1. Дюк В., Самойленко А. Data Mining: Учебный курс. – СПб: Питер, 2001. – 368 с. 2. Han J., Kamber M. Data Mining: Concepts and Techniques. – Simon Fraser University, 2000. 3. А. Шахиди. Data Mining – добыча данных // http://www.basegroup.ru. 4. New Zealand Digital Library // http://www.cs.waikato.ac.nz/~nzdl. 5. Marti A. Hearst. Untangling Text Data Mining. // http:www.sims.berkeley.edu/~hearst. 6. Lent B. Discovering trends in text databases // www.almaden.ibm.com. 7. Agrawal R., Bayardo R., Sricant R. Athena: Mining-based interactive management of text databases // www.almaden.ibm.com. 8. Ahonen H., Heinonen O., Klemettinen M., Verkamo I. Mining in the Phrasal Frontier // 1st European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'97), Trondheim, Norway, 1997. 9. Ahonen H., Heinonen O., Klemettinen M., Verkamo I. Applying Data Mining Techniques in Text Analysis.// Report C-1997-23, University of Helsinki, Department of Computer Science, 1997. 10. Киселев М. Средства добычи знаний в бизнесе и финансах // Открытые системы. – 1997. – № 4. – С. 41–44. 11. Буров К. Обнаружение знаний в хранилищах данных // Открытые системы. – 1999. – № 5–6. – С. 67–77. 12. Белоногов Г.Г., Кузнецов Б.А. Языковые средства автоматизированных информационных систем. – М.: Наука, 1983. – 290 с. 13. Mark Dixon. An Overview of Document Mining Tecnology. 14. Шахиди А. Введение в анализ ассоциативных правил // http://www.basegroup.ru.

УДК 681.3

А. М. Пелещишин
Національний університет "Львівська політехніка",
кафедра "Інформаційні системи та мережі"

ПОШУК, КОНСОЛІДАЦІЯ ТА АНАЛІЗ ДАНИХ У ГЛОБАЛЬНІЙ СИСТЕМІ WORLD WIDE WEB

This paper considers main problems of information searching, consolidation and analysis in World Wide Web. Survey of present researchs has done. Some approaches to resolve these problems are proposed.

Розглядаються проблеми пошуку, консолідації та аналізу інформації, що існує у глобальній службі WWW. Зроблено огляд стану існуючих досліджень і запропоновано деякі підходи до вирішення вказаних проблем.

1. ПОСТАНОВКА ЗАДАЧІ

Інтернет та глобальна система WWW сьогодні є безпрецедентним і унікальним інформаційним ресурсом, який об'єднує в собі величезні об'єми інформації. Проте сьогодні слабо розроблені реально працюючі механізми ефективного використання цього масиву інформації. Фактично існуючі системи обмежуються елементарним пошуком ресурсів, що, можливо, є релевантними запитам. Складніша синтетична обробка даних, що містяться у WWW, лише започатковується.

Кількість користувачів Інтернет зараз у світі сягає півмільярда і продовжує швидко зростати. Це зростання відбувається не тільки за рахунок традиційно високотехнологічних

країн (Північної Америки, Європи, далекосхідного регіону). Останнім часом спостерігається стрімкий ріст популярності Інтернет і в країнах третього світу. Але, з другого боку, через декілька років слід очікувати поступового сповільнення росту популярності Інтернет аж до стабілізації на рівні показників природного приросту населення планети.

Другим важливим показником є кількість атомарних інформаційних ресурсів з унікальним суттєвим наповненням (сторінок). Сьогодні основними джерелами такої інформації є пошукові машини, роботи яких сканують велику кількість сторінок; наприклад у базі даних Google (найповніша з усіх пошукових систем) зареєстровано понад 3 мільярди сторінок. Але немає підстав вважати, що це число в повній мірі характеризує кількість сторінок у WWW. Імовірно, що реальна кількість сторінок значно (у рази) переважає 3 мільярди.

Складно також визначити орієнтовну кількість логічних об'єднань сторінок у інформаційні системи (сайти). Найповніший світовий каталог сайтів ODP сьогодні містить у своїй базі даних інформацію про понад 3 мільйони сайтів. Проте зауважимо, що високі вимоги, що ставляться до сайтів, описаних у світових каталогах та "ручний" принцип їхнього поповнення значно обмежують розміри каталогів. Тому реальна кількість існуючих сайтів на порядки перевищує кількість описаних в ODP чи в інших світових каталогах.

Жодних підстав вважати, що ріст кількості сторінок чи сайтів зупиниться чи хоча б сповільниться, сьогодні немає.

2. АНАЛІЗ ДОСЛІДЖЕНЬ

Великі обсяги інформації та розміри аудиторії, що наповнює та використовує WWW, і визначають базові напрями наукових досліджень з розв'язання основних проблем Інтернет і WWW. Головними з них є:

- **Пошук інформації в Інтернет.** Проблема породжується надзвичайно великими об'ємами даних в WWW.
- **Аналіз аудиторії Інтернет та конкретних сайтів.** Проблема породжується великою глобальною аудиторією користувачів Інтернет.
- **Організація ефективного обміну інформацією в World Wide Web**. Проблема породжується відсутністю ефективних стандартів обміну даними в більшості розподілених систем, що функціонують у Інтернет і WWW та необхідністю обробки великих розподілених масивів даних.

Дослідження у наведених напрямках становлять значну частину сучасних досліджень у галузі Веб-технологій. Серед них слід виділити дослідження, що лягли в основу сучасних пошукових систем [1, 4, 5, 6, 9, 10, 13, 14, 15], дослідження з організації моніторингу та інтелектуального аналізу трафіка сайтів [3, 7, 16, 17], дослідження з відкритих стандартів обміну інформацією та консолідації даних у WWW [7, 10, 15, 16, 21, 25, 26, 27], з розширення алгебр та мов запитів [8, 11, 12, 17, 18, 19, 20, 22, 23].

