлення у Web потрібної їм інформації є популярні пошукові системи, такі як Google, Yandex, Rambler та ін.

Розмір проіндексованого пошуковиками всесвітнього Web-простору становить принаймні 13,23 мільярдів сторінок [1], а повний обсяг Web ще більше. При цьому гостро встає проблема масштабованості рішень [2, 3, 4, 5]: постійно збільшується розходження між проіндексованою частиною Web та його «повним» об'ємом, ускладнюється підтримка актуальності індексу. Отже, звичний екстенсивний підхід до нарощення можливостей пошуку з максимальним охопленням сторінок стає все менш виправданим через вартість дискових, мережених та обчислювальних потужностей, близьких до необхідних [5, 6]. Разом з тим, більшість популярних пошукових систем розраховані на пересічного користувача і, як наслідок, не можуть повністю задовольнити потреб конкретних користувачів, оскільки контекст запиту не враховується пошуковими механізмами, і кожному користувачеві надається однаковий усереднений результат [5, 7]. Одним з небагатьох винятків є Google Personalized Search, який разом з тим, базується на збереженні історії попереднього пошуку користувача, що не завжди прийнятне з точки зору інформаційної безпеки. При цьому використання у пошукових системах загального вжитку метрик цінності сторінок, заснованих на популярності, яка виражена кількістю посилань на сторінку в умовах неконтрольованого хаотичного розвитку Web-простору не завжди забезпечує якість отриманої інформації [8, 9, 10]. Крім того, залежність від популярного пошуковика накладає обмеження на автоматизоване здійснення запитів, що перешкоджає застосуванню сучасного інструментарію аналізу даних.

Отже, на сьогодні є актуальним питання створення моніторингової інформаційно-аналітичної системи (МІАС) для відслідковування інформації з тематики, яка задана користувачем, в межах окремого сегменту Web, придатної для використання у корпоративному середовищі. Вхідні дані МІАС подаються у вигляді початкового списку сайтів потрібної тематики (наприклад, сайтів з новинами та блогів) та пошуковими запитами; вихідні дані — список релевантних фрагментів тексту з посиланнями на джерело. Апаратні ресурси, які можуть бути надані для системи, зокрема пропускна здатність каналу зв'язку, обмежені тим, що немає можливості відмовитися від інших задач і виділити МІАС всі корпоративні ресурси. Велике значення має також вартість створення або ліцензування сторонніх програмних засобів, необхідних для реалізації подібної системи, її складність, простота інтеграції з іншими загальновживаними інформаційними системами у корпоративному середовищі, можливість її розробки у прийнятні терміни.

Одним з найбільш опрацьованих рішень для здійснення тематичного пошуку у Web на сьогодні є пошук із застосуванням сфокусованого краулера [11, 12, 13]. Сфокусований краулер – краулер, який

прогнозує ймовірність того, що посилання вказує на релевантну сторінку ще перед її завантаженням, що дозволяє у більшості випадків не завантажувати нерелевантні сторінки [15]. Зазвичай дані краулери мають менше пошукового шуму (з точки зору реального користувача) у результатах пошуку, ніж звичайні рішення [5], що зокрема дає змогу застосовувати сучасні інтелектуальні засоби аналізу даних, які є неефективними на великих об'ємах інформації. Таким чином, при сфокусованому краулінгу підвищення новизни, швидкості отримання та глибини охоплення релевантної підмножини Web відбувається за рахунок зменшення покриття Web [2, 5]. Це можливо завдяки тематичній близькості – малої відстані за посиланнями при розташуванні тематичної інформації у Web, що дає змогу обходу потрібних сторінок загалом залишаючись у межах релевантної підмножини Web. При цьому проблема масштабованості пошуку у сучасному Web вирішується шляхом розподілу та децентралізації задачі пошуку між користувачами [6, 2].

Типовий сфокусований краулер містить (див. рис. 1) модуль черги «сторінок до завантаження», яка впорядкована за обраною розробником метрикою цінності, модуля списку відвіданих сторінок, модулів відбору URL, модуль ранжування, модуль оцінки вмісту та модуль визначення частот повторного звернення. При цьому порядок використання модулів зводиться до наступного.

