УДК 316:303.621.36:004.77

Д.С. Глібов, А.С. Чуприна

Харківський національний університет радіоелектроніки, Харків

ПОШУК ЗАПОЗИЧЕНОЇ ІНФОРМАЦІЇ В ІНТЕРНЕТІ, ВИКОРИСТОВУЮЧИ АЛГОРИТМИ: TF, LONGSENT, WINNOWING

В статті наводиться опис основних типів алгоритму пошуку запозиченої текстової інформації, представляються результати дослідження алгоритмів TF, LongSent та Winnowing для пошуку нечітких дублікатів в Інтернеті. Особливість пошуку у тому, що алгоритми використовуються окремо для кожного абзацу тексту вхідних документів. Якість роботи алгоритмів розглянуто через метрики: точність, повнота, F-міра.

Ключові слова: *дублікат, алгоритм, шингл, подібність.*

Вступ

Теперішній розвиток інформаційних технологій дав користувачам Інтернету доступ до великої кількості інформації. Це призвело до зростання її дублювання. Особливо це помітно в таких галузях, як освіта та ЗМІ.

Таким чином, задача знаходження подібних запозичень набуває значущу актуальність.

Документи з запозиченою інформацією поділяються на «чіткі» та «нечіткі» дублікати. Чіткий дублікат – точна копія оригінального документа. Нечіткий дублікат – документ, в якому частково змінено зміст та/або форматування.

Задача пошуку чітких дублікатів є тривіальною на відміну від пошуку нечітких.

Огляд існуючих алгоритмів. Усі методи пошуку нечітких дублікатів можна умовно поділити на:

1. Синтаксичні методи (в яких здійснюється вибір послідовностей символів, слів та/або речень).

2. Лексичні (семантичні) методи (в яких здійснюється вибір представницьких мовних одиниць).

Основним синтаксичним методом є метод шинглування, коли вхідний документ представляють набором усіх підланцюгів послідовностей слів визначеної довжини. Такі ланцюги слів, що вибираються з певним зміщенням по лінійній структурі тексту, називають шинглами. Кількість можливих побудованих шинглів для одного речення дорівнює: кількість_слів_речення – довжина_шинглу + 1. Кожному ланцюгу співставляється хеш-код, при виборі якого забезпечуються наступні важливі властивості: рівність ланцюгів гарантує рівність кодів (тобто кодування є хеш-функція), а рівність кодів каже про високу подібність ланцюгів. Найбільш поширеними є хеш-коди SHA1 та Rabin. Необхідною умовою стає мінімальна кількість колізій для хеш-функцій. Два вхідних документи є подібними, якщо вони мають множину однакових шинглів. Такий метод має різні модифікації для більш швидкого пошуку подібностей: «супершингли» та «мегашингли» [1].

В методах лексичного типу реалізується відбір множини представницьких слів, виходячи з показників значимості цих слів. У такій множині не включаються слова зі заздалегідь фіксованого списку стоп-слов.

Список стоп-слів містить в себе прийменники, сполучники, ввідні слова і т.п. Показниками значимості служать частотні характеристики: для подальшого аналізу вибираються слова, в яких їхні частоти знаходяться в деякому інтервалі, тому що високочастотні слова можуть бути неінформативними, а низькочастотні – друкарськими помилками або випадковими словами.

Існує ряд методів, які засновані на лексичному представленні тексту: метод I-Match (обчислюється сигнатура (хеш-значення) для слів зі середнім значенням інверсної частоти слів в документах IDF), метод «опорних» слів (за визначеними правилами для кожного документа формується множина слів, для яких обчислюється сигнатура) [1].

Для знаходження запозиченої інформації також використовують алгоритми, що побудовані на принципах інформаційного пошуку: TF, TF*IDF, TF*RIDF, OptFreq [2].

