УДК 004.1'3

Т.О. САВЧУК, П.В. НОВАЛЕНКО
Вінницький національний технічний університет, Україна

ВИКОРИСТАННЯ МОДИФІКОВАНИХ МЕТОДІВ ПОШУКУ ТЕМАТИЧНИХ СПІВТОВАРИСТВ ДЛЯ ПІДВИЩЕННЯ ПАРТИНЕНТНОСТІ ПОШУКУ У WEB-РЕСУРСАХ

В статті запропоновано використання модифікованих методів пошуку тематичних співтовариств з метою ефективного пошуку тематичної інформації у web-ресурсах. За рахунок запропонованих модифікацій означених методів з урахуванням аналізу текстів та семантичної структури web-сторінки підвищується точність та партинентність пошуку на 5%.

Ключові слова: тематичний пошук, алгоритм HITS, web-ресурси, методи пошуку тематичних співтовариств.

T.O. SAVCHUK, P.V. NOVALENKO
Vinnytsia National Technical University, Ukraine

THE USAGE OF MODIFIED METHODS OF INFERRING WEB COMMUNITIES FOR THE INCREASE OF SEARCH PARTINENT AT WEB-RESOURCES

In this article the modified methods usage of inferring web communities in aim of effective search of topical information at web-resources is suggested. Due to offered modifications of the considered methods with taking into account the analysis of the text and semantic structure of web-page, the precision and pertinence of search to 5% is arised.

Keywords: thematic search, algorithm HITS, web-resources, methods of inferring web-communities.

Вступ

Задача тематичного інформаційного пошуку у загальному виді полягає в тому, щоб у заданому просторі пошуку знайти документи, релевантні інформаційній потребі користувача, заданої у вигляді запиту. Тому, у зв'язку з наявністю потужних об'ємів інформації доступних в мережі Інтернет, вирішення цієї задачі є пріоритетним для забезпечення своєчасного доступу до інформації, що цікавить користувача.

Постановка задачі

Нехай є простір пошуку:

$G = \langle O, L \rangle$ – орієнтований граф об'єктів,

де O – множина об'єктів пошуку, а L – множина орієнтованих посилань між об'єктами.

Кожний об'єкт $o_i \in O$ має вектор атрибутів:

$$O = \{o_j = \langle a_{j1},...,a_{jk} \rangle, k = k(j)\}, \quad (1)$$

Для різних об'єктів набори атрибутів можуть відрізнятися.

Простір пошуку в кожний момент часу являє собою лише деяке наближення всього графа Web. Це наближення може змінюватися в процесі пошуку:

$$G_t = \langle O_t, L_t \rangle = G(t), \qquad t = 0,1,2... \quad (2)$$

Користувач формулює своє початкове представлення про інформаційну потребу у вигляді початкового запиту:

$$b_0 = \{b_{01},...,b_{0m}\}, \quad (3)$$

який задає множину необхідних атрибутів об'єктів пошуку. Запит користувача може змінюватись в процесі пошуку $b_t = b(t)$.

Функція пошуку A здійснює інформаційний пошук у просторі G_t по запиту b_t:

$$A(b_t, G_t) = \delta \subset O_t \quad (4)$$

Результатом пошуку є множина δ об'єктів пошуку із простору пошуку O_t.

Користувач оцінює результати пошуку на відповідність своєї інформаційної потреби:

$$E(b_t, \delta) = \delta' \subseteq \delta \quad (5)$$

Таким чином, для розв'язку задачі тематичного ІП в Інтернет необхідно побудувати функцію пошуку $A(b_t, G_t)$;

Пошук припиняється, коли множини δ'_{t+1} і δ_1, будуть співпадати:

$$\left\| \delta'_{t+1} - \delta'_t \right\| < \varepsilon \quad (6)$$

Точність пошуку визначається відповідно до традиційного визначення точності: $\frac{\left|\delta'_{t+1}\right|}{\left|\delta'_t\right|}$.