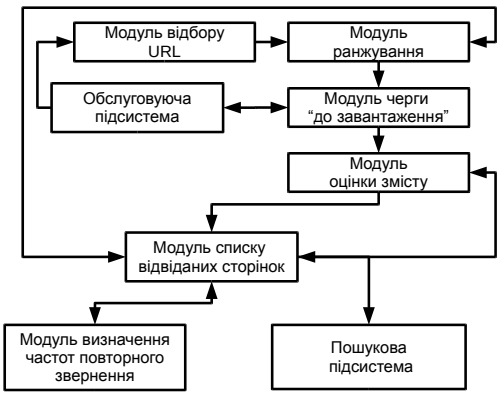

Рис 1. Структурна схема сфокусованого краулера

Перед першим запуском краулера черга має містити стартові адреси сторінок. Під час роботи краулера URL з черги послідовно вибираються та сторінки, на які вони вказують, завантажуються на сервер, на якому запущено краулер. З цих сторінок відбираються адреси для продовження обходу та вносяться до черги з дотриманням її пріоритетності. При цьому частину часу краулер витрачає на повторний огляд вже відвіданих сайтів з частотою, яка враховує власну частоту оновлення джерела. Обслуговуюча частина, яка у даній статті не розглядається, складається з тих низькорівневих модулів, які технічно необхідні для реалізації роботи краулера: бази даних, яка зберігає дані черги та списку відвіданих сторінок, різноманітну статистику та налаштування, модуль завантаження сторінок, який взаємодіє з Web-серверами по протоколу HTTP тощо.

Існує багато модифікацій алгоритмів пошуку та метрик цінності сторінок. Розглянемо основні алгоритми та підходи до обходу Web для того, щоб мати можливість вибору найбільш доцільних для вирішення задачі побудови МІАС та запропонуємо класифікацію даних алгоритмів (див. рис. 2).

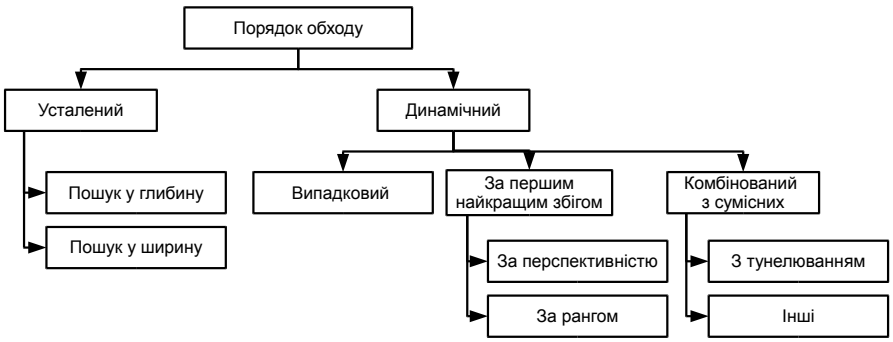

Рис 2. Класифікація алгоритмів обходу Web

Web можна розглядати як граф, у якому вузли представлені сторінками, а дуги – посиланнями [16], тому для пошуку у ньому застосовуються стандартні алгоритми пошуку у глибину, пошуку у ширину та пошуку за першим найкращим збігом. Пошук за першим найкращим збігом у контексті Web-простору полягає у виборі для переходу посилань-ребер на вершини-сторінки графа Web у порядку їх оцінки, яка може складатися, наприклад, з рангу за обраною метрикою та/або ймовірності того, що у близькій перспективі вони приведуть до високорангових сторінок. Одним з способів визначення перспективності є застосування класифікатора, тренованого під час роботи краулера [17] або попередньо на т. з. контекстному графі сторінок [18], які містять посилання на релевантні сторінки. Недоліком використання контекстного графу є необхідність залучення іншої пошукової системи з вичерпним скануванням сегментів Web, у яких буде виконуватися пошук [16]. Згідно експериментів [17], тренування павука під час роботи за технологією навчання з підкріпленням дозволяє отримати 75% шуканих документів при проходженні в 2-3 рази меншої кількості посилань у порівнянні з звичайним пошуком в ширину. Недоліком є складність забезпечення гнучкості класифікатора у виборі тем, а також не врахування того, що деякі сторонні документи найчастіше ведуть до тематичних [18]. Подібну задачу вирішує технологія тунелювання, яка передбачає встановлення порогу терпіння, після якого краулер припиняє прохід через низькорангові сторінки [7]. Це дає змогу вийти на потрібний кластер (підмножина результатів пошуку, об’єднаних одною темою) сторінок, навіть якщо він не має прямого зв’язку з тим кластером, у якому знаходиться краулер. Крім того, застосовуються модифікації даних алгоритмів та комбіновані рішення, зокрема, перехід на окремих ділянках, таких як вибір частини посилань на сторінці часто виконують випадковим чином. Це дає можливість обмежити кількість сторінок до обробки і на великих масивах гіпертексту не перешкоджає прийнятному покриттю. До недоліків пошуку за першим найкращим збігом відноситься залежність від якості використовуваної евристики, а також від дотримання вимог, які евристика ставить до користувача системи, наприклад, вдалого вибору стартових сторінок [18].

