УДК 004.912

К.т.н., доцент Орлова М.М., студент Зайцев І.О.

Національний технічний університет України
«Київський політехнічний інститут»

МОДИФІКОВАНИЙ АЛГОРИТМ ВИЯВЛЕННЯ НЕЧІТКИХ ДУБЛІКАТІВ WEB-ДОКУМЕНТІВ

Abstract

Mariia N. Orlova, assoc. prof., PhD; Zaitsev Igor, student
Comparative of near-duplicate detection algorithms of Web documents

The work is comparative experimental investigation of most popular modern methods of near-duplicate detection for textual documents. Algorithms compared by three parameters: completeness, precision and computer resource usage. A new algorithm, having higher quality parameters than the existing approaches is proposed.

Вступ

Проблема виявлення нечітких дублікатів web-документів є однією з найважливіших і найважчих задач аналізу веб-даних і пошуку інформації в Інтернеті. Актуальність цієї проблеми визначається різноманітністю додатків, у яких потрібно враховувати «подібність», наприклад, текстових документів — це й поліпшення якості пошукового індексу та архівів пошукових систем за рахунок видалення надлишкової інформації, і об'єднання статей у сюжети за подібністю їх по змісту, і фільтрація поштового й пошукового спаму, і визначення порушень авторських прав при незаконному копіюванні.

Основною перешкодою для успішного розв'язку цієї задачі є гігантський обсяг даних, що містяться у базах даних сучасних пошукових машин. Такий обсяг робить неможливим «пряме» її вирішення шляхом попарного порівняння текстів документів за прийнятний час. Тому останнім часом велика увага приділяється розробці методів зниження обчислювальної складності алгоритмів пошуку дублікатів web-документів за рахунок зменшення обсягу документу до «змістового» мінімуму: попередня обробка тексту (видалення html-тегів, розділових знаків, прийменників, сполучників і т. д.), хешування певного фіксованого набору «вагомих» слів або речень документа та інші методи.

Додатковими проблемами виявлення нечітких дублікатів web-документів є множини текстів з невеликими змінами відносно вихідного документа та короткі тексти, у яких важко виділити «змістову» частину.

Мета та задачі дослідження

Метою даного дослідження є розробка модифікованого алгоритму пошуку нечітких дублікатів web-документів з покращеними параметрами повноти, тобто кількості знайденних пар дублікатів текстів, та точності визначення дублікатів.

Для цього необхідно:

1. Провести аналіз якості найбільш відомих, різноманітних, ефективних з обчислювальної точки зору алгоритмів.
2. Виявити їх переваги, недоліки та обмеження.
3. Зберігши переваги, мінімізувати або усунути недоліки базових алгоритмів, покращити якісні параметри.

Теоретичні відомості

Основні алгоритми визначення нечітких дублікатів web-документів використовують синтаксичний або лексичний принципи обробки тексту документа.

Синтаксичний підхід використовується в алгоритмі «шинглів» [1, 2, 3], «надбудовах» над ним [4, 5]: алгоритмах «мегашинглів» і «супершинглів» та їх модифікаціях по довжині шингла, способу перекриття тексту послідовностями слів – із взаємним перетином і без, способу вибірки підмножини шинглів з повної множини для наступного кроку алгоритму (наприклад, «супершинглування» логарифмічної вибірки [6]).

Алгоритми, засновані на лексичних принципах [7, 8, 9], поділяються на «локальні» – не використовують загальну статистику колекції документів і «глобальні», які спираються на частотні характеристики слів всієї колекції текстів.

Перевагами «локальних» алгоритмів є: абсолютна незалежність від додавання нових документів у колекцію документів та невеликі обчислювальні витрати.

Недоліками «локальних» алгоритмів є: невисока точність і повнота виявлення дублікатів.

Перевагами «глобальних» алгоритмів є: висока точність і повнота виявлення дублікатів (за умови наявності великої початкової колекції

документів), причому обидві характеристики зростають зі збільшенням розміру обробленої колекції документів.