Проте означені вище дослідження часто мають обмежений характер і торкаються лише деяких аспектів комплексної задачі ефективної та корисної обробки даних, що містяться в WWW та в супутніх базах даних. Такий підхід, як правило, призводить до отримання лише часткових результатів, застосування яких на практиці є функціонально обмеженим та дає лише короткочасну вигоду [3, 16].

3. МЕТА ДОСЛІДЖЕННЯ

Досліджено основні комплексні підходи до розв'язання актуальних задач пошуку, аналізу та консолідації даних у глобальному інформаційному просторі World Wide Web.

4. ОСНОВНИЙ МАТЕРІАЛ

4.1. Пошук інформації в Інтернет

Актуальним напрямом досліджень, що породжується об'ємами інформації у WWW, є організація пошукових систем, що дозволяють здійснювати пошук інформації у WWW або певних його сегментах. Фактично цей напрям досліджень є базовим для інших досліджень як WWW взагалі, так і принципів побудови сайтів зокрема.

Ряд інших досліджень у галузі Веб-технологій базуються на результатах досліджень щодо пошуку інформації в Інтернет. Крім того, багато досліджень визначаються не тільки теоретичними результатами, а й наслідками практичного впровадження теоретичних результатів.

Так, зокрема, часто зазначають значний вплив змін у пошукових технологіях на принципи структурування сайтів та на методи організації міжсайтової взаємодії.

Використання пошукових систем (пошукових машин) сьогодні є найпопулярнішим методом навігації користувача Інтернет по системі World Wide Web (не враховуючи прямих заходів на сайти). При аналізі прямих заходів слід також враховувати, що першого разу відвідувач потрапив на сайт за якимсь посиланням (або його офлайновим аналогом), і більша частина таких "перших" переходів здійснюється саме з пошукових систем. Отже, пошукові системи сьогодні є головним засобом організації системи навігації по WWW.

Сам по собі пошук інформації в гіпертекстових масивах даних є достатньо добре розв'язаною з теоретичної точки зору задачею. Проте насправді задача у випадку World Wide Web радикально ускладнюється величезними об'ємами вхідної результуючої інформації. Так, на світових пошуковцях результатом пошуку за поширеними у мові словами чи словосполученнями можуть бути мільйони сторінок. Зрозуміло, що пересічного користувача Інтернет може цікавити лише незначна частина сторінок, що відображають зміст запиту. Для виділення з множини релевантних сторінок найважливіших пошуковими системами сьогодні використовується рангування сторінок у результуючій множині.

При рангуванні сторінок пошукові системи використовують різні технології [1, 9, 15, 10], які поділяються на декілька груп, з яких основними є:

- Системи на основі лексичного підходу.
- Системи на основі структурного гіпертекстового підходу.
- Системи на основі оцінки популярності ресурсів.
- Системи на основі експертного оцінювання ресурсів.

Першу групу становлять системи, що використовують лінгвістичний підхід до впорядкування інформації. Для кожної сторінки оцінюється ступінь релевантності її тексту (та деяких інших текстів – зокрема текстів зовнішніх посилань на сторінку) до запиту. Далі формується список релевантних сторінок на основі ступеня релевантності.

При використанні лінгвістичного підходу до рангування результатів пошуку використовуються лінгвістичні методи, які дозволяють на основі тексту сторінки робити припущення щодо її відповідності запиту (релевантності). Зокрема аналізується:

- частота появи ключових слів запиту в сторінці;
- частота появи словоформ ключових слів у сторінці;

- частота появи синонімів до ключових слів у сторінці;
- робиться спроба співставлення семантики запиту до семантики сторінки.

Крім того, можуть аналізуватися ключові слова, транслітеровані з одного алфавіту в інший, переклад слів з однієї мови на іншу та інші модифікації ключових слів.

Слід зазначити, що лінгвістичний аналіз гіпертекстових Веб-сторінок суттєво відрізняється від традиційного лінгвістичного аналізу природного тексту. Ці відмінності породжуються двома обставинами:

- Додатковими властивостями та характеристиками гіпертексту.
- Активністю середовища, що підлягає аналізу.

Перша обставина дозволяє набагато повніше проводити лінгвістичний аналіз сторінок. До тексту додаються додаткові властивості [5, 15], зокрема:

- Структурні характеристики тексту (заголовок тексту, заголовки розділів, назви зображень та вбудованих об'єктів).
- Семантичні характеристики тексту (виділення важливих фрагментів тексту, кодів, цитат, адрес).
- Візуальні характеристики тексту (розміри та стилі написання, колір, позиціонування на екрані, шрифти).
- Тексти зовнішніх посилань на сторінку.
- Узагальнені характеристики тексту – метадані сторінки (ключові слова, короткий опис, частота оновлення).

Перераховані додаткові характеристики визначаються розробником сторінки за допомогою спеціальних тегів розмітки тексту.

На основі додаткових характеристик тексту пошуковій машині вдається побудувати точніші та ефективніші критерії рангування [5]. Ці критерії є синтетичними та поєднують у собі вищенаведені характеристики. Оцінка відповідності запиту (релевантності), за яким і здійснюється рангування, визначається як функція виду

$$\mathrm{Rel}(Pg,Sq)=\mathrm{Rel}(\mathrm{Text}(Pg,Sq),\mathrm{Str}(Pg,Sq),\mathrm{Sem}(Pg,Sq),\mathrm{Vis}(Pg,Sq),\mathrm{Ref}(Pg,Sq),\mathrm{Meta}(Pg,Sq)),$$

де Pg – сторінка; Sq – пошукова фраза; $\mathrm{Rel}(Pg,Sq)$ – релевантність сторінки пошуковому виразу; $\mathrm{Text}(Pg,Sq),\mathrm{Str}(Pg,Sq),\mathrm{Sem}(Pg,Sq),\mathrm{Vis}(Pg,Sq),\mathrm{Ref}(Pg,Sq),\mathrm{Meta}(Pg,Sq)$ – релевантність основного тексту, структурних елементів тексту, семантики тексту, тексту з врахуванням візуальних характеристик, текстів зовнішніх посилань та метаданих сторінки пошуковому виразу відповідно.

