УДК 681.301

Наук. співробітник Михайлюк О.С., магістрант Нетецький К. І.

**Національний технічний університет України
«Київський політехнічний інститут»**

ДОСЛІДЖЕННЯ АЛГОРИТМІВ ВИЯВЛЕННЯ ДУБЛЮВАНЬ У ТЕКСТОВИХ ІНФОРМАЦІЙНИХ ОБ'ЄКТАХ

Вступ

Головну цінність Internet становить інформація, що знаходиться на його необмеженому просторі. Причому справді цінною є лише унікальна інформація, не розтиражована дослівно на десятках сайтів. Сьогодні ж повідомлення багаторазово дублюються на численних нових веб-сторінках, у той час як веб-сторінки з новою і дійсно корисною інформацією з'являються не надто часто. Додаткові проблеми виникають внаслідок можливого порушення змістовної цілісності інформації при її безпосередньому або опосередкованому тиражуванні, оскільки безліч сайтів містить просто уривки оригінальних текстів, окремі абзаци тощо, що часто призводить до неправильного тлумачення відомостей, описаних у першоджерелі.

Постановка задачі

Одним із завдань, з яким стикається будь-яка інформаційно-пошукова система, є визначення взаємної схожості різних документів. Виявлення дублікатів дозволяє видаляти повтори текстів у пошуковому відгуку, зменшувати розміри індексу шляхом усунення надлишковості, виявляти плагіат і розпізнавати спам. У спрощеному вигляді задача зводиться до з'ясування того, чи є два довільні документи дублікатами. Нижче розглянуто деякі з існуючих алгоритмів виявлення дублікатів текстів.

Методи встановлення дублювань електронних текстових документів

Одним з методів виявлення дублікатів є метод "описових слів" [1]. При його реалізації набирається деяка множина слів N, що називається "описовою множиною". Ця множина слів повинна покривати максимально

можливу кількість документів, при цьому кількість слів у самій множині повинна бути мінімальною.

Для кожного слова встановлюється гранична частота *Vi* з якою воно зустрічається у документах. Це значення повинно задовольняти наступну умову: різниця між граничною частотою і частотою, з якою слово зустрічається в конкретному документі, не повинна бути занадто малою.

У свою чергу, слова повинні бути підібранні таким чином, щоб відносна стабільність відповідного компонента вектора була максимальна, тобто ймовірність для цього слова перевищити встановлений поріг (на збільшення або зменшення), мінімальна - у випадку невеликих змін документа. Оптимальна кількість слів *N* визначається експериментально.

Після визначення *Vi* для кожного слова по кожному документу підраховується вектор, де *i*-тий компонент вектора дорівнює одиниці, якщо величина відносної частоти *i*-того слова з "описової множини" цього документа більша, ніж вибрана гранична частота, інакше компонент дорівнює нулю. Після цього, отримані вектори порівнюються за допомогою функції Левенштейна [2], і якщо різниця між ними не перевищує 8% то документи вважаються однаковими. Вірогідність отриманої інформації про наявність дублікатів при використанні цього методу залежить від вдалого вибору "описових слів" і від оцінок граничних частот, тобто значною мірою залежить від «людського фактора».

Другий метод, що розглядається, - метод сигнатур. У роботі [3] автором пропонується метод отримання сигнатур документа, які будуються на основі певного набору статистичних параметрів документа, обраних з міркувань стійкості до певних форм зміни документа. Наприклад, приблизну кількість речень у документі можна визначити за кількістю великих літер, ком і крапок; загальна довжина тексту в символах за винятком пробілів і стоп-слів дає загальну оцінку обсягу документа тощо.

Автором методу проводилися дослідження можливості використання різноманітних параметрів текстів для опосередкованого виявлення дублювань [3]. Так, було підраховано кількість входжень у текст документа кожного символу з наступного набору: . , - _ : ; ! ? ( ) і символ пробілу. Образ документа подається у вигляді вектора розмірністю 11 елементів, *i*-тим компонентами якого була кількість входжень відповідного символу. В іншому тесті з документа видалялися буквено-цифрові символи, залишаючи спецсимволи, пробіли й переведення рядків. Таким чином була превірена послідовність спецсимволів. Так само в документах були підраховані середня довжина слова та речення, а також їх загальна кількість. Шляхом конкатенації отриманих значень складалася

сигнатура вигляду: [середня довжина слова] - [середня довжина речення] - [кількість слів] - [загальне число речень]. Також в одному з тестів були виділені з тексту документа й зчеплені два найдовших речення. Факт наявності дублювання встановлювався шляхом застосування функції Левенштейна.