В більшості випадків оцінка якості роботи алгоритмів нечітких дублікатів складається з трьох метрик: точність, повнота та F-міра:

– точність (precision) – це відношення кількості правильно знайдених алгоритмом документів-дублікатів до кількості знайдених документів. Ця метрика показує тільки ті документи (у відсотковому відношенні), які дійсно є дублікатами;

– повнота (recall) – це відношення кількості правильно знайдених алгоритмом документів-дублікатів до загальної кількості документів-дублікатів в колекції. Ця метрика не враховує кількість недублікатів, що помилково визначені як дублікати;

– F-міра – це загальна метрика, що об'єднує метрики повноти та точності [3]:

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

Мета дослідження. Перевірити роботу алгоритмів TF, LongSent та алгоритму просіювання (Winnowing) для пошуку запозиченої інформації, використовуючи для кожного абзацу тексту. Ці алгоритми відносяться до типу «локальні» алгоритми, тому що не використовують загальну статистичну інформацію про вхідний документ чи колекції, в яких знаходиться документ.

Викладення основного матеріалу

Абзац – частина письмового мовлення (тексту), що складається з одного чи кількох речень, пов'язаних між собою за змістом. Таким чином, розбиваючи текст на смислові частини та шукаючи подібності в Інтернеті, можна більш глибоко дослідити текст, ніж сучасні системи пошуку плагіату.

Для пошуку запозиченої інформації в Інтернеті використовувалася пошукова система yandex (сервіс Яндекс.XML). Сервіс має такі особливості:

– можливість відсилати 1000 безкоштовних запитів протягом однієї доби, що є дуже важливим при дослідженні трьох алгоритмів на великих об’ємах даних (на відміну від Google Search API, який дозволяє робити 100 щоденних запитів безкоштовно);

– велика довжина запиту: 40 слів даних (на відміну від Google Search API, рядок запиту якого має не перевищувати 15 слів) [5].

Опис методів та алгоритмів

– TF – будується частотний словник абзацу, який упорядковується за зменшенням частот. Потім вибираються та з’єднуються в алфавітному порядку в один рядок 6 слів з найбільшим значенням tf; якщо слова мають однакове значення tf, то упорядковуються за довжиною.

– LongSent – абзац ділиться на речення, які впорядковуються за спаданням довжини, що виражена кількістю слів, а при рівності довжин – у алфавітному порядку. Потім вибираються та зчіплюються в рядок в алфавітному порядку 2 найдовших речення [1].

– Winnowing – абзац ділиться на шингли довжиною, що дорівнює 4, з яких потім вибираються шингли з найменшими контрольними сумами CRC32 за таким приниципом:

1. «Вікно» розміром w рухається вздовж послідовності $h_1 \ldots h_n$, на кожному кроці зміщуємо вікно на одну позицію вправо.

2. Назначаємо мітку хеш-значення – мінімальне h_j у вікні.

3. Якщо в одному вікні два хеш-значення приймають мінімальне значення, то праве значення стає наступною міткою [4].

Для дослідження було обрано довжину вікна w = 7.

Обробка вхідних документів

Перед аналізом вхідні документи з текстовим контентом піддаються попередній обробці, яка складається з таких частин:

1. В html-файлах видалялась html-розмітка.

2. Видалялись з текстів стоп-слова (прийменники, сполучники, займенники) та стоп-символи (розділові знаки).

3. Змінювався регістр слів на нижній.

4. Виконувався стеммінг (обробка закінчень слів) над словами, що залишилися в тексті.

Для пошуку подібної інформації в абзацах в Інтернеті обчислювалися рядки кожного абзацу за згаданими вище алгоритмами та за допомогою сервісу Яндекс. XML відправлялися до пошукової системи yandex.

Для більш детального аналізу подібні між собою документи ділилися на речення, з яких потім будували шингли довжиною 4. Після цього проводилось міжшинглове порівняння двох файлів.

Для розуміння алгоритмів нижче наведено приклад одного з абзаців вхідного документа, що підлягав аналізу:

«Характеристики алгоритма относительно некоторых требований приложений, сформулированные выше, зависят от конкретных условий, в которые будет поставлен алгоритм. Так, степень компрессии зависит от того, на каком классе изображений алгоритм тестируется. Аналогично, скорость компрессии нередко зависит от того, на какой платформе реализован алгоритм. Преимущество одному алгоритму перед другим может дать, например, возможность использования в вычислениях алгоритма технологий нижнего уровня, типа MMX, а это возможно далеко не для всех алгоритмов. Так, JPEG существенно выигрывает от применения технологии MMX, а LZW нет.»