Недоліками «глобальних» алгоритмів є великі обчислювальні витрати, що зростають пропорційно збільшенню розміру обробленої колекції документів.

Постановка задачі

Важливими вимогами до алгоритмів, що використовують частотні характеристики слів всієї колекції документів є як її великий розмір, так і її якість. Для забезпечення вірних «середніх» частотних характеристик слів, документи початкової колекції текстів мають обиратися з якомога більшої кількості напрямків знань. Наприклад, не можна створювати колекцію текстів тільки на основі медичних документів, так як частота відповідних термінів буде перевищувати середню частоту вживання цих слів.

Ще однією вимогою, що випливає з головного недоліку «глобальних» алгоритмів, є вимога до «пакетності» додання нових документів у коллекцію.

Цих недоліків позбавлені алгоритми, що спираються тільки на «локальні» характеристики тексту, тому базовими для дослідження є два алгоритми, що засновані на обробці слів та речень тексту.

А1. Перший алгоритм засновано на обробці слів тексту. Для документа будується частотний словник, впорядкований по зменшенню частот слів. Сигнатурою документа є контрольна сума *n* об'єднаних у рядок найчастіше вживаних слів, впорядкованих за алфавітом. Головними недоліками алгоритму є залежність сигнатури документу від: найчастіше вживаних слів та їх порядку об'єднання у рядок. Тобто, штучне підвищення частоти появи будь-якого слова у тексті, при якому воно потрапить в *n* найчастіше вживаних слів, змінить сигнатуру документа.

А2. Другий алгоритм засновано на обробці речень тексту. Документ розбивається на речення, які сортуються по зменшенню довжини речення у словах. При рівності довжин – в алфавітном порядку. Сигнатурою документа є контрольна сума *m* об'єднаних у рядок найдовших речень. Головними недоліками алгоритму є залежність сигнатури документа від: найдовших речень та порядку слів у них. Тобто, зміна чи перестановка будь-якого слова в будь-якому з *m* найдовших речень змінить сигнатуру документа.

Завдання полягає в розробці алгоритму, що використовує «локальні» характеристики тексту, потребує мінімальних обчислювальних ресурсів, має підвищену точність і повноту виявлення дублікатів. Додатковими

вимогами до якості алгоритму є: стійкість до «невеликих» змін вихідного документа і впевнена обробка коротких документів.

Модифікований алгоритм пошуку нечітких дублікатів web-документів

Модифікований алгоритм пошуку нечітких дублікатів web-документів дозволяє суттєво підвищити повноту при збереженні високої точності виявлення дублікатів текстів.

Алгоритм складається з наступних кроків:

1. Для кожного документа колекції web-документів формується не більше трьох записів наступного виду:

sign1, id, length, sentences, sign1, sign2, sign3, wsign1, ..., wsign5
sign2, id, length, sentences, sign1, sign2, sign3, wsign1, ..., wsign5
sign3, id, length, sentences, sign1, sign2, sign3, wsign1, ..., wsign5, де

sign1...sign3 — сигнатури трьох найдовших речень документа, впорядковані по зменшенню довжини,

id — ідентифікатор документа,

length — довжина документа у словах,

sentences — число речень у документі,

wsign1...wsign5 — сигнатури п'яти найдовших слів документа, впорядковані по зменшенню довжини слів.

2. Для коротких документів, що складються з одного або двох речень, сигнатури відсутніх речень та слів замінюються нулями.

3. Отриманий файл сигнатур сортується і для записів із співпавшими сигнатурами найдовших речень (тобто першими полями) формуються послідовності виду:

id1, length1, sentences1, sign11, sign21, sign31, wsign11,..., wsign51
id2, length2, sentences2, sign12, sign22, sign32, wsign12,..., wsign52
...