Очевидним способом визначення такої функції є лінійна згортка показників:

$$\mathrm{Rel}(Pg,Sq)=\sum_{f\in F}K_f\,f(Pg,Sq),\ \text{де}\ F=\{\mathrm{Text},\mathrm{Str},\mathrm{Sem},\mathrm{Vis},\mathrm{Ref},\mathrm{Meta}\},$$

де K_f – коефіцієнти згортки при відповідних показниках релевантності, які визначаються розробниками та адміністраторами пошукових систем.

Як правило, великими є значення коефіцієнтів при показниках релевантності деяких структурних тегів (зокрема заголовків сторінки та її розділів) та коефіцієнтів релевантності текстів зовнішніх посилань.

Незначними є значення коефіцієнтів при узагальнених показниках тексту та при структурних показниках, що відображаються лише при певних умовах (типу атрибутів ALT вбудованих об'єктів).

Оцінювання візуальних характеристик тексту базується на принципі – “чим краще видно – тим краще” [5, 15], проте сьогодні воно значно ускладнюється поширенням технологій верстання на основі каскадних таблиць стилів та програмних кодів, що виконуються в середовищі броузера (JavaScript, Flash тощо). Пошукові системи часто не в стані встановити істинні візуальні характеристики тексту, тому деякі системи просто ігнорують їх або ж визначають малі ваги для візуальних характеристик.

Поза увагою пошукових машин залишається текстова інформація, розміщена в графічних растрових файлах та (за незначним винятками) в інших мультимедійних файлах.

Як видно з наведеного вище списку, пошукові системи, що використовують лінгвістичні підходи до рангування результатів пошуку, при визначенні рангу практично повністю базуються на внутрішній інформації, що розміщена на сторінці, або на інформації, що може виявитися недостовірною. Це призводить до можливості *спаму* пошукових систем.

Поява спаму, орієнтованого на засмічення пошукових систем, обумовлена в першу чергу значною очевидною диспропорцією між кількістю інформаційних ресурсів Інтернет та кількістю користувачів. Сьогодні уже існує багатократна перевага кількості ресурсів над кількістю користувачів. І ця диспропорція буде лише підсилюватися (враховуючи закономірне сповільнення росту аудиторії Інтернет та постійне зростання кількості ресурсів).

Тобто, можуть існувати сторінки, на які впродовж довгого часу (а можливо взагалі ніколи) не заглядав жоден відвідувач. Проте для багатьох сайтів така ситуація є неприйнятною, і їх адміністрація вживає відповідних заходів. Враховуючи те, що більшість користувачів Інтернету користується для навігації пошуковими системами, доцільним і логічним виглядає спроба збільшення рангу сторінок сайту в результатах пошуку пошукових машин.

Враховуючи “замкнутість” пошукових систем “лінгвістичного” класу на внутрішньому вмісті сторінки, покращення досягається достатньо просто – потрібно змінити текст сторінок сайту та його додаткові характеристики. Сама по собі така модифікація не є некоректною чи неетичною. Як правило, такі модифікації лише покращують читабельність, зручність сторінок та чіткіше окреслюють їхню семантику. Проте виникає також можливість спаму – побудови спеціальних ресурсів, єдиною метою яких забезпечення високих позицій в результатах пошуку.

З метою захисту від спаму в пошукових машинах використовуються дещо модифіковані алгоритми обчислення релевантності, які дозволяють визначати та відсіювати сторінки, що містять:

- незв’язний текст;
- слова з надмірною частотою появи;
- автогенерований текст;
- текст, що має неприродні візуальні властивості (нечитабельний або, навпаки, текст з надзвичайно великими літерами тощо);
- примітивну переадресацію на інші сторінки.

Значна частина можливих модифікацій властивостей тексту спамерського характеру сьогодні практично не може бути розпізнана за допомогою автоматизованих алгоритмів (використання зовнішніх таблиць стилів, складних програмних одиниць, що виконуються в середовищі). У такому разі доводиться використовувати ручну модерацію ресурсів.

Проте активність текстового середовища є тим новим фактором, що робить практично неможливим ефективне використання лише лінгвістичних методів в пошуку Веб-сторінок в

Інтернет. Дійсно, які б не були досконалі методи рангування сторінок у результатах пошуку, якщо вони базуються виключно на інформації, розміщеній всередині сторінки, автор сторінки завжди може модифікувати сторінку так, щоби вона була максимально релевантною бажаному запиту. Навіть врахування текстів зовнішніх посилань не здатне змінити ситуацію – адже сьогодні створення додаткових допоміжних сторінок з потрібними текстами посилань не складає жодних труднощів і може бути навіть повністю автоматизоване.

На перешкоді створенню сторінок “спамерського” класу стоїть закритість алгоритмів рангування в пошукових системах. Невідомість деяких параметрів (зокрема, максимально допустимої щільності ключових слів) ускладнює задачу створення оптимальних сторінок. Проте існує можливість створення значної кількості сторінок, які будуть відрізнятися між собою лексичними показниками в певних розумних межах. І тоді за результатами рангування можна встановити практично оптимальні показники (зокрема ту ж максимально допустиму щільність ключових слів).

Отже, лише за умови жорсткої ручної модерації пошукові системи лінгвістичного класу здатні давати задовільні результати пошуку за словами та словосполученнями, які вважаються “конкурентними”. Як правило, такими є широковживані слова, назви товарів та послуг, популярні власні назви та загальні поняття. Хоча за словами та словосполученнями, які не мають комерційного забарвлення, або за багатослівними словосполученнями результати роботи таких пошукових систем можуть бути задовільними та успішними.

Якщо для систем, що розглядалися вище, ідеологічною основою роботи було “знаходження максимально релевантного ресурсу до запиту користувача”, для систем інших класів є “знаходження достатньо релевантних ресурсів при мінімумі пошукового сміття”.