Аналіз сучасних засобів та алгоритмів тематичного пошуку у web-ресурсах

Існуюча практика інформаційного пошуку в Інтернет склалась таким чином, що й для традиційного, і для тематичного пошуку більшість користувачів звертаються до тих самих промислових інформаційно-пошукових систем (ІПС). У той же час для розв'язку задачі тематичного пошуку в Інтернет були розроблені математичні методи, які практично невідомі рядовим користувачам Інтернет. Враховуючи це, подальший виклад матеріалу даного розділу побудовано в такий спосіб: описані існуючі промислові ІПС в Інтернет, методи розв'язку задачі тематичного пошуку в Інтернет та сформульовані висновки по застосуванню різних промислових ІПС і існуючих методів для розв'язку задачі тематичного ІП в Інтернет [1].

Промислові ІПС в Інтернет можна розділити на три основні класи, які показані у таблиці 1.

Таблиця 1

Класифікація інформаційно-пошукових систем в Інтернет

№	Клас	Приклади
1	Системи пошуку по ключових словах	Google, Yandex
2	Класифікатори	ODP, Yahoo
3	Метапошукові системи	Dogpile, Yippy, META.UA

1) На сьогодні СПКС являються найпоширенішими ІПС в Інтернет. СПКС часто інтегруються із класифікаторами (так зроблено, наприклад, у СПКС Yandex), надаючи можливість пошуку по ключових словах у класифікаційній системі й у сторінках, що містяться в класифікаторі [2].

Виробники більшості існуючих на сьогоднішній день промислових СПКС тримають інформацію про архітектуру і методи роботи систем повністю закритою з комерційних міркувань. Основною проблемою при розробці СПКС є забезпечення роботи з надвеликим об'ємом даних (близько 1 трлн. сторінок) і одночасна обробка запитів від великої кількості користувачів (більше 500 млн. запитів щодня), тому СПКС являють собою складні розподілені програмні комплекси.

Одним з найбільш кардинальних кроків у розвитку СПКС стало використання аналізу гіперпосилань для підвищення якості ранжування сторінок. Цей підхід був вперше реалізований у вигляді алгоритму Pagerank [3], який обчислює ранг сторінки - Pagerank - з урахуванням рангів сторінок, що посилаються на неї, на основі наступного рекурсивного визначення:

$$PR(A) = p / N + (1 - p) * (PR(T1) / C(T1) + ... + PR(Tn) / C(Tn)), \quad (7)$$

де $PR(A)$ - ранг Pagerank сторінки A;

p - константа, зазвичай 0.1-0.2;

$T1$ - ранг Pagerank сторінки $T1$, що вказує на сторінку A;

$C(T1)$ - число вихідних посилань зі сторінки $T1$;

n - число сторінок, що вказують на A;

N - число сторінок у базі даних.

Ця формула відображає процес нескінченного випадкового блукання по Web. Користувач переміщається по Web шляхом навігації по гіперпосиланнях. У якийсь момент користувач може "стомитися" і перейти на деяку сторінку, задавши її адресу в браузері. У кожний момент часу користувач перебуває на деякій сторінці й може з імовірністю p перейти на випадкову сторінку або з імовірністю *1-p* прослідувати по гіперпосиланнях з даної сторінки. Імовірність того, що він прослідує по одній з посилань із сторінки A, дорівнює $(1 - p)/C(A)$.

Для обчислення рангів Pagerank потрібен аналіз усієї бази даних СПКС. Кожній сторінці спочатку присвоюється одиничний ранг, після чого здійснюється кілька ітерацій обчислення, у ході кожної з яких ранги всіх сторінок обновляються по наведеній вище формулі. Ранги періодично переобчислюються в міру оновлення бази даних СПКС (приблизно один раз на місяць для Google). Ранг Pagerank відбиває популярність сторінки, і його легко підвищити навмисне - достатньо створити велику множину сторінок, що посилаються на задану сторінку.

2) Класифікатори надають користувачам можливість самостійно шукати сторінки, що зберігаються в базі даних класифікатора й розподілені по розділах деревоподібної класифікаційної системи, аналогічно роботі із предметним покажчиком бібліотеки.