Після обробки текст має такий вигляд: «характеристики алгоритма относительно некоторых требований приложений сформулированные выше зависят конкретных условий будет поставлен алгоритм. степень компрессии зависит классе изображений алгоритм тестируется. аналогично скорость компрессии нередко зависит платформе реализован алгоритм. преимущество одному алгоритму другим может дать возможность использования вычислениях алгоритма технологий нижнего уровня типа mmx возможно далеко всех алгоритмов. jpeg существенно выигрывает применения технологии mmx lzw нет».

Частотний словник для алгоритму TF:

– алгоритм (8);

– некоторых (3);

– зависят (3);

– компрессии (2);

– технологий (2);

– возможность (2) ;

– mmx (2).

Отсортована колекція речень для LongSent за зменшенням слів:

1) «преимущество одному алгоритму другим может дать возможность использования вычислениях алгоритма технологий нижнего уровня типа mmx возможно далеко всех алгоритмов» (19 слів).

2) «характеристики алгоритма относительно некоторых требований приложений сформулированные выше зависят конкретных условий будет поставлен алгоритм» (14 слів).

Далі наведений результат формування шинглів довжиною 4:

– характеристики алгоритма относительно некоторых (1185057092),

– алгоритма относительно некоторых требований (2931669778),

– относительно некоторых требований приложений (1981075345),

– некоторых требований приложений сформулированные (2093578939),

– требований приложений сформулированные выше (37153806),

– приложений сформулированные выше зависят (1287399703),…,

– сформулированные выше зависят конкретных (2607234763),

– выигрывает применения технологии mmx (631784109),

– применения технологии mmx lzw (3836519715),

– технологии mmx lzw нет (3262629560).

Для алгоритму Winnowing маємо такі шингли:

– требований приложений сформулированные выше (37153806),

– условий будет поставлен алгоритм (370112271),

– поставлен алгоритм степень компрессии (338630055),

– алгоритм степень компрессии зависит (337541172),

– компрессии зависит классе изображений (296036706),

– алгоритм тестируется скорость компрессии (377740370),

– нередко зависит платформе реализован (186250871),

– зависит платформе реализован алгоритм (1015526783),

– реализован алгоритм преимущество одному (2101935448),

– одному алгоритму другим может (2166437444),

– возможность использования вычислениях алгоритма (324302963),

– типа mmx возможно далеко (278385477),

– далеко алгоритмов jpeg существенно (451223588).

Результати роботи

Результат роботи TF: «алгоритм некоторых зависят компрессии технологий возможность».

Результат роботи LongSent: «преимущество одному алгоритму другим может дать возможность использования вычислениях алгоритма технологий нижнего уровня типа mmx возможно далеко всех алгоритмов характеристики алгоритма относительно некоторых требований приложений сформулированные выше зависят конкретных условий будет поставлен алгоритм».

Результат роботи Winnowing: «требований приложений сформулированные выше условий будет поставлен алгоритм степень компрессии зависит классе изображений алгоритм тестируется скорость компрессии нередко зависит платформе реализован алгоритм преимущество одному алгоритму другим может возможность использования вычислениях алгоритма типа mmx возможно далеко алгоритмов jpeg существенно».

Результати проведених експериментів

В якості еталона 100%-ї повноти використовувалися сукупності дублікатів, знайденими усіма алгоритмами.

В якості еталона 100%-ї точності використовувалися результати, що були отримані міжшингловим порівнянням з порогом подібності 75%.

Результати дослідження пошуку плагіату в Інтернеті наведені в табл. 1.

Таблиця 1

Результати роботи алгоритмів LongSent, Winnowing та TF

Алгоритм	Повнота	Точність	F-міра
LongSent	0,92	0,98	0,95
Winnowing	0,90	0,98	0,94
TF	0,50	0,60	0,55

Крім того, для дослідження був взятий ще один алгоритм: FullText (табл. 2).