Незважаючи на простоту алгоритму, пошук у ширину після меншої, ніж у інших алгоритмів, кількості ітерацій виявляє сторінки, найбільш популярні у сенсі посилань на них [2, 10]. Це зручно для пошукових систем загального вжитку, оскільки це відповідає метриці оцінки якості, яка у них здебільшого використовується. В той же час алгоритм є ресурсоємким, оскільки по мірі просування вниз по графу кількість вузлів швидко зростає.

При використанні пошуку у глибину для Web-простору подальший обхід з сторінки-вершини по ребрах-посиланнях не виконується не лише при відсутності посилань, але й при виконанні певного евристичного правила. Такий пошук дозволяє швидко знайти деяку кількість прийнятних сторінок, які підходять під умову, але для охоплення, яке може бути порівняне з пошуком в ширину потребує значно більшого часу [19].

Розглянемо основні метрики релевантності для того, щоб мати можливість вибору такої, яка підходить найбільше для поставленої задачі. На рис. 3 подано запропоновану класифікацію метрик релевантності.

Жоден краулер, навіть ті, що використовуються глобальними пошуковими системами, не охоплює весь Web [1]. Окрім того, порядок обходу визначає час, через який сторінка опиниться у пошуковому індексі і з’явиться у результатах пошуку. Тому виникає потреба у ранжуванні сторінок за одною чи

декількома метриками. При практичній реалізації краулера вони задають порядок впорядкування черги “до завантаження” та враховуються при визначенні доцільності переходу за посиланням. Це особливо актуально для сфокусованих краулерів, які базуються на алгоритмі пошуку за першим найкращим збігом. У спеціалізованих краулерах, для рекурсивного завантаження окремих сайтів ранжування може взагалі не виконуватися. У більших масштабах це є обґрунтованим лише за наявності відповідних ресурсів.

Одним з підходів є використання метрик, які оцінюють сторінки за популярністю на основі так званих зворотніх посилань, яка виражена числом посилань, що вказують на дані сторінки. Більш розвинутими метриками даної групи є метрики PageRank та HITS. Їх перевагою є те, що вони добре відповідають запитам пересічного користувача, недоліком — те, що популярність далеко не завжди пояснюється релевантністю, особливо у випадку інформації на теми, що становлять інтерес лише для фахівців вузького профілю [8, 9, 10].

В основі PageRank лежить ідея не лише рахувати кількість зворотніх посилань, але й враховувати те, що посилання з сторінки високого рангу важить більше, ніж посилання з низькорангової сторінки. Це реалізується шляхом акумулювання кількості зворотніх посилань, тобто враховуються всі посилання, через які можна пройти на сторінку. PageRank добре себе зарекомендував і покладений в основу алгоритмів, які використовуються більшістю пошукових систем. Але по мірі сканування графу Web через необхідність зберігання інформації про посилання збільшується вимоги до ресурсів [9].