4. Послідовності впорядковуються по зростанню довжини документа, а при рівності довжин по ідентифікатору документа. При цьому відношення довжин текстів двох сусідніх елементів послідовності не повинно перевищувати деякого порогу, обумовленого заданим мінімальним коефіцієнтом подібності документів. Наприклад, для коефіцієнта 0.85, оптимальне значення порога дорінює 1.15.

5. При перевищенні порогу формування поточної послідовності закінчується і починається формування нової. У результаті вихідний файл сигнатур перетвориться у множину порівняно коротких послідовностей, які містять монотонно зростаючі елементи довжин документів.

6. Файл послідовностей сортується, з нього виключаються дублі та послідовності, які цілком входять в інші, більш довгі послідовності. Після такої нормалізації виходить файл послідовностей невеликого розміру, у якому входження елементів у різні послідовності становить біля 10%.

7. Оскільки всередині послідовності елементи впорядковані по зростанню довжин документів, для кожного елемента існує локальний окіл невеликого розміру, яка обмежена тим самим граничним значенням відношення довжин речень сусідніх елементів послідовності.

Пошук дублікатів документа для кожного елемента послідовності здійснюється за наступними правилами:

- пошук відбувається тільки у межах локального околу;
- пари порівнюються, тільки якщо відношення кількості речень у них не перевищує деякого порога (1.20), а з п'яти сигнатур найдовших слів співпадають не менше двох у будь-якому порядку;

Два документи вважаються дублікатами, якщо в них співпадають сигнатури найдовших речень або для документів, що складаються більш ніж з п’яти речень із трьох сигнатур найдовших речень спідпадають дві у будь-якому порядку.

Висновки

Модифікований алгоритм має наступні переваги перед базовими алгоритмами:

1. Залежність сигнатури документа від двох параметрів: найдовших слів і найдовших речень тексту.
2. Незалежність від частоти вживання слів.
3. Відсутня потреба сортування основних параметрів по спаданню довжини, так як сукупність найдовших слів та речень двох документів порівнюються як множнини.
4. Дуже висока стійкість до незначних змін тексту вихідного документа, обумовлена збільшенням кількості параметрів та способу їх порівняння.
5. Передбачені окремі умови для коротких документів.

Проведені дослідження показали наступні параметри модифікованого алгоритму:

- Повнота виявлення нечітких дублікатів web-документів близько 95%, що на 10% перевищує повноту алгоритму А2 і на 35% перевищує повноту алгоритму А1.

- Точність алгоритму становить приблизно 95%, що на 15% перевищує точність алгоритму A2 і не менше точності алгоритму A1.

Література

1. *A. Broder, S. Glassman, M. Manasse and G. Zweig*. Syntactic clustering of the Web //Proc. of the 6th International World Wide Web Conference, April 1997.
2. *A. Broder*. On the resemblance and containment of documents // Compression and Complexity of Sequences (SEQUENCES'97), IEEE Computer Society, 1998. - pp. 21-29.
3. *A. Broder*. Algorithms for duplicate documents, 1999.
4. *D. Fetterly, M. Manasse, M. Najork*. A Large-Scale Study of the Evolution of Web Pages //WWW2003, Budapest, Hungary, May 20-24, 2003.
5. *A. Broder, M. Charikar*. Min-wise independent permutations // Proceedings of the thirtieth annual ACM symposium on Theory of computing, 1998.
6. *И. Сегалович, Д. Тейблюм, А. Дилевский*. Принципы и технические методы работы с незапрашиваемой корреспонденцией, 2003.
7. *A. Kolcz, A. Chowdhury, J. Alspector*. Improved Robustness of Signature-Based Near-Replica Detection via Lexicon Randomization /KDD 2004.
8. *K. Church, W. Gale*. Poisson mixtures // Natural Language Engineering, 1995. - pp. 163–190.
9. *S.-T. Park, D. Pennock, C. Lee Giles, R. Krovetz*. Analysis of Lexical Signatures for Finding Lost or Related Documents // SIGIR’02, Tampere, Finland, August 11-15, 2002.