Дані, отримані в тестах при використанні кількості та послідовності спецсимволів у документах були однаковими і показали найкращий результат, що було підтверджено визначенням середнього гармонійного повноти і точності. Під повнотою автор мав на увазі відношення загальної кількості знайдених “релевантних” пар дублікатів до загальної кількості “релевантних” дублікатів, а під точністю - відношення загальної кількості знайдених дублікатів до загальної кількості знайдених пар. Тести, що використовують параметри слів та речень (середню довжину і загальну кількість), порівняно з попередніми показали на 20-25% гіршу точність при незмінній повноті. Найгірші результати показали тести, що базувались на підрахунку кількості та послідовності заголовних літер, а також на підрахунку загальної довжини тексту (як після видалення стоп-слів і пунктуації, так і без видалення).

І останній розглянутий метод – застосування т.з. алгоритму лусок [4]. Ідея алгоритму полягає у наступному. Для кожного десятислів’я тексту розраховується контрольна сума, яку автори назвали лускою.

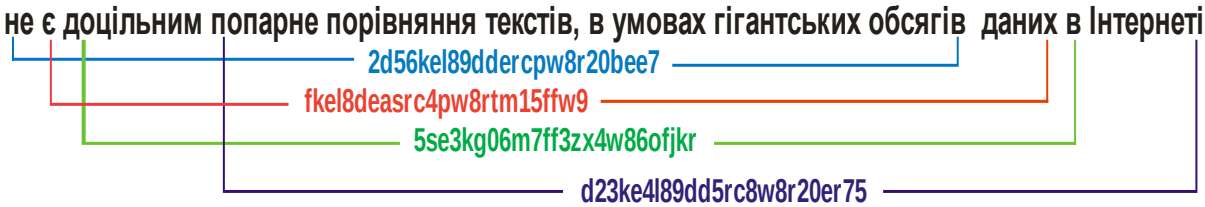

Рис.1 Підрахунок контрольних сум

Десятислів’я йдуть з перекриттям, тобто так, щоб усі варіанти були врахованні. Потім з усієї множини контрольних сум (очевидно, що їх стільки ж, скільки слів у документі мінус 9) відбираються тільки ті, які діляться на деяку константу (наприклад, 25). Оскільки значення контрольних сум розподілені рівномірно, критерій виборки ніяк не прив'язаний до особливостей тексту. Ясно, що повтор навіть одного десятислів’я - вагома ознака дублювання, якщо ж їх багато, скажімо, більше половини, то з високою вірогідністю можна стверджувати що має місце копія. Очевидно, що у такий спосіб можна визначати відсоток перекриття текстів, виявляти всі його джерела й т.і.

Недоліком алгоритму “лускування” є малоефективна робота з невеликими текстами. Так само в алгоритмі від вибору підрядків залежить ймовірність випадкових повторів, тобто значення розмірів підрядків

повинне бути досить великим, але при цьому й досить малим, щоб типові зміни в тексті не зруйнували більшу частину “лусок”.

Висновок

Очевидно, що розглянуті алгоритми не є ідеальними при розв’язанні задачі визначення нечітких дублікатів. З метою підвищення ефективності можливі варіанти об’єднання декількох алгоритмів. Наприклад, за допомогою методу “описових слів” можна визначити, до якого класу належать документи, що перевіряються, оскільки кожний згенерований вектор однозначно визначає цей клас. Після цього визначити дублікати в конкретному класі документів, використовуючи методи сигнатур, що базуються на аналізі спецсимволів. У цьому випадку можливе збільшення результативності визначення дублікатів у конкретному класі документів.

І наостанок необхідно зазначити, що дублювання текстів в інформаційних потоках не завжди є негативним явищем з точки зору користувача, який використовує Internet для бізнес-цілей. Прикладом такого винятку є, наприклад, визначення рейтингу торгової марки, коли підраховується кількість републікацій прес-релізів. Також можливе використання кількості дублювань як ознаки «міри важливості» того чи іншого повідомлення тощо.

Література

1. *Ильинский, С.В.* Эффективный способ обнаружения дубликатов web документов с использованием инвертированного индекса / С.В. Ильинский, А.Д. Мелков, М.С. Кузьмин, И.В. Сегалович // ВебПроекты, №19, 2008 http://webmastera.org/files/File/secur/FindClonDoc.pdf
2. *Романовский И. В.* Дискретный анализ: Учебное пособие для студентов специализирующихся по прикладной математике и информатике / 3-е изд., перераб. и доп. — СПб.: Невский Диалект, 2003. - 320 с.
3. *Косинов, Д.И.* Использование статистической информации при выявлении схожих документов / Д. И. Косинов // Интернет-математика 2007 : сборник работ участников конкурса. - Екатеринбург: Изд-во Урал. ун-та, 2007. - С. 84-91.
4. *Andrei Z. Broder*. Identifying and Filtering Near-Duplicate Documents, COM’00 // Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching, 2000. – P. 1-10.