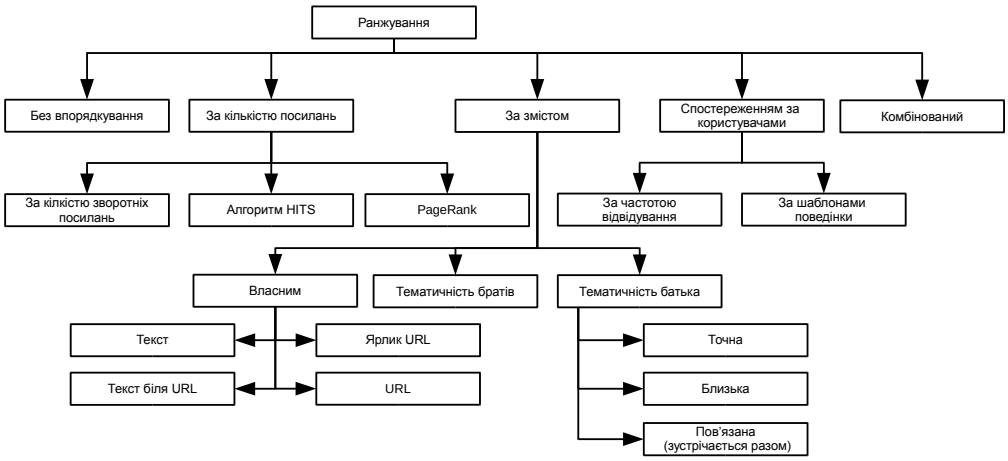

Рис. 3. Класифікація метрик релевантності

Алгоритм HITS ділить сторінки на два взаємозалежні класи: “хаби”, які розглядаються як переліки посилань та ”авторитети” з певної інформації. При цьому на першій ітерації алгоритму ранги авторитетів та хабів відповідно дорівнюють $V_{авт}=1$ та $V_{хаб}=1$, а далі для кожного авторитета та хаба обчислюються:

$$V_{авт}=\sum_{i=1}^{n}V_{хаб\,i}\ ,\ V_{хаб}=\sum_{j=1}^{m}V_{авт\,j}\ ,\ \text{де}$$

n – кількість сторінок, що вказують на даний авторитет,

m – кількість сторінок, на які вказує даний авторитет,

i, j – номери сторінок.

Далі обчислення рангів виконується циклічним перерахунком. Даний алгоритм погано працює при непотизмі -- автори сторінок домовляються про взаємодопомогу шляхом встановлення у себе посилань один на одного, що є типовим явищем для сучасного Web [20].

Перспективною метрикою є оцінка за змістом. Зазвичай виконують оцінку тексту сторінки, тексту, який знаходиться поряд з URL, назву посилання, текст URL як такий (наприклад, cycling.com з більшою ймовірністю стосується велосипедного спорту ніж інші сайти) або в їх комбінації. При цьому становлять

інтерес як теми сторінок, які обрано користувачем, так і ті, що є близькими або часто зустрічаються поряд з ними. Крім того, оцінка тексту може виконуватися як для самої сторінки, так і для сторінок “батьків”, що мають посилання на неї, а також сторінок-“братів”, на які є посилання з тих самих “батьків”. Це дозволяє скористатися можливостями, які надає наявність тематичної близькості у Web. [21, 6] Перевагою змістовного підходу є можливість оцінити інформацію за її власною цінністю, недоліком — складність точної оцінки, оскільки застосовуються квазісемантичні методи, які лише імітують з певним ступенем наближеності “розуміння” змісту тексту.

Іншу групу складають метрики, які передбачають спостереження за користувачами. При цьому виявляються користувачі з заданими інформаційними інтересами, після цього сторінка або сайт оцінюються за частотою відвідування даними користувачами та тим, чи вона переглядалася ними поряд з вже визначеними релевантними сайтами. Наприклад, якщо сайт переглядався під час Web-серфінгу сайтів, присвячених історії, ймовірно, що він має історичну тематику. Недоліком даної групи метрик є втручання у особисте життя користувачів, а також необхідність мати доступ до журналів сервера, який використовується значною кількістю користувачів, що є необхідним для забезпечення репрезентативності [22].

Виконаємо приблизний розрахунок для підтвердження доцільності використання сфокусованого краулера. При цьому припустимо, що сфокусований павук може обробляти тематичні підмножини Web з завантаженням незначної кількості не релевантних сторінок, якою можна знехтувати.