База даних класифікатора заповнюється або вручну, або класифікатор використовує базу даних СПКС. Як ми вже відзначали, у багатьох випадках класифікатор входить до складу СПКС і остання дозволяє робити пошук по ключових словах, по класифікаційній системі й по вмісту бази даних [4].

Індексування в більшості класифікаторів здійснюється вручну користувачами, наприклад, в Yahoo, ODP і інших.

3) Метапошукові системи дозволяють здійснювати ІП в Інтернет по ключових словах і використовують у своїй роботі існуючі СПКС і бази даних. Метапошукові системи не містять власного індексу, а при обробці користувацького запиту звертаються до декількох СПКС і базам даних [5]. При цьому метапошукова система виконує наступні дії:

- відповідає за вибір використовуваних СПКС;
- може модифікувати запит;
- ранжує, сортує, фільтрує й групує результати за своїм розсудом.

Такий підхід має наступні переваги:

- він дозволяє підвищити охоплення Інтернет при пошуку. Враховуючи, що жодна з існуючих СПКС не охоплює більш 56% Інтернет, у той час як використання 11 СПКС дозволяє охопити 92% Інтернет;
- користувачеві не потрібно знати відмінності мов запитів, використовуваних СПКС;
- з'являється додаткова можливість підвищити точність пошуку за рахунок аналізу інформації про одну й ту ж сторінку з декількох джерел. Ранжування сторінок може здійснюватися на основі алгоритмів, відмінних від використовуваних конкретними СПКС.

За допомогою набору ключових слів часто неможливо виразити інформаційну потребу, тому розвиток метапошукових систем спрямований на більш точне з'ясування інформаційної потреби. Незважаючи на слабку підтримку тематичного пошуку у промислових ІПС, існує велика кількість дослідницьких розробок, спрямованих на вирішення задачі тематичного ІП в Інтернет. Ці методи можна розділити на наступні напрямки:

- традиційні методи пошуку тематично близьких документів в електронних бібліотеках, що ґрунтуються на аналізі тексту документів;
- методи, специфічні для Web: тематичні роботи, методи пошуку тематичних співтовариств в Web.

Традиційні методи орієнтовані на пошук в електронних бібліотеках і не використовують інформації про гіпертекстову структуру Web. Стосовно Web такі методи використовуються, зокрема, у промислових ІПС. Методи, специфічні для Web, розширюють традиційні методи залученням додаткової інформації про структуру Web з метою підвищення якості пошуку.

Усі методи тематичного ІП, специфічні для Web, ґрунтуються на наявності гіперпосилань між web-сторінками й на наступних припущеннях про семантику гіперпосилань [3]:

- коли автор сторінки А ставить на ній гіперпосилання на іншу сторінку В, він рекомендує читачеві прочитати не тільки А, але ще й В;
- якщо дві сторінки з'єднані гіперпосиланням, то ймовірність того, що вони відносяться до однієї теми вище, чим у випадку відсутності гіперпосилання.

Таким чином, гіперпосилання є відображенням думок авторів сторінок і можуть бути основою для оцінки тематичної близькості документів, тобто певної семантичної інформації. Важливо відзначити, що спільність цих припущень обмежена: посилання між документами використовуються й в інших цілях, прикладом тому є навігаційні посилання й реклама.

Нижче розглянуті методи тематичного ІП, специфічні для Web: тематичні роботи й методи пошуку тематичних співтовариств.

Робота тематичного робота багато в чому схожа на роботу звичайного мережного робота, використовуваного в СПКС. Відмінність полягає лише в тому, що робот СПКС обходить усі сторінки, досяжні по гіперпосиланнях, а задачею тематичного робота є обхід тільки сторінок, що належать до заданої теми. Мірою ефективності тематичного робота є відношення кількості скачаних сторінок, що належать до теми, до загальної кількості скачаних сторінок. Використання спрямованого обходу досить актуально, тому що при ненаправленому рекурсивному обході сторінок тематична спрямованість губиться дуже швидко. Визначено, що при обході всіх посилань зі сторінки на глибину 2, темі вихідного документа відповідає тільки 10% документів [4].

