2. ВИДИ ПОШУКОВИХ СИСТЕМ WWW

Використання пошукових систем (пошукових машин) на 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 перший раз відвідувач потрапив на сайт по якомусь посиланню (або по його офлайновому аналогу). І більша частина таких "перших" переходів є здійснюється саме з пошукових систем. Таким чином, пошукові системи на 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 по поширених у мові словах чи словосполученнях можуть бути мільйони сторінок. Зрозуміло, що для рядового користувача Інтернет можуть складати інтерес лише незначна частина сторінок, що відображають зміст запиту. Для виділення з множини релевантних сторінок найважливіших пошуковими системами на сьогодні використовується рангування сторінок у результуючій множині.

При рангуванні сторінок пошукові системи використовують різні технології, які поділяються на декілька груп, з яких основними є:

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

2.1. Пошукові системи лінгвістичного класу

Першу групу складають 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 відзначити, 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, 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. Ці критерії є синтетичними та поєднують у собі вище наведені характеристики.

Оцінювання візуальних характеристик тексту базується на принципі – "що краще видно - то краще". Проте на сьогодні значно ускладнюється поширенням технологій верстки 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. На сьогодні уже має місце багатократна перевага кількості ресурсів над кількістю користувачів. І ця диспропорція буде лише підсилюватися (враховуючи закономірне сповільнення росту аудиторії Інтернет та постійне зростання кількості ресурсів).

Таким чином, 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 приймає відповідні заходи. 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 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 володіє неприродними візуальними властивостями (невидимий, або навпаки, текст з надзвичайно великими літерами тощо);
- примітивну переадресацію на інші сторінки.

Значна частина можливих модифікацій властивостей тексту спамерського характеру на 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 таким чином, щоби вона була максимально релевантною бажаному запиту. Навіть врахування текстів зовнішніх посилань

не здатне змінити ситуацію – адже на сьогодні створення додаткових допоміжних сторінок з потрібними текстами посилань не складає жодних труднощів і може бути навіть повністю автоматизоване.

Таким чином, лише за умови жорсткої ручної модерації пошукові системи лінгвістичного класу здатні давати задовільні результати пошуку по словах та словосполученнях, які вважаються “конкурентними”. Як правило, такими є широковживані слова, назви товарів та послуг, популярні власні назви та загальні поняття. Хоча по словах та словосполученнях, які не мають комерційного забарвлення, або по багатослівних словосполученнях 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 безпосереднього відношення до запиту користувача. З одного боку, це може призводити до пониження релевантності тексту сторінок, що ранговані найвище, до запиту. Проте, з другого боку, ці характеристики усувають з розгляду сторінки, що носять явно спамерський характер, і відповідно знаходяться поза “корисною” частиною WWW.

2.2. Пошукові системи структурного класу

Другу групу пошукових систем складають 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 на сьогодні підходом по “зважуванню” сторінки є підхід, запропонований Lawrence Page. 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.

Найвідомішою пошуковою системою, що використовує PR для рангування є Google (http://google.com/) - найпотужніша по багатьох критеріях (об'єм БД, частота переіндексування, популярність серед користувачів) пошукова машина на сьогодні. Ряд інших провідних світових та тематичних пошукових систем також користають алгоритмами рангування, що використовують PR.

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

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 наступними факторами.

- Вимагається PR сторінки, більший за певну величину. Пошукові системи в принципі не включають в результати пошуку по "конкурентних" запитах, сторінки з малим PR.
- При рангуванні по "конкурентних" запитах вирішальну роль грає не лінгвістічна релевантність, а показники PR сторінки, PR сторінок, що посилаються, рангів посилань та текстів посилань.
- При рангуванні по неконкурентних запитах (особливо по багатослівних, з складними умовами, на національних мовах тощо) сторінки

зі спамом мають шанс потрапити в результуючу вибірку (за умови точної відповідності). Проте, такі сторінки ефективні лише при умові перекриття значної множини таких запитів (по формулі “один запит – одна сторінка”). А це вимагає індексування пошуковою системою великого числа сторінок спаму. Тому у пошукових системах використовується правило по пріоритетності індексації сторінок згідно до рангу посилань на сторінку. Таким чином, 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 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 недопустимим – сайт втрачає в очах відвідувачів та попадає 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 наступні:

- Сайти з сторінками з високим PR, що є реально слабо релевантні запиту, часто опиняються вище за дійсно релевантні, але з меншим PR.