Перш за все, варто описати середовище, у якому повинна функціонувати система. Це, перш за все, україномовний та російськомовний сегмент Інтернет. Враховуючи складність визначення його розміру, для приблизної оцінки можна скористатися даними інформаційного бюлетеню [23], який описує характеристики сайтів українською, російською, білоруською та казахською мовою, а також будь-якою мовою, розміщених на національних доменах .am, .az, .by, .ge, .kg, .kz, .md, .ru, .su, .tj, .ua та .uz станом на 2009 рік. За його даними число сайтів складає 15 млн., або 6.5 % усього Інтернету, з нього текст складає 140 тисяч гігабайт. (200 тис. Гб. з урахуванням дублювання інформації), причому 88% тексту міститься на менш ніж 1% сайтів. Надалі під терміном «Web» будемо розуміти його вищезгаданий сегмент.

По-друге, варто врахувати частоти змін сторінок, які потрібно відслідковувати. З роботи [24] видно, що переважна більшість сторінок або взагалі не змінюється, або зміни відбуваються надзвичайно рідко.

Система складається з двох частин: моніторингової системи і краулера, який поповнює списки моніторингу. Нехай пропускна здатність каналу доступу до Інтернет 1 Гбіт/с, яка є типовою для корпоративного середовища. Нас цікавить лише інформація на певну тему, тому обсяг даних для завантаження по каналу можна значно зменшити шляхом застосування сфокусованого краулера. Через властивість Web-сторінок на одну тему утворювати взаємопов’язані підмножини-кластери, сфокусований краулер може уникати завантаження нетематичних сторінок. Крім того, ¼ сторінок Web займає пошуковий спам, здебільшого орієнтований на пошуковики, які ранжують сторінки за популярністю посилань. Оскільки наш критерій це оцінка змісту, ми можемо відфільтрувати такі сторінки.

Згідно [24] існує значна диспропорція у розподілі частот змін сторінок Web (див. табл. 1).

Таблиця 1. Частоти оновлення сторінок Web

Номер групи сторінок	1	2	3	4	5	6	7	8	9	10	11	12
Частка сайтів (%)	75,79	11,32	3,23	1,61	0,16	0,16	1,61	0,16	0,16	0,16	1,61	4,03
Середня частота оновлення	0	0,05	0,15	0,25	0,35	0,45	0,55	0,65	0,75	0,85	0,95	1

Під частотою зміни розуміється відношення $v = N_{змін}/(N_{зав} - 1)$ кількості змін сторінки $N_{змін}$ до кількості її успішних завантажень $N_{зав}$ (за виключенням першого, яке ще не мало зразка для порівняння).

Як бачимо, більшість сторінок (75,79%) взагалі не змінюються. По мірі сканування краулером, такі сторінки мають поступово перейти до бази, не потребуючи частого пересканування.

Слід поставити два питання: по-перше, чи встигне краулер обробляти усю необхідну інформацію при заданому каналі і, по-друге, який буде час від появи нової тематичної інформації у Web до її виявлення моніторинговою системою. Нехай краулер виконує повторне сканування щодобово, оскільки затримка виявлення інформації у межах доби є прийнятною для більшості користувачів. При цьому для вже відвіданих раніше тематичних даних враховуються частоти зміни джерел. Нетематичні дані, які не містять корисної для нас інформації, можна оглядати у довільному порядку у вільний від іншого час. Окрім того, слід врахувати те, що з'являються нові Web-сторінки, для яких на момент виявлення частоти зміни невідомі.

Розглянемо перше питання. Візьмемо розмір Web за даними на 2009 рік [23]. На етапі ініціалізації для тренування класифікаторів анти-спаму, тематичності та виявлення дублікатів повністю завантажимо україно-російський сегмент Web обсягом W_{Web} через канал зі швидкістю V, що займе

$$T_{ініц} = \frac{W_{Web}}{V} = \frac{2 \cdot 10^5}{0{,}12} = 1666666{,}67 \text{ сек} = 19{,}29 \text{ діб}$$

Наступним етапом буде оцінка частот змін відібраної тематичної неспамової оригінальної інформації. Цей огляд необхідно продовжувати впродовж періоду, достатнього для заміру частот зміни сторінок. Припустимо, що середня частка одної теми у Web може бути приблизно оцінена за частотою запитів такої інформації при пошуці. Автори [XZ] класифікували запити за тематикою і в середньому частка одної теми складає близько 0,07.