При його використанні весь оброблений текст відправлявся до сервісу «Яндекс.XML».

На прикладі, описаному вище, маємо такий запит для алгоритму FullText: «характеристики алгоритма относительно некоторых требований приложений сформулированные выше зависят конкретных условий будет поставлен алгоритм степень компрессии зависит классе изображений алгоритм тестируется аналогично скорость компрессии нередко зависит платформе реализован алгоритм. преимущество одному алгоритму другим может дать возможность использования вычислениях алгоритма технологий».

Таблиця 2

Результати роботи (алгоритм FullText)

Алгоритм	Повнота	Точність	F-міра
FullText	0,92	0,98	0,94

Висновки

Таким чином, було досліджено 4 алгоритми. За результатами дослідження видно, що найкраще, ніж метод TF, спрацювали методи LongSent, Winnowing та FullText.

В майбутньому є необхідність експериментального порівняння «глобальних» алгоритмів, що спираються на частотні характеристики по всій колекції документів, що підлягають аналізу (таких як TF*IDF, TF*RIDF, OptFreq).

Список літератури

1. Kolcz A. Improved Robustness of Signature-Based Near-Replica Detection via Lexicon Randomization [Електронний ресурс] / A. Kolcz, A. Chowdhury, J. Alspector // KDD 2004. – Режим доступу до ресурсу: WWW/ URL: http://ir.iit.edu/~abdur/publications/470-kolcz.pdf. 2004. – 20.04.2012 р. – Загол. з екрану.

2. Зеленков Ю.Г. Сравнительный анализ методов определения нечетких дубликатов для WEB-документов [Текст] / Ю.Г. Зеленков, И.В. Сегалович // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: труды 9-й Всероссийской научной конференции RCDL'2007: сб. работ участников конкурса. – Переславль-Залесский, 2007. – Т. 1. – С. 166-174.

3. Агеев М. Официальная метрика РОМИП 2006 [Електронний ресурс] / М. Агеев, И. Кураленок, И. Некрестьянов. – Режим доступу до ресурсу: WWW/ URL: http://romip.ru/romip2008/appendix_a_metrics.pdf – 20.04.2012 р. – Загол. з екрану.

4. Aiken A. Winnowing: local algorithms for document fingerprinting [Текст] / A. Aiken, S. Schleiner, D. Wikerson // In Proc 2003 ACMSIGMOD Int. Conf. on Management of Data, CA, June 9-12. – P. 76-85.

5. Яндекс Xml [Електронний ресурс] / 2012. – Режим доступу до ресурсу: WWW/ URL: http://help.yandex.ru/xml/ – 20.04.2012 р. — Загол. з екрану.

Надійшла до редколегії 8.05.2012

Рецензент: д-р техн. наук, проф. Є.П. Путятин, Харківський національний університет радіоелектроніки, Харків.

ПОИСК ЗАИМСТВОВАННОЙ ИНФОРМАЦИИ В ИНТЕРНЕТЕ, ИСПОЛЬЗУЯ АЛГОРИТМЫ: TF, LONGSENT И WINNOWING

Д.С. Глебов, А.С. Чуприна

В статье приводится описание основных типов существующих алгоритмов поиска заимствованной текстовой информации, представляются результаты исследования алгоритмов TF, LongSent и Winnowing для поиска нечетких дубликатов в Интернете. Особенность исследования заключается в том, что алгоритмы используются отдельно для каждого абзаца текста входного документа. Качество работы алгоритмов рассмотрено через метрики: точность, полнота, F-мера.

Ключевые слова: *дубликат, алгоритм, шингл, схожесть.*

SEARCH ADOPTED INFORMATION IN THE INTERNET USING ALGORITHMS TF, LONGSENT, WINNOWING

D.S. Glibov, A.S. Chuprina

This paper present description of the main types of adopted text information search algorithms and the results of study TF, LongSent and Winnowing algorithms for searching near duplicates in the Internet. The feature of the study is the algorithms are used for each paragraphs of input docum`ents text separately. The quality of algorithms was appraised by metrics: accuracy, completeness, F-measure.

Keywords: *duplicate, algorithm, shingle, similarity.*