Такий результат досягається за допомогою введення до розгляду додаткових характеристик Веб-сторінок, які не мають прямого відношення до запиту користувача. З одного боку, це може призводити до пониження релевантності тексту сторінок, що ранговані найвище, до запиту. Проте, з іншого боку, ці характеристики усувають з розгляду сторінки, що мають явно спамерський характер і відповідно знаходяться поза “корисною” частиною WWW.

Другу групу пошукових систем становлять системи, що рангують результати на основі структури WWW, яка визначається гіпертекстовими посиланнями. У системах такого класу для сторінки визначаються додаткові характеристики. Це може бути “вага” сторінки, її “авторитетність” тощо.

Найпопулярнішим сьогодні підходом по “зважуванню” сторінки є підхід, запропонований Lawrence Page [5, 6]. Цей підхід базується на визначенні величини PR (Page Rank, ранг сторінки) для кожної сторінки. PR характеризує імовірність перебування відвідувача на сторінці. Для визначення цієї імовірності використовується структура посилань WWW. Фактично, будується математична модель навігації користувача по посиланнях. Модель базується на апараті ланцюгів Маркова. PR отримується як результат розв’язання відповідної СЛАР (враховуючи величезні можливості, можуть використовуватися певні спеціалізовані алгоритми або взагалі – пряме моделювання ланцюга Маркова; у будь-якому разі не ставиться вимога отримання точного розв’язку системи).

У первинному варіанті вирази для обчислення PR мають такий вигляд

$$PR(Pg_i) = \frac{(1-d)}{N} + d\sum_{j=1}^{N} Link(Pg_j, Pg_i)\left(\frac{PR(Pg_j)}{C(Pg_j)}\right)$$

де

$$\sum_{j=1}^{N} PR(Pg_j) = 1$$

$$Link(Pg_j, Pg_i) = \begin{cases} 1, & \text{Pg}_\text{j} \text{ посилається на Pg}_\text{i} \\ 0, & \text{Pg}_\text{j} \text{ не посилається на Pg}_\text{i} \end{cases}$$

$LR(Pg_j) = \dfrac{PR(Pg_j)}{C(Pg_j)}$ – величина (ранг посилання), що передається зі j-ї сторінки по одному посиланню; $C(Pg_j)$ – кількість посилань на j-й сторінці; $PR(Pg_j)$ – ранг j-ї сторінки.

Хоча кожна пошукова система, імовірно, використовує власні модифікації.

Найвідомішою пошуковою системою, що використовує PR для рангування, є Google (http://google.com/) – найпотужніша за багатьма критеріями (об'єм БД, частота переіндексування, популярність серед користувачів) сучасна пошукова машина. Ряд інших провідних світових та тематичних пошукових систем також користуються алгоритмами рангування, що використовують PR.

При використанні PR при рангуванні пошукова машина обраховує повну релевантність сторінки до запиту як агреговану монотонну функцію, параметрами якої лінгвістична релевантність, ранг сторінки, лінгвістична релевантність зовнішніх посилань та їхні ранги та лінгвістична релевантність сторінок, на яких містяться ці посилання.

$$\text{Rel}(Pg, Sq) = \text{Rel}(\text{LRel}(Pg, Sq), \text{PR}(Pg), \text{LR}(Pg), \text{LnRel}(Pg, Sq)),$$

де $\text{LRel}(Pg, Sq)$ – лінгвістична релевантність сторінки; $\text{LnRel}(Pg, Sq)$ – лінгвістична релевантність зовнішніх посилань на сторінку.

Виділення лінгвістичної релевантності зовнішніх посилань обумовлене тим, що кожне посилання має окрему вагу і, відповідно, ступінь врахування своєї лінгвістичної релевантності запиту.

Ранг сторінки в такому разі є мірою врахування лінгвістичної релевантності сторінки запиту.

Пошукові системи структурного класу мають суттєво вищий захист від пошукового спаму. Дійсно, ті методи брутального підвищення релевантності сторінок до певних запитів, які застосовуються до попереднього класу пошуковців, виявляються неефективними. Це обумовлюється такими факторами.

- Вимагається PR сторінки, більший за певну величину. Пошукові системи в принципі не включають в результати пошуку за "конкурентними" запитами сторінки з малим PR.
- При рангуванні за "конкурентними" запитами вирішальну роль грає не лінгвістична релевантність, а показники PR сторінки, PR сторінок, що посилаються, рангів посилань та текстів посилань.
- При рангуванні за неконкурентними запитами (особливо за багатослівними, зі складними умовами, на національних мовах тощо) сторінки зі спамом мають шанс потрапити в результуючу вибірку (за умови точної відповідності). Проте такі сторінки ефективні лише за умови перекриття значної множини таких запитів (за формулою "один запит – одна сторінка"). А це вимагає індексування пошуковою системою великого числа сторінок спаму. Тому у пошукових системах використ-

товується правило пріоритетності індексації сторінок згідно з рангом посилань на сторінку. Отже, більша частина спамерських сторінок взагалі не потрапляє до БД пошукової системи.

Як видно з вищесказаного, пошукові системи цього класу на відміну від "суто лінгвістичних" систем беруть за основу інформацію, яка є зовнішньою стосовно сторінки і, відповідно, володіє певною незалежністю та об'єктивністю. Незважаючи на те, що сайт (керований одним власником) може містити значну кількість сторінок, сторінки з високим PR можуть з'являтися лише за умови значної кількості зовнішніх посилань (з інших сайтів) на сторінки сайту (явище накопичення PR буде далі розглянуто). Сторінки спаму фактично втрачають сенс, адже для забезпечення PR на них повинні вказувати зовнішні посилання з авторитетних сторінок. У такому разі брутальний спам легко ідентифікується модераторами БД пошукових машин. Для власників сайтів наявність посилань на сторінки з явним спамом є неприпустимою – це компрометує сайт в очах відвідувачів; сайт наражається на небезпеку ігнорування пошуковою системою (принцип "поганих сусідів").

Певною слабкістю алгоритму PageRank є апріорне (і некоректне) припущення щодо рівноцінності усіх посилань, що ведуть зі сторінки, незалежно від їхніх візуальних характеристик. Можна припустити, що в майбутньому розробники пошукових систем при визначенні рангів посилань використовуватимуть і їхні візуальні характеристики.