Нехай коефіцієнт відбору $K_{відб}$ складається з коефіцієнтів виключення дублювання інформації [23], коефіцієнту виключення спаму [23andex] та коефіцієнту середньої частки інформації одної тематики у Web:

$$K_{відб} = K_{бездубл} \cdot K_{неспам} \cdot K_{тем} = 0{,}7 \cdot 0{,}75 \cdot 0{,}07 = 0{,}04$$

А сам час щодобового огляду буде

$$T_{щод} = \frac{W_{Web} \cdot K_{відб}}{V} = \frac{2 \cdot 10^5 \cdot 0{,}04}{0{,}12} = 66666{,}67 \text{ сек} = 0{,}77 \text{ діб}$$

Наступним етап -- моніторинг інформації, який полягає у огляді створеного за добу об'єму тематичної інформації, огляду $N_{попер.}$ подібних попередніх об'ємів з оцінкою змінилися вони чи ні, огляд тематичної інформації, яка вже занесена у базу разом з відповідними частотами її зміни, а у решту часу – огляд, на всякий випадок, довільним чином решти інформації Web (нетематичної, спамової і т. п.). Нехай $N_{попер.}$ = 10, оскільки 10 діб достатньо, щоб виявити сторінки, що постійно оновлюються. Приріст тематичної інформації за рік можна приблизно обчислити пропорційно збільшенню кількості сторінок:

$$W_{рік} \approx W_{2000} - W_{1999} = \left(\frac{N_{сайт.2000} \cdot N_{ст.2000}}{N_{сайт.1999} \cdot N_{ст.1999}} - 1 \right) \cdot W_{Web} \cdot K_{відб} =$$

, де

$$= \left(3 \cdot \frac{139}{255} - 1 \right) \cdot 2 \cdot 10^5 \cdot 0{,}04 = 5200 \text{ Гб}$$

$N_{сайт.2000}$, $N_{сайт.1999}$ -- кількість сайтів у відповідні роки (за звітом Яндекса);

$N_{ст.2000}$, $N_{ст.1999}$ -- кількість сторінок на сайті.

Відповідно, приріст тематичної інформації за день $W_{нов} = 14{,}21 \, Гб$.

Час на завантаження обов'язкової (тематичної) інформації $T_{обов.}$ (без урахування часу обходу решти) на початковий день (він буде поступово рости за рахунок $W_{нов}$):

$$T_{обов.i}(i=0)=\frac{W_{нов}\left(+N_{попер}\right)+(W_{Web}K_{відб}+iW_{нов})\sum_i P_iQ_i}{V}=$$

$$=\frac{14{,}21\cdot 11+2\cdot 10^5\cdot 0{,}04\cdot(0{,}76\cdot 0+0{,}11\cdot 0{,}05+...+0{,}04\cdot 1)}{0{,}12}=, \text{ де}$$

$$=\frac{156{,}31+680}{0{,}12}=6969{,}25 \text{ сек}=0{,}08 \text{ діб}$$

$N_{попер.}$ -- кількість попередніх «нових» порцій інформації;

P_i -- частка інформації даної групи у загальному об'ємі;

Q_i -- середня частота зміни (і відвідування) такої інформації.

Розглянемо час на завантаження обов'язкової інформації через 2 роки після початкової точки, тобто у 2011 році при $i \approx 2\cdot 366 = 732$:

$$T_{обов.i}(i=732)=$$

$$=\frac{156{,}31+(2\cdot 10^5\cdot 0{,}04+14{,}21\cdot 732)\cdot(0{,}76\cdot 0+0{,}11\cdot 0{,}05+...+0{,}04\cdot 1)}{0{,}12}=$$

$$=\frac{156{,}31+1404{,}23}{0{,}12}=13004{,}5 \text{ сек}=0{,}11 \text{ діб}$$

Отже, 0,11 < 1 і краулер буде в змозі щодобово оглядати обов'язковий об'єм інформації.