Тематичний робот складається з наступних компонентів:

- мережний робот, який відповідає за скачування сторінок;
- база даних URL, яка містить у собі список відвіданих сторінок і черга URL ще не відвіданих сторінок;
- аналізатор документів, який направляє мережного робота шляхом упорядкування черги сторінок;
- підсистема класифікації, яка визначає відповідність документів темі.

Робота тематичного робота будується в такий спосіб. Для початку роботи робот повинен мати визначення теми, по якій він повинен, шукати документи. Це визначення задається у вигляді набору сторінок, які належать до цієї теми (так званої початкової множини сторінок). Початкова множина сторінок може бути задана користувачем вручну, при цьому користувач повинен самостійно знайти кілька сторінок в Інтернет, які належать до теми, що його цікавить [4].

Експерименти показують, що при використанні тематичного робота співвідношення числа релевантних сторінок до числа відвіданих варіюється в межах 0.4-0.5, тоді як для нетематичного робота, що працює з тими ж початковими даними, це співвідношення не перевищує 0.1.

Використання класифікатора замість простої вказівки користувачем початкової множини сторінок має наступні переваги:

- дозволяє підвищити точність класифікації, тому що відомий контекст пошуку;
- дозволяє розв'язати проблему пошуку вхідних даних для робота, тому що відобразити тему, що цікавить користувача, на класифікатор іноді простіше, чим знайти документи-приклади;
- дозволяє знаходити теми, тісно пов'язані з темою, що цікавить користувача, і, отже, більш повно представити тему, що цікавить користувача.

У той же час слід зазначити, що розмір ODP становить близько 590,000 категорій, що ускладнює

роботу користувача по формуванню початкової множини документів. Крім того, даний підхід використовує як зразки тільки Web-Сторінки, а при наявності класифікатора можна було б використовувати ще й звичайні документи, що не містять гіперпосилань.

У роботі Брайана Девісона була експериментально обґрунтована властивість тематичної локальності Web, згідно з якою більшість сторінок з'єднана посиланнями із близькими по темі сторінками. На підставі цього факту можна зробити припущення про існування в Web груп зв'язаних між собою сторінок, що належать до однієї теми - так званих тематичних співтовариств. Дослідження показали вірність цього припущення, у результаті чого з'явився новий підхід до тематичного ІП - пошук тематичних співтовариств [6].

Поняття тематичного співтовариства, як і поняття релевантності, неможливо визначити формально, тому в різних роботах можна знайти багато різних визначень тематичного співтовариства. Більшість із них базується на визначенні тематичного співтовариства за допомогою структурного шаблону (сигнатури співтовариства) - підграфа графа Web певного виду. Так, наприклад, сигнатурою співтовариства може бути повний двочастковий підграф. При цьому: шаблон дозволяє ідентифікувати лише частину сторінок співтовариства, так зване ядро співтовариства, а границя співтовариства зазвичай не визначається.

Пошук тематичних співтовариств по заданій темі дозволяє знайти групи сторінок, у яких сторінки багато посилаються один на одного. На основі властивості тематичної локальності можна припустити, що сторінки усередині групи належать до однієї теми. Таким чином, пошук тематичних співтовариств можна розглядати як метод тематичного інформаційного пошуку в Інтернет [7].

Нижче розглянуті основні методи пошуку тематичних співтовариств.

Алгоритм HITS, розроблений Дж. Клейнбергом, багато в чому схожий на Pagerank, але на відміну від останнього він аналізує не граф Web, а невеликий його підграф. У ході своєї роботи HITS взаємодіє із СПКС для побудови аналізованого підграфа.