Рангування на основі PageRank окрім явних переваг має ряд недоліків і побічних ефектів, зокрема такі:

- Сайти зі сторінками з високим PR, що є реально слабкорелевантні запиту, часто опиняються вище за дійсно релевантні, але з меншим PR.
- Сторінки (як правило з високим PR) можуть часто знаходитися за такими запитами, які є в принципі не передбачені власником сторінки (лише за рахунок текстів і рангів зовнішніх посилань). Це є певним джерелом небезпеки, а саме можлива компроментація сайту – його високе рангування за небажаними словами (нецензурні вислови, образливі терміни, поняття з негативним змістом).

Для ефективного обчислення PR пошукові системи за умов величезних об'ємів інформації змушені використовувати спеціальні алгоритми, які імовірно базуються на ряді спрощень, допусків та додатково визначеній початковій інформації. У такому разі величезного значення набувають каталоги сайтів, які крім своєї основної функції (надання користувачам адрес та описів сайтів) грають велику роль при обраховуванні PR (імовірно, що існує особливе визначення рангів посилань зі світових каталогів). Саме існування каталогів якісних ресурсів є запорукою стійкості при розв'язанні складних обчислювальних задач щодо визначення PR.

З метою повнішого врахування системної функції каталогів ресурсів розроблено алгоритм HITS [4, 13, 14], який розглядається як альтернативний до PageRank. Алгоритм також базується на дослідженні структури WWW на основі посилань. Виділяються додаткові характеристики інформаційних ресурсів. Це "авторитетність" ресурсу (authority weight) та його "концентрованість" (hub weight). Рангування результатів пошуку за бажанням користувача може відбуватися з врахуванням як одного, так і другого показника.

Алгоритм HITS цікавий також тим, що дозволяє отримувати значні додаткові результати по структуруванню WWW та надає нові види послуг при пошуку в WWW [0]. Зокрема, на алгоритмі HITS базується визначення Веб-спілок – сильно зв'язаних сукупностей однотематичних сайтів. При пошуку інформації з'являються ефективні можливості

повторного уточненого пошуку. Проте суттєвим недоліком цього алгоритму вважається його певна нестійкість проти пошукового спаму . Тому сьогодні алгоритм HITS більше застосовується у тематичному пошуку за добре модерованими базами даних гіпертекстових документів (наприклад, при пошуку наукової інформації – http://researchindex.org/)/

Популярність описаних підходів (PageRank, HITS та їхні модифікації) до рангування результатів пошуку відіграє визначальну роль у процесі структурування WWW, який відбувається останні роки. Гіпертекстові посилання з елемента зручності інтерфейсу користувача поступово перетворюються у головний інструмент організації усього інформаційного ресурсу WWW [14]. Зокрема, адміністрації сайтів вводять суворі правила оформлення зовнішніх посилань, зростають вимоги до системи навігації всередині сайту, навіть організовується торгівля посиланнями.

До недоліків пошукових систем цього класу слід віднести неточності пошуку при складних запитах, що складаються з кількох слів. У такому разі можливе завищення у результатах пошуку реально слаборелевантних сторінок, котрі проте мають високі системні показники.

Часткове усунення цього недоліку можливе шляхом уточнення виду функції інтегрованої релевантності.

Визначення системних показників (таких, як PR) дає можливість також розв'язати ряд технічних проблем, що постають при побудові пошукових машин. Це зокрема:

- Інтелектуальна поведінка робота пошукової системи [2];
- Виявлення плагіату в WWW.

Основними проявами першої проблеми є визначення терміну повторного заходу робота на сторінку та глибини індексації сайту.

Без визначення системних показників сайту та його сторінок задача визначення частоти заходів на сторінку роботом здійснювалася на основі:

1) мета-інформації сторінки (мета-тег "Revisit-after");

2) прогнозованої частоти оновлення сторінки на базі попередніх даних.

Обидва варіанти практично не передбачають захисту від некоректних неправдивих показників, які характерні для пошукового спаму. Крім того, використання другого варіанта може призвести до "зациклення" робота на сторінці з динамічно оновлюваною інформацією (сторінки з лічильниками, з випадковим текстом, сторінки форумів та чатів).

При відомих системних показниках сторінки пошукова система до вказаних критеріїв застосовує додатково і "вагу" сторінки. Тоді "вага" сторінки визначає допустиму верхню межу можливої частоти переіндексації сторінки роботом. Отже, для важливих та популярних сторінок (наприклад, авторитетних інформаційних агентств) зберігається часта переіндексація, а пошуковий спам та несуттєві динамічні сторінки переіндексовуються в загальному порядку.

Плагіат в WWW пошукова система виявляє за принципом порівняння "авторитетності" сторінок з однаковими або схожими між собою текстами і є побічною дією функції визначення дзеркал сайту. За показник авторитетності може братися або PageRank ресурсу, або authority weight. Оригіналом вважається ресурс з вищою авторитетністю. Ресурси-копії вилучаються з результатів пошуку.

Такий підхід дозволяє вирішити проблему плагіату, коли "слабкий" сайт використовує інформацію з "сильного" (достатньо популярна ситуація). Проте у протилежному випадку ("сильний" краде у "слабкого") такий підхід є неправильним. Щоправда, слід

зазначити, що ця ситуація зустрічається у чистому вигляді досить рідко (як правило, "сильний" сайт при запозиченні інформації зі "слабкого" якимось чином її інтегрує та перетворює, що не дозволяє пошуковій системі вважати інформаційні ресурси тотожними).

Третю групу пошукових систем утворюють системи на основі аналізу популярності інформаційних ресурсів. Формально вони базуються на тій же ідеології, що і системи попереднього класу на основі PageRank. Але замість визначення імовірності перебування користувача на сторінці лише за структурою WWW, вони використовують доступну їм інформацію про реальну популярність ресурсу. Для ресурсів, про які інформація є недоступною, здійснюється екстраполяція на основі гіпертекстових посилань. Тобто, системи цього класу повніше враховують реальну відвідуваність сайтів, виходячи з апріорного твердження, "що на сайт, де є цікава інформація, частіше заходять відвідувачі".