- Сторінки (як правило з високим PR) можуть часто знаходитися по таких запитах, які є в принципі не передбачені власником сторінки (лише за рахунок текстів і рангів зовнішніх посилань). Це є певним джерелом небезпеки, а саме можлива компрометація сайту – його високе рангування по небажаних словах (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 має місце особливе визначення рангів посилань зі світових каталогів). Саме існування каталогів якісних ресурсів є запорукою стійкості при розв'язанні складних обчислювальних задач по визначенню PR.

З метою повнішого врахування системної функції каталогів ресурсів розроблено алгоритм HITS, який розглядається як альтернативний до PageRank. Алгоритм також базується на дослідженні структури WWW на основі посилань. Виділяються ряд додаткових характеристик 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. 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 на сьогодні алгоритм HITS більше застосовується у тематичному пошуку по добре модерованих базах 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 елементу зручності інтерфейсу користувача поступово перетворюються у головний інструмент організації усього інформаційного ресурсу WWW. Зокрема, адміністрації сайтів вводять строгі 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;
- 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 :

1) на основі мета-інформації сторінки (мета-тег “Revisit-after”);

2) на основі прогнозованої частоти оновлення сторінки на основі попередніх даних.

Обидва варіанти практично не передбачають захисту від некоректних неправдивих показників, які характерні для пошукового спаму. Крім того, використання другого варіанту може призвести до “зациклення” робота на

сторінка з динамічно обновлюваною 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. Таким чином для важливих та популярних сторінок (наприклад авторитетних інформаційних агенств) зберігається часта переіндексація, а пошуковий спам та несуттєві динамічні сторінки переіндексовуються в загальному порядку.

Виявлення плагіату в WWW пошуковою системою здійснюється по принципу порівняння “авторитетності” сторінок з однаковими або схожими текстами між собою 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 відмітити, що ця ситуація зустрічається в чистому виді 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).

2.3. Пошукові системи з аналізом популярності інформаційних ресурсів

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 по структурі WWW, вони використовують доступну їм інформацію про реальну популярність ресурсу. Для ресурсів, про які інформація є недоступною,

здійснюється екстраполяція на основі гіпертекстових посилань. Таким чином, 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, повинне бути встановлений “лічильник” даної системи).
- До апаратного забезпечення системи ставляться високі вимоги – адже вона в реальному часі повинна обліковувати відвідувачів сайтів.

Для усунення наведених обмежень на 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 по он-лайн рекламі або “просто так”, без певної мети.

2.4. Системи на основі експертного та колаборативного оцінювання ресурсів

Четверту групу пошукових систем складають гібридні системи, де результати пошуку формуються з урахуванням оцінки ресурсів користувачами. Зокрема, може мати місце:

- оцінка користувачами якості пошуку;
- оцінка користувачами ресурсів (зокрема – голосування);
- відбраковування користувачами незадовільних ресурсів;

- безпосереднє формування бази даних ресурсів.

У попередні роки подібні системи не мали високої популярності, хоча елементи були присутні в усіх пошукових системах (зокрема відзначення користувачами спаму та експертне оцінювання ресурсів). Проте за останній час з появою Веб 2.0 та ростом активності користувачів такі системи почали набувати суттєвого значення для WWW.

Системи даного класу часто є ближчими по своїй суті до каталогів сайтів, ніж до пошукових систем. Проте, на відміну від каталогів, у навігаційних сервісах Веб 2.0 значно демократичніші правила поповнення бази даних та інші принципи рейтингування. Як наслідок, за обсягами БД та якістю наповнення вони наближаються до пошукових систем.

До таких систем відносяться:

- служби онлайн-закладок (наприклад Del.Isio.Us);
- колективні агрегатори новин (наприклад blogspot.com);

колективні реєстри ресурсів (наприклад digg.com).