УДК 004.89+336.713

# ПІДВИЩЕННЯ ПАРТИНЕНТНОСТІ ПОШУКУ У WEB-РЕСУРСАХ З ВИКОРИСТАННЯМ МЕТОДІВ ПОШУКУ ТЕМАТИЧНИХ СПІВТОВАРИСТВ

*Савчук Тамара, Новаленко Павло*

Вінницький національний технічний університет

## Анотація

*З метою ефективного пошуку тематичної інформації у web-ресурсах доцільним є використання методів пошуку тематичних співтовариств, які засновані на аналізі структури гіперпосилань. В даній роботі запропоновано модифікацію означених методів з урахуванням аналізу текстів та семантичної структури web-сторінки.*

## Abstract

*With the aim of effective thematic information retrieval in web- resources expedient are the uses of methods of search of thematic communities, that based upon the analysis of the structure of hyperlinks. In this work modification of the marked methods is offered taking into account the analysis of texts and semantic structure of web- page.*

## Вступ

Актуальністю тематичного інформаційного пошуку визначається наявність потужних об’ємів інформації доступних в мережі Інтернет в сукупності з великими темпами її росту. Це в свою чергу становить вирішення задач тематичного інформаційного пошуку не тільки пріоритетним, а й життєво необхідним для забезпечення своєчасного доступу до інформації, що цікавить користувача.

## Застосування методів пошуку тематичних співтовариств для тематичного пошуку в web-ресурсах

Для здійснення тематичного інформаційного пошуку в Інтернет було запропоновано застосовувати так звані методи пошуку тематичних співтовариств.

Основною особливістю цих методів є те, що вони засновані на наступному припущенні про семантику посилань між сторінками:

– Коли автор ставить на своїй сторінці А посилання на чужу сторінку В, він рекомендує читачеві А прочитати ще й В.

– Якщо дві сторінки з'єднані посиланням, то ймовірність того, що вони відносяться до однієї теми вище, ніж у випадку відсутності посилання.

Таке припущення було досить правдоподібне в той час, коли в Інтернеті не було реклами, зараз же поширення банерних мереж, лічильників та автоматично створюваних посилань істотно знижує кількість сторінок, відповідних цьому припущенню і знижує якість роботи цих методів.

Напрямки покращення тематичного пошуку у web-ресурсах при використанні методів пошуку тематичних співтовариств можна розділити на дві категорії: додавання нових евристик в алгоритм і комбінування їх з аналізом тексту.

Модифікована модель роботи цих методів виглядає наступним чином:

1. Користувач задає тему пошуку у вигляді набору ключових слів. Цей набір надсилається системі пошуку за ключовими словами, яка повертає множину сторінок *R*. На основі *R* будується множина *S*, яке розширюється шляхом додавання всіх сторінок, які посилаються на сторінки з *R* і сторінок, доступних з *R* по посиланнях.

Також пропонується аналізувати структуру окремих сторінок. Великі 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.

2. Будується підграф *G* графа Web на основі множини *S*.

3. Граф *G* аналізується з метою виділення в ньому сторінок двох типів: так званих авторитетних (authorities) і індексних сторінок (hub pages). Авторитетність сторінки розуміється за аналогією з бібліографічним цитуванням: чим більше сторінка цитується авторитетними сторінками, тим більш вона авторитетна. Індексні сторінки містять посилання на авторитетні сторінки.

Потрібно 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):

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

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

Такий спосіб перешкоджає потраплянню в підграф *G* не релевантних, рекламних та навігаційних посилань, дозволяє видаляти семантично незначні посилання, розбивати не тільки індексні сторінки, але й авторитетні й, у підсумку, уникнути "зсуву теми".

## Висновки

Експериментально доведено, що при вирішенні задачі тематичного пошуку в web-ресурсах, пропоновані методи пошуку тематичних співтовариств, засновані на зазначених вище припущеннях про семантику гіперпосилань стають незастосовні без використання додаткових евристик, які відсіюють гіперпосилання, що завідомо не відповідають припущенням про семантику. Модифікований підхід дозволяє знизити кількість попадань в аналізований граф гіперпосилань, що не задовольняють припущеннями про семантику, а також дозволяє дещо знизити 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.

## Список використаних джерел:

1. Borodin A., Roberts G., Rosenthal J., Tsaparas P. Finding Authorities and Hubs From Link Structures on the World Wide Web. Tenth World Wide Web Conference, Hong Kong, 2001.

2. 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.