Розглянемо друге питання. Візьмемо розмір тематичного Web на 2009 рік. Спочатку розглянемо математичне очікування часу виявлення інформації на пересічній сторінці, яка відвідується з періодом T. Нам нічого не відомо про конкретний час, коли до неї вносяться зміни, але вони відбуваються в середньому з частотою $1/T$. Припустимо, що час внесення змін має неперервне рівномірне розподілення. Математичне очікування часу виявлення зміни X на окремій сторінці, враховуючи останню обставину, дорівнює:

$$M(X)=\frac{T}{2}$$

Вирахуємо частку змін за добу c від кожної групи сторінок і занесемо у табл. 2, спираючись на те, що зміни відбуваються пропорційно частоті зміни сторінок v:

$$\frac{c_i}{c_j}=\frac{v_i}{v_j},\quad c_j=\frac{c_j}{\sum_i c_i}, \text{ де і, j – номер групи сторінок (див. табл. 2).}$$

Таблиця 2 . Частка змін за добу від кожної групи сторінок

	Вже занесені у базу сторінки												Нові сто-рін-ки
Номер групи сторінок	1	2	3	4	5	6	7	8	9	10	11	12	13
Середня частота оновлення v	0	0,05	0,15	0,25	0,35	0,45	0,55	0,65	0,75	0,85	0,95	1	1*
Півперіод зміни $T/2$	0	10	3,34	2	1,43	1,11	0,91	0,77	0,67	0,59	0,53	0,5	0,5

Частка з числа щодобових змін C	0	0,01	0,02	0,04	0,05	0,06	0,08	0,09	0,11	0,12	0,14	0,14	0,14

* - для нових сторінок частота невідома, під час визначення виконується щоденне відвідування (при найпершому завантаженні вважаємо, що сторінка змінена)

Математичне очікування часу виявлення певної інформації Y від часу її появи у Web:

$$M(Y) = \sum_{i=1}^{13} p(Y_i)Y_i = \sum_{i=1}^{13} c(Y_i)\frac{T_i}{2} = 0{,}01 \cdot 10 + 0{,}02 \cdot 3{,}34 + \ldots + 0{,}14 \cdot 0{,}5 =$$, де

$= 0{,}88$ діб

i -- номер групи сторінок (див. табл. 2);

$p(Y_i)$ -- ймовірність того, що час виявлення буде Y_i;

$c(Y_i)$ -- частка числа змін сторінок одної групи сторінок у їх загальній кількості.

Отже, час від появи інформації на тему, що задана краулеру користувачем, до її виявлення системою буде 0,88 діб < 1 доби, що є прийнятним для пересічного користувача.

Таким чином, стаття присвячена розв'язанню задачі створення системи моніторингу тематичних Web-ресурсів для цілей вітчизняного користувача. У статті запропоновано класифікацію основних алгоритмів обходу ресурсів за ознакою об'єктів, які враховуються при виборі порядку обходу, пропонується класифікація метрик ранжування сайтів за ознакою об'єктів, на основі яких виконується оцінювання. Була проведена попередня оцінка часу виявлення інформації на тему, що задана краулеру користувачем, і показано, що вона є прийнятною для більшості задач.

Література

1. World Wide Web Size. [Електронний ресурс]. – Режим доступу: http://www.worldwidewebsize.com/.
2. Pant G., Srinivasan P., Menczer F. Exploration versus Exploitation in Topic Driven Crawlers. ACM TOIT Volume 4 Issue 4. -- 2004.
3. Risvik K. M., Michelsen R.. Search Engines and Web Dynamics. Computer Networks Vol. 39 Iss. 3. – Elsevier. – 2002.
4. Aggarwal C. C., Al-Garawi F., Yu P. S. On the Design of a Learning Crawler for Topical Resource Discovery. ACM TOIS Vol. 19 Issue 3. -- 2001.
5. Chakrabarti S., van den Berg M., Dom B. Focused crawling: a new approach to topic-specific Web resource discovery. WWW 1999. – Toronto, 1999.
6. Menczer F., Pant G., Srinivasan P. Topical Web Crawlers: Evaluating Adaptive Algorithms. ACM TOIT Vol. 4 Iss. 4. -- 2004.
7. Micarelli A. Gasparetti F. Adaptive Focused Crawling // The adaptive web. -- Springer-Verlag Berlin. -- 2007.
8. Rennie J., McCallum A. K. Using reinforcement learning to spider the web efficiently. Proc. International Conference on Machine Learning (ICML). – 1999.
9. Cho J., Adams R. E. Page Quality: In Search of an Unbiased Web Ranking. Proc. ACM SIGMOD. – 2005.
10. Baeza-Yates R., Castillo C., Marin M., Rodriguez A. Crawling a Country: Better Strategies than Breadth First for Web Page Ordering. WWW 2005. -- Chiba.
11. Ester M., Kriegel H. P., Schubert M. Accurate and Efficient Crawling for Relevant Websites. Proc. 30th International Conference on very large Databases (VLDB 2004). – Toronto. -- 2004.
12. Pant G., Menczer F. Topical Crawling for Business Intelligence. ECDL. – 2003.
13. Hesham A. Self Ranking and Evaluation Approach for Focused Crawler Based on Multi-Agent System. IAJIT. – 2008.
14. Liu H., Milios E., Janssen J.. Probabilistic Models for Focused Web Crawling. WIDM 2004. -- 2004.

