XII наукова К О Н Ф Е Р Е Н Ц І Я ТЕРНОПІЛЬСЬКого ДЕРЖАВНого ТЕХНІЧНого УНІВЕРСИТЕТу ІМЕНІ ІВАНА ПУЛЮЯ 14– 15 травня 2008 р. Тези конференції

УДК 004.415.5

*__О.Назаревич__, **__О.Чернух__, *__В.Яцишин__
(*Тернопільський державний технічний університет імені Івана Пулюя)
(**ПВНЗ "Європейський університет", Тернопільська філія)

ПОРІВНЯЛЬНИЙ АНАЛІЗ МЕТОДІВ ВИЗНАЧЕННЯ НЕЧІТКИХ ДУБЛІКАТІВ ДЛЯ WEB-ДОКУМЕНТІВ

Актуальність задачі визначення нечітких дублікатів WEB-документів визначається різноманітністю сфер застосування, в яких необхідно враховувати "подібність", зокрема: поліпшення якості індексу і архівів пошукових систем; об'єднання повідомлень новин в сюжети на основі подібності цих повідомлень за змістом; фільтрація спаму; проблема плагіату або копірайта; і ряд інших.

Одними з перших досліджень в даній області є роботи *U. Manber* і *N. Heintze.* В якості міри подібності двох документів використовується відношення числа загальних підстрічок до розміру файла або документа.

В 1997 р. *A. Broder et al*. запропонували новий, "синтаксичний" метод оцінки подібності двох документів, заснований на представлені документу в вигляді множини всеможливих послідовностей фіксованої довжини *k*, які складається із сусідніх слів. Такі послідовності були названі "шинглами". Два документа рахувалися подібними, якщо їх множини шинглів істотно перетиналися. Подальшим развитком концепцій *A. Broder et al.* були дослідження *D. Fetterly et al.*, *A. Broder et al.* Суттєві переваги алгоритму *D. Fetterly et al.* в порівнянні з дослідженнями *A. Broder et al*. полягала в тому, що, по-перше, будь-який документ (в том числі і дуже малий) завжди можна представити вектором фіксованної довжини, і, по-друге, подібність визначається простим порівнянням координат вектора і не потребує (як в *A. Broder*) виконання теоретико-множинних операцій.

Інший сигнатурний підхід, заснований вже не на синтаксичних, а на лексичних принципах, був запропанований *A. Chowdhury et al.* в 2002 р. і покращений в 2004 р. Основна ідея поягає в обчисленні дактілограми *I-Match* для представлення змісту документів. Два документа рахуються подібними, якщо в них співпадають *I-Match* сигнатури. Алгоритм має високу обчислювальну ефективність та високу ефективність в порівняні невеликих документів. В редакції 2004 р. покращено стійкість роботи алгоритму в ситуаціях коли відбуваються невеликі зміни змісту документа, за рахунок можливості багаторазового випадкового перемішування основного словника. Інший схожий підхід описаний в патенті *US Patent 6,658,423 W. Pugh* з компанії *Google*.

Ще одним сигнатурним підходом, також на основі лексичних принципів, є метод "опорних" слів, запропонований *С. Ильинским* та ін. Даний метод, дозволяє, почавши з виборкі в сотні тисяч слів, залишити набір в 3-5 тисяч, розрахунок сигнатур по якому із застосуванням повнотекстового індексу здійснюється на міліардному індексі декілька хвилин на пошуковому кластері *Яндекса.*

В 2007 р. Зеленков Ю.Г, Сегалович И.В (Yandex team) в своїй публікації запропонували серію алгоритмів серед яких два: *метод "декомпозиції" (A11) та метод "3+5" (A12)* показали найкращі експериментальні результати. В якості основних критеріїв роботи даних алгоритмів були вибрані: *повнота, точність та F-міра.*