Клейнберг відходить від моделі популярності сторінки, запропонованої в Pagerank, і вводить поняття значимості сторінки. Найбільш значимими сторінками запропоновано вважати ті сторінки, на які найбільше посилаються інші значимі сторінки. Такі сторінки називаються авторитетними (authorities). Авторитетні сторінки є найбільш значимими в рамках заданої теми, тому на них часто посилаються інші сторінки, що належать до даної теми. Ця властивість дозволяє виявити так звані індексні сторінки (hub pages), які посилаються на кілька авторитетних сторінок, що належать до однієї теми. Разом ці два типи значимих сторінок утворюють відношення взаємного посилення (mutually reinforcing relationship), тобто якісна авторитетна сторінка посилається на багато якісних індексних сторінок і якісна індексна сторінка посилається на багато якісних авторитетних сторінок. Таким чином, метою аналізу в HITS є пошук найбільш якісних авторитетних сторінок і найбільш якісних індексних сторінок [8].

Робота алгоритму будується в такий спосіб:

- на першому етапі будується так званий сфокусований підграф Web, який містить сторінки, отримані шляхом посилання запиту до СПКС;
- на другому етапі проводиться аналіз сфокусованого підграфа Web і обчислюються найбільш значимі документи.

Алгоритм HITS локальний при аналізі (тобто використовує підграф Web), але для побудови підграфа вимагає глобального індексу СПКС, щоб одержувати гіперпосилання на сторінку.

Алгоритм SALSA аналізу гіперпосилань базується на ідеях HITS, але використовує для фази аналізу стохастичний метод, заснований на ланцюгах Маркова. Алгоритм використовує як вхідні дані множина Sσ з HITS і аналогічно HITS обчислює АР- і НР-ваги сторінок. Аналогічно алгоритму Pagerank алгоритм SALSA відображає процес випадкового блукання по підграфу графа Web, індукованому на множині Sσ, але з наступним обмеженням: перехід від однієї сторінки до іншої здійснюється шляхом двох переміщень по гіперпосиланнях у різних напрямках: по одній з посилань уперед і потім з отриманої сторінки по одній з посилань назад [9].

При використанні евклідової норми SALSA у точності відповідає першому кроку алгоритму HITS у випадку, якщо в графі міститься рівно одне співтовариство. При наявності декількох співтовариств HITS виділяє найбільш головне з них, а SALSA виділяє найбільш значимі сторінки з різних співтовариств. У результаті, SALSA успішно працює в ряді випадків, коли HITS стає непридатним, наприклад, коли є маленька група сильно зв'язаних сторінок.

Аналіз предметної області показав, що для розв'язання задачі тематичного пошуку у web-ресурсах найкраще підходять методи пошуку тематичних співтовариств і тематичні роботи. Але враховуючи велике поширення реклами в Інтернеті, банерних мереж, лічильників та автоматично створюваних посилань виникає необхідність в удосконаленні даних методів для підвищення якості їх роботи.

Удосконалення методів пошуку тематичних співтовариств для підвищення партинентності пошуку у web-ресурсах

Напрямки покращення тематичного пошуку у web-ресурсах при використанні методів пошуку тематичних співтовариств можна розділити на дві категорії:

додавання нових евристик в алгоритм;

комбінування з аналізом тексту.

Необхідно враховувати, що одним автором може бути створено кілька сайтів, які містять багато посилань один на одного. У цьому випадку authority-вага сторінки, на яку посилаються багато (n) сторінок з деякого сайту, збільшується в n раз. Для таких сторінок вплив кожного посилання слід зменшувати в n раз і аналогічним образом робити з hub-вагою. Також необхідно використовувати ваги гіперпосилань і не видаляти внутрішні гіперпосилання, а присвоювати їм малі ваги. Використання ваг дозволяє комбінувати HITS з аналізом тексту.

Вага посилання обчислюється шляхом аналізу фрагмента тексту навколо посилання (100 байт навколо посилання):

$$w = 1 + n(t), \tag{8}$$

де n(t) - кількість входжень термів із запиту в розглянутий фрагмент.

Для перешкоджання потраплянню в підграф G не релевантних, рекламних та навігаційних посилань пропонується аналізувати структуру окремих сторінок. Великі індексні сторінки необхідно ділити на частини на підставі розмітки HTML і аналізувати як окремі сторінки. Розвинувши даний підхід, потрібно будувати все дерево об'єктної моделі web-сторінки і в ньому вибирати потрібний варіант розбивки таким чином, щоб кожна окрема частина документа або відповідала темі, або ні. Такий спосіб дозволяє видаляти семантично незначні посилання, розбивати не тільки індексні сторінки, але й авторитетні й, у підсумку, уникнути "зсуву теми".