15. Baeza-Yates R., Castillo C. Balancing Volume, Quality and Freshness in Web Crawling. In Hybrid Intelligent Systems 2002. -- IOS Press. – Santiago. -- 2002.

16. Kleinberg J. M., Kumar R., Raghavan P., Rajagopalan S., Andrew S. Tomkins .The web as a graph: Measurements, models, and methods. Proc. International Conference on Combinatorics and Computing #1627 in LNCS. -- Springer-Verlag . -- 1999.

17. Rennie J., McCallum A. K. Using Reinforcement Learning to Spider the Web Efficiently. Proc. ICML-99. -- Morgan Kaufmann Publishers. -- San Francisco.

18. Diligenti M., Coetzee F. M., Lawrence S., Giles C. L., Gori M. Focused Crawling Using Context Graphs. Proc. VLDB 2000.

19. De Bra P., Houben G. J., Kornatzky Y., Post R. Information Retrieval in Distributed Hypertexts. Proc. IMIRSM RIAO 94.-- New York . – 1994. -- pp. 481-491.

20. De Vocht J., Experiments for the Characterization of Hypertext Structures. Masters Thesis. -- Eindhoven Univ. of Technology. --1994.

21. Rungsawang A., Angkawattanawit N. Learnable topic-specific web crawler. -- J. Network and Computer Applications 28(2). – 2005. -- pp. 97-114

22. Aggarwal C. C. Collaborative Crawling: Mining User Experiences for Topical Resource Discovery. Proc. KDD Conference. -- 2002.

23. Контент Рунета. Yandex. [Електронний ресурс]. – Режим доступу: http://download.yandex.ru/company/yandex_on_content_autumn_2009.pdf.

24. Kim S. J., Lee S. H. An Empirical Study on the Change of Web Pages. Proc. the 7th APWeb. 2005. -- 632-642

Замятин Д. С., Михайлюк А. Ю., Михайлюк В. А., Петрашенко А. В. Сфокусированный краулинг как средство снижения ресурсоемкости поиска в Web

В статье исследуется проблема создания системы мониторинга тематических Web-ресурсов для корпоративной среды. Предложена классификация основных алгоритмов обхода ресурсов. Разработана классификация метрик ранжирования сайтов по признаку объектов, на основе которых выполняется оценивание. Проведено предварительный рассчет оценки пригодности сфокусированного краулинга для данной задачи. Ист. 24

Ключевые слова: поисковые системы, Web краулеры, тематический поиск.

Zamyatin D. Mykhailiuk A. Mykhailiuk V. Petrashenko A. Focused search as a mean of Web search resource usage.

The article examines the problem of topical Web-resources monitoring for corporate environment. Classification of major crawling algorithms is proposed. Classification of rank metrics by objects which are taken into consideration is developed. Preliminary calculation of focused crawling fitness estimation for chosen task is carried out. Ref 24

Keywords: search engines, Web crawlers, topical search.

Стаття подана
06.04.2011.

Замятін Денис Станіславович, к.т.н., доцент кафедри спеціалізованих комп'ютерних систем Національного технічного університету "Київський політехнічний інститут"
Михайлюк Антон Юрійович, кандидат технічних наук, завідувач науково-дослідної лабораторії інформатизації освіти Київського університету імені Бориса Грінченка
Михайлюк Вадим Антонович, магістрант кафедри спеціалізованих комп'ютерних систем факультету прикладної математики Національного технічного університету "Київський політехнічний інститут"
Петрашенко Андрій Васильович, кандидат технічних наук, доцент кафедри спеціалізованих комп'ютерних систем Національного технічного університету "Київський політехнічний інститут"

Рецензент Тарасенко Володимир Петрович, д.т.н., проф., Національний технічний університет "Київський політехнічний інститут"