Пошукові системи цього класу є мало поширеними. Основними причинами є:

- Головна теза ("хороший сайт – хороша відвідуваність") далеко не завжди себе виправдовує. Зокрема, високу відвідуваність сайт може мати не за рахунок цікавої та корисної інформації, а за рахунок активної реклами. Крім того, існують можливості організації беззмістовних відвідувань сайту (наприклад, за певну плату).
- Ефективно така система може функціонувати лише за умови масової підтримки власників сайтів (адже на сайтах, що дають початкову інформацію, має бути встановлений "лічильник" даної системи).
- До апаратного забезпечення системи висуваються високі вимоги – адже вона в реальному часі повинна обліковувати відвідувачів сайтів.

Для усунення наведених обмежень сьогодні пропонується інший спосіб обліку відвідувачів сайтів – встановлення спеціального ПЗ на комп'ютерах користувачів, яке б в режимі он-лайн передавало на технічний сервер системи інформацію про відвідування того чи іншого сайту. Таке ПЗ теоретично могло б фільтрувати заходи на сайт за он-лайн рекламою або "просто так", без певної мети.

Четверту групу пошукових систем становлять гібридні системи, де результати пошуку формуються з кількох джерел – інформація, підібрана експертами за дуже конкурентними запитами (найвищий пріоритет при рангуванні), інформація з якісних світових каталогів (середній пріоритет) та інформація з пошукової машини вищенаведених класів (найнижчий пріоритет).

Основними недоліками таких систем є:

- їхня нездатність існувати автономно (без підтримки інших систем);
- висока вартість утримання (витрати на роботу експертів).

Враховуючи недоліки, такі системи орієнтуються в першу чергу на комерційні розділи Інтернету.

Серед модифікацій систем такого класу слід виділити системи пошуку послуг та товарів, які за визначеними експертами правилами збирають потрібну інформацію з Веб-сторінок.

4.2. Аналіз аудиторії Інтернет та конкретних сайтів

Необхідність аналізу аудиторії Інтернет та конкретних сайтів викликана такими обставинами:

- Для оптимізації фізичної структури Інтернет та World Wide Web є необхідною інформація про характеристики (географія, час роботи, об'єкти доступу, технічні характеристики) користувачів.

- Для реалізації проектів з впровадження нових технологій і стандартів в системі WWW необхідно володіти аналітичною інформацією про програмні та технічні засоби, що використовуються користувачами, методи роботи в Інтернет, вікові та соціальні показники, методи підключення до Інтернет.
- Для проведення рекламних і маркетингових кампаній потрібно володіти інформацією про загальні характеристики аудиторії, на котру орієнтовано акцію, об'єкти зацікавлення цієї аудиторії в WWW (сайти, портали, пошукові слова).
- Для перебудови існуючих сайтів необхідно мати інформацію про характер поведінки відвідувача на сайті та поза ним. Зокрема, це шляхи навігації, точки входу на сайт та виходу зі сайту, ключові слова, за якими шукається та знаходиться сайт, показники повернення на сайт та відгуку на сайт.
- Для побудови нових ефективних сайтів потрібно мати інформацію про прогнозовану аудиторію сайту, про методи навігації прогнозованої аудиторії (можливі сайти-партнери, розділи каталогів, пошукові слова та фрази), про технічні та програмні засоби потенційних відвідувачів, регіональні, лінгвістичні, культурні, вікові, соціальні та статеві особливості аудиторії.

Аналіз аудиторії WWW може здійснюватися на двох основних типах масивів даних [16]:

- Внутрішній інформації сайтів (журнали доступу, облік замовлень, опитування та анкети на сайті);
- Зовнішній інформації щодо сайту.

Перший варіант найчастіше є доступний для власників існуючих сайтів, які бажають оптимальним чином перебудувати свій сайт. Особливістю цього варіанта є добра структурованість інформації та її повнота – у журнали доступу заносяться усі події, що відбуваються на сайті.

Інформація журналів доступу до сайту обробляється методами статистичного аналізу та data mining [7].

4.3. Організація ефективного обміну інформацією та її консолідації в World Wide Web

Основним методом видобування інформації сьогодні з World Wide Web є пошук за допомогою пошукових систем. Для цього можуть використовуватися як достатньо прості інтуїтивні, так і складні та потужні мови запитів, які дозволяють достатньо точно специфікувати необхідний ресурс.

Проте пошукові машини є високоспеціалізованими інформаційними системами, що обмежують доступ до даних у WWW лише пошуком конкретних сторінок. Насправді множина можливих операцій над інформацією у WWW набагато ширша ніж пошук. Можливими та доцільними є практично усі види запитів, які існують у традиційних базах даних. Проте на заваді реалізації таких систем – ряд перешкод, зокрема такі фактори [16, 17]:

- Складність та висока вартість програмного та апаратного забезпечення для глобальних систем такого класу, яка значно перевищує витрати на пошукові системи (адже основні принципи, що закладено в пошукових системах, є орієнтовані на ефективне та економне виконання одного виду запитів).
- Відсутність стандартів щодо подання, структурування даних, відсутність єдиних узгоджених класифікаторів.

- Наявність локальних інтересів власників сайтів, які не завжди зацікавлені в інтеграції своїх даних у глобальну систему. Крім того, складними та не повністю вирішеними питаннями є організація розрахунків за обмін інформацією та захист інтелектуальної власності.
- Імовірною також є поява ресурсів спамерського характеру, основною задачею яких є дезінформування користувача з метою досягнення власниками ресурсу певних корисних цілей (заманювання відвідувачів на сайт, усунення конкурентів тощо).