Також можуть бути використані промислові інформаційно-пошукові системи як джерела інформації про граф Web, для використання в методах пошуку тематичних співтовариств.

Висновки

Таким чином, за рахунок запропонованих модифікацій знижено кількість попадань в аналізований граф гіперпосилань, що не задовольняють припущеннями про семантику, а також дещо зменшено розмір аналізованого графу. Проведені експерименти показали, що поєднання аналізу тексту з аналізом структури гіперпосилань підвищує точність та партинентність пошуку на 5%, але ціною збільшення складності, що доводить доцільність використання модифікованого методу пошуку тематичних співтовариств для розв’язання поставленої задачі.

Література

1. Маннинг К. Д. Введение в информационный поиск / К. Д. Маннинг, П. Рагхаван, Х. Шютце. М.: Вильяме, 2011. - 528 с.

2. Davison B. Recognizing Nepotistic Links on the Web / B. Davison // AAAI Workshop on Artificial Intelligence for Web Search. Technical Report WS-00-01, AAAI Press. - 2000. - P. 23-28.

3. Райдингс К. Растолкованный PageRank. http://digits.ru/articles/promotion/pagerank.html. (пер. Садовский А.)

4. Chakrabarti S., Van Den Berg M., Dom B. Focused crawling: A new approach to topic-specific Web resource discovery. Eights World Wide Web Conference, Toronto, May 1999.

5. Сегалович, И.В. Как работают поисковые системы // Мир Internet.v -2002.-№ 10.-С. 24-32.

6. Flake G. Self-Organization of the Web and Identification of Communities / G. Flake, S. Lawrence, C. Giles, F. Coetzee // IEEE Computer. 2002. - Volume 35, № 3. - P. 66-71.

7. Gibson D. Inferring Web communities from link topology / D. Gibson, J. Kleinberg, P. Raghavan // Proc. 9th ACM Conference on Hypertext and Hypermedia. ACM Press. - New York. - 1998. - NY, USA. - P. 225-234.

8. Хензингер М. Анализ гиперссылок в Web. Открытые системы, 2001, N10. http://www.osp.ru/2001/10/050.htm.

9. Kleinberg J. Authoritative sources in a hyperlinked environment. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.

References

1. Manning C. D. Introduction to Information Retrieval / C. D. Manning, P. Raghavan , H. Schütze.: Cambridge University Press, 2009. - 496c.

2. Davison, B. Recognizing Nepotistic Links on the Web / B. Davison // AAAI Workshop on Artificial Intelligence for Web Search. Technical Report WS-00-01, AAAI Press. - 2000. - P. 23-28.

3. Ridings C. PageRank Explained. http://digits.ru/articles/promotion/pagerank.html. (tr. Sadovsky A.)

4. Chakrabarti S., Van Den Berg M., Dom B. Focused crawling: A new approach to topic-specific Web resource discovery. Eights World Wide Web Conference, Toronto, May 1999.

5. Segalovich I. How Search Engines Work // Mir Internet. -2002.-№ 10.-C. 24-32.

6. Flake G. Self-Organization of the Web and Identification of Communities / G. Flake, S. Lawrence, C. Giles, F. Coetzee // IEEE Computer. 2002. - Volume 35, № 3. - P. 66-71.

7. Gibson D. Inferring Web communities from link topology / D. Gibson, J. Kleinberg, P. Raghavan // Proc. 9th ACM Conference on Hypertext and Hypermedia. ACM Press. - New York. - 1998. - NY, USA. - P. 225-234.

8. Henzinger M. Hyperlink Analysis for the Web. Internet Computing, 2001, N10. http://www.osp.ru/2001/10/050.htm.

9. Kleinberg J. Authoritative sources in a hyperlinked environment. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.

Рецензія/Peer review : 6.11.2014 р. Надрукована/Printed :2.1.2015 р.