Дослідження щодо консолідації та ефективного обміну інформацією проводяться за такими напрямками:

- Розробка нових стандартів обміну інформацією в глобальному середовищі [21].
- Розробка методів структурування даних, отриманих з слабоструктурованих гіпертекстових документів.
- Розробка методів фільтрації сумнівної та неякісної інформації.
- Розробка принципів взаємодії учасників глобального середовища обміну інформацією (постачальників і користувачів) між собою.
- Розробка мов запитів до World Wide Web.

Розглянемо детальніше кожен з напрямків.

Розробка нових стандартів обміну інформацією в глобальному середовищі. Цей напрям сьогодні в основному зосереджений на побудові мов обміну даними, що базуються на технології XML. Практично для кожної сфери людської діяльності сьогодні розроблено відповідні спеціалізовані мови. Базовими принципами для таких мов є:

- уніфікована ієрархічна деревоподібна структура ресурсу (документа в термінах XML);
- незалежність візуального (звукового, графічного, мультимедійного тощо) відображення від структури та вмісту документа;
- висока спеціалізованість мови.

Основними перевагами рішень на основі технології XML та створених за наведеними принципами мов є:

- простота консолідації глобально розподілених масивів даних;
- семантична відкритість документів;
- можливість конвертації документів з одних мов у інші (технологія XSLT);
- уніфіковані методи ідентифікації атомарних даних у документі (технології XPath, XPointer).

Сьогодні ефективним є застосування XML-технологій при обміні метаданими (мова RDF), при обміні та консолідації новин та журналів (RSS), у векторній графіці та мультимедії (SMIL, SVG), обміні науковою, бібліографічною та комерційною інформацією.

Розробка методів структурування даних, отриманих з слабоструктурованих гіпертекстових документів. Дослідження в цьому напрямку в основному зосереджені на створенні методів виявлення регулярних структур у HTML-сторінках. У ресурсах цього класу зосереджено основні об’єми інформації WWW, проте сама мова слабо структурована, розмітка слабо інформативна та семантично закрита (хоча мова формально відноситься до класу XML-мов). Серед основних типів структур, які доцільно виявляти в таких сторінках є визначення пар “величина – значення” та визначення структур табличного типу [25, 26].

Основними перешкодами при визначенні регулярних структур є:

- відсутність семантики у таблицях HTML;
- некоректне використання таблиць (як засобу візуальної верстки);
- використання графічних образів для передання текстової інформації.

Крім локальної задачі виявлення регулярних структур у межах сторінки важливими є задачі глобальної структуризації, які полягають у визначенні повторюваних регулярних структур на різних сторінках (зокрема з різних сайтів) та знаходженні відображень між різними регулярними структурами, що, як правило, розміщаються на різних сайтах.

Розробка методів фільтрації сумнівної та неякісної інформації. Основними дослідженнями в цьому напрямку є побудова методів та алгоритмів, які б дозволяли фільтрувати недоброякісну або недостовірну інформацію та спам. Підходи, що використовуються при вирішенні цієї проблеми, є аналогічними до підходів, що застосовуються у пошукових системах, а саме

- виявлення лінгвістично некоректної інформації;
- фільтрування за системними ознаками;
- експертна оцінка.

Розробка принципів взаємодії учасників глобального середовища обміну інформацією (постачальників і користувачів) між собою. Ця область сьогодні є вкрай погано дослідженою і на практиці відсутність таких принципів постійно породжує конфлікти інтересів в World Wide Web. Класичними прикладами таких конфліктів є:

- Конфлікти між власниками інтелектуальної власності в WWW;
- Конфлікти між каталогами, пошуковими та метапошуковими системами;
- Конфлікти між спеціалізованими сервісами для адміністраторів сайтів та загальними службами;
- Конфлікти між рекламодавцями та споживачами рекламної продукції в WWW.

Розробка мов запитів до World Wide Web. Крім досліджень у галузі розробки методів і засобів консолідації даних у WWW, активно проводяться дослідження щодо побудови мов запитів до баз консолідованих ресурсів у WWW [18, 24]. Такі мови умовно можна поділити на два основні класи:

- мови класу XML;
- мови квазіреляційного типу.

Мови першого класу орієнтовані на ієрархічне структуроване формування запитів до XML-документів. Основною мовою такого класу сьогодні є XMLQuery. Недоліком цих мов є висока складність формування запиту. Як наслідок, вони можуть застосовуватися в першу чергу при розробці спеціалізованих аплікацій [8, 11, 19].

Мови другого класу орієнтовані на користувача і в своєму синтаксисі намагаються копіювати SQL та інші засоби формування запитів реляційного типу. Проте, на відміну від SQL, мови запитів до WWW повинні враховувати слабку структурованість, розподілений характер і наявність додаткових характеристик середовища. Як наслідок, у таких мовах наявні додаткові синтаксичні конструкції (порівняно з SQL), послаблено типізацію даних та введено нові функції та предикати [12].

Сьогодні серед інших розроблено такі мови запитів квазіреляційного типу: **WebSQL** [20], **Squel** [23], **Quilt**, **UnQL**, **XDuce**, **XML-QL**, **DSQL** [22], **XQL, W3QL** та інші.

ВИСНОВКИ

World Wide Web є сьогодні величезним інформаційним ресурсом, що складається з міліардів сторінок, об'єднаних у сайти. Основні кількісні показники WWW та темпи їхнього збільшення дають підставу виділяти як важливі задачі пошуку, консолідації іноформації, що розміщено у World Wide Web.

Слід зазначити наявність великої кількості досліджень з пошуку інформації та їхніх реалізацій у вигляді глобальних пошукових машин. Проте принциповими обмеженнями систем цього класу є вкрай вузька спеціалізованість та виконання лише деяких базових функцій з обробки даних. Подальший розвиток пошукових систем доцільно, зокрема, проводити з врахуванням результатів теоретичних досліджень з консолідації та аналізу даних у WWW. Зокрема, це використання квазіреляційних підходів до маніпулювання даними та методів класу Data Mining з інтелектуальної аналітичної обробки інформації та синтезу нових даних.

Ефективне розв'язання задач пошуку, консолідації та аналізу інформації вимагає також використання супутніх до WWW баз даних. Це у першу чергу журнали доступу до сайтів, що ведуться на локальному та глобальному рівнях. Побудова та аналіз таких баз даних є однією з основ побудови дійсно ефективних та якісних сайтів в системі World Wide Web.

1. Baeza-Yates R., Riebero-Nero B. Modern Information Retrieval. ACM Press. – New York, 1999. – 501 p. 2. Bayeza-Yates R.,Castillo C. Balancing Volume, Quality and Freshness in Web Crawling. Center for Web Research, Departament of Computer Science, University of Chile, 2002. – 24 p. 3. Bayeza-Yates R.,Castillo C. Relating Web Structure and User Search Behavior. Center for Web Research, Departament of Computer Science, University of Chile, 2002. – 24 p. 4. Borodin A., Roberts G., Tsaparas P., Rosenthal J. Finding Authorities and Hubs From Link Structures on the World Wide Web. Proceedings of WWW10 Conference, Hong Kong, 2001 http://www10.org/cdrom/papers/314/. 5. Brin S., Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Science Department, Stanford University, Stanford, 1998. – 20 p. 6. Brin S., Page L. The Page Rank Citation Ranking: Bringing Order to Web. Computer Science Department, Stanford University, Stanford, 1998. – 17 p. 7. Cooley R., Mobasher B., Srivastava J. Web Mining: Information and Pattern Discovery on the World Wide Web. Departament of Computer Science, University of Minnesota, Minneapolis, 1997. – 10 p. 8. Fernandez M., Simeon J., Wadler P. An Algebra for XML Query. ATT Labs, Bell Labs, Avaya Labs. 2000. – 36 p. 9. Hu W.-C., Chen Y., Smalz M., Ritter G. An Overview of World Wide Web Search Technologies. Departament of Computer Science. Auburn University, 2000. – 6 p. 10. Huang L. A Survey On Web Information Retrieval Technologies. Computer Science Departament, State University of New York. – 2000. – 33 p. 11. Karvounaxakis G., Alcxaki S., Plexousakis D., Scholl M. RQL: A Declarative Query Language for RDF. Technical Report. Greece Institute of Computer Science. – 2001. – 27 p. 12. Kemper A., Query Languages for the WWW. Lehrstuhl fur Dialogorientierle Systeme, Fakultat for Mathematik und Informatik, Universitat Passau. Hauptseminar WS 97/98: Datenbanken – Konzepte und Anwendungen, 1998. – 23 p. 13. Kleinberg J. Authoritative Sources in a Hyperlinked Environment. Proceedings of the ACM-SIAM Symposium on Discrete Algorithms. – 1998. – 34 p. 14. Kleinberg J., Kumar R., Raghavan P., Rajagopalan S., Tomkins A. The Web as a graph: measurements, models, and methods. Lecture Notes in Computer Science. Department of Computer Science, Cornell University, 1999. – 18 p. 15. Kobayashi M., Takeda K. Information Retrieval on the Web. IBM Research, IBM Tokyo Research Laboratory. IBM Japan.

2000. – 47 p. 16. Kosala R., Blockeel H. Web Mining Research: A Survey. Department of Computer Science, Katholieke Universiteit Leuven SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM, 2000. – 15 p. http://w. 17. Lacroix Z., Shahuguet A., Chandrasekar R. Information Extraction & Database Techniques: a user-oriented approach to querying the Web. IRCS & CIS, University of Pennsylvania. – 1998. – 16 p. 18. Mendelzon A., Mihaila G., Milo T. Querying the World Wide Web. Department of Computer Science and CSRI, University of Toronto, Toronto, Canadca, 1996. – 14 p. 19. Pankowski T. XML-SQL: An XML Query Language Based on SQL and Path Tables. Chair of Control, Robotics and Computer Science, Poznan University of Technology. – 2002. – 15 p. 20. Patamarjarnkul A. A Customized Web Search Engine Using a Tiny WebSQL Query Language. Project Report. Auburn University, Alabama May, 2000. – 73 p. 21. Sarawagi S. Nagaralu S. Data mining models as services on the internet. Indian Institute of Technology. Bombay 2000. – 24 p. 22. Sengupta A., Dalkilic M. DSQL – An SQL for Structured Documents. Extended Abstract, Department of A&IS, Kelley School of Business, Indiana University, 2002. – 4 p. 23. Spertus E., Stein L. Squeal: A Structured Query Language for the Web. MIT Artificial Intelligence Lab. Ninth International World-Wide Web Conference, May 2000. – 12p. 24. Stein L., Spertus E., College M. Squeal: A Structured Query Language for the Web. MIT Artificial Intelligence Lab, 2000. – 12 p. 25. Wang Y. Hu J. A Machine Learning Based Approach for Table Detection on The Web. Dept. of Electrical Engineering, Univ. of Washington 2001. – 13 p. 26. Wang Y. Hu J. Detecting Tables in HTML Documents. Dept. of Electrical Engineering, Univ. of Washington. – 2001. 12 p. 27. Witvoet O., Rauber A., Aschenbrenner A., Bruckner R., Putting the World Wide Web into a Data Warehouse:A DWH-based Approach toWeb Analysis. Department of Software Technology and Interactive Systems, Vienna University of Technology,Vienna, Austria. – 2002.

УДК 683.1

А.М. Пелещишин, Т.Б. Гулка
Національний університет “Львівська політехніка”,
кафедра “Інформаційні системи та мережі”

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ПОВЕДІНКИ ВІДВІДУВАЧІВ ВЕБ-САЙТІВ (ФОРМАЛЬНА МОДЕЛЬ)

Web-site visitors behaviour intelligent analysis main concepts are formally defined. Examples of some tasks formalizations are described based on given definitions.

Зроблено формальне визначення основних понять задачі інтелектуального аналізу поведінки відвідувачів веб-сайтів, наведено приклади формалізації ряду задач на основі поданих визначень.

ПОСТАНОВКА ПРОБЛЕМИ

Однією із основних задач інтелектуального аналізу поведінки користувача сайту (зокрема на основі журналів доступу до сайтів) є задача визначення невідомих характе-