УДК 681.301

К.т.н, доцент Замятін Д.С., магістрант Гриненко А.Ю.

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

МЕТОД ТЕМАТИЧНОЇ КЛАСИФІКАЦІЇ ТЕКСТОВИХ РЕСУРСІВ НА ОСНОВІ АЛГОРИТМУ ШИНГЛІВ

Abstract

Denis S. Zamiatin, assoc. prof., PhD; Artem Grynenko, student
Method of text classification based on algorithm of shingles

This paper presents the method of automatic text classification based on algorithm of shingles known in near-duplicate detection. The evaluation of text classification quality was obtained. The results of experiments carried out which show mostly positive results are introduced. There are several ways to improve the method proposed.

Вступ

Розвиток та широке впровадження інформаційних технологій справляють трансформуючий вплив на всі сфери сучасного життя включаючи економіку, владу, науку та освіту. На сьогодні обсяги інформації, що зберігається у традиційній формі, ускладнюють ефективну роботу з нею – зберігання, розповсюдження, пошук, облік тощо. Вирішення проблем ефективного аналізу й обробки інформації лежить на шляху використання сучасних засобів обчислювальної техніки та інформаційних технологій і потребує подання інформації в електронній формі. Електронна форма подання дозволяє зберігати інформацію найбільш надійно і компактно, розповсюджувати її набагато оперативніше і ширше, а, крім того, надає можливості маніпулювання нею, яких не могло бути при традиційній – паперовій формі зберігання [1]. Найрозповсюдженішими прикладами електронних сховищ текстової інформації є електронні архіви, бібліотеки, системи документообігу та інші. Обробка й аналіз змісту текстів звичайно були ручною роботою, але зі збільшенням обсягів інформації це стало неефективним і виникла потреба в автоматизації цих процесів. Однією з актуальних задач аналізу текстових даних є класифікація – визначення тематики заданого тексту і віднесення його до однієї з наперед визначених категорій.

Постановка задачі

Метою даної роботи є оцінка якості методу автоматичної класифікації тестів за допомогою алгоритму шинглів. Формально задачу класифікації можна представити так: нехай задано скінчену множину категорій $C = \{c_1,...,c_n\}$, скінчену множину документів $D = \{d_1,...,d_m\}$, F – цільова функція, яка по парі $< d_i, c_j >$, визначає, чи відноситься об'єкт d_i до категорії c_j. Задача класифікації полягає у побудові функції F', яка максимально наближена до F.

Дана класифікація називається чіткою бінарною, тобто мається на увазі, що існують тільки дві категорії, які не перетинаються. Існує інший вид класифікації, в якому враховується те, що досліджуваний текст може відноситися не тільки до однієї, а відразу до декількох категорій з різним ступенем належності, тобто категорії можуть перетинатися між собою. Такий вид класифікації називається ранжирування. Найпростіший спосіб щоб перейти від функції ранжирування до точної класифікації – обрати категорію c_j, для якої функція ранжирування приймає найбільше значення [2].

Метод розв'язання задачі

Тут задачу класифікації запропоновано вирішувати за допомогою алгоритму шинглів. Відомі застосування даного алгоритму для вирішення задачі пошуку нечітких дублікатів [3]. Але враховуючи подібність обох цих задач, було зроблено припущення, що використання алгоритму шинглів для класифікації текстів матиме позитивний результат. Основна ідея алгоритму шинглів полягає в розбитті тексту на послідовності слів однакової довжини – шингли (від англ. shingle – черепиця). Важливою особливістю даного алгоритму є те, що шингли виділяються не один за одним, а накладаються для запобігання втрати інформації.

Нижче наведено приклад розбиття фрази на шингли довжиною в чотири слова:

Вихідна фраза: *просуваючись вперед наука невпинно перекреслює сама себе*

1-й шингл: *просуваючись вперед наука невпинно*

2-й шингл: *вперед наука невпинно перекреслює*

3-й шингл: *наука невпинно перекреслює сама*

4-й шингл: *невпинно перекреслює сама себе*

Реалізація класичного алгоритму шинглів передбачає кілька основних етапів:

1. Канонізація тексту

На початковому етапі вхідні тексти приводяться до канонічного вигляду. Тобто з них видаляються усі знаки пунктуації, елементи форматування, зайві пробіли, HTML теги та стоп-слова (сполучники, частки, займенники, вигуки і т.д.), а також всі слова приводяться до початкової форми.

2. Розбиття тексту на шингли

Даний етап передбачає розбиття канонізованого тексту на підрядки однакової довжини (шингли). Найчастіше в якості кроку обираються символи або слова і значно рідше - речення. Шингли повинні виділятися з тексту не один за одним, а накладатися. Мінімальна відстань (d), між двома сусідніми шинглами дорівнює 1 слово/символ і гарантує, що при розбитті тексту на шингли не буде жодних втрат інформації.

3. Знаходження контрольних сум шинглів

Після того, як отримано множину шинглів для текстового документу, необхідно обчислити їх контрольні суми. Для цього використовують хеш-функції: md5, crc, sha тощо.

4. Побудова образу документу

В класичному алгоритмі шинглів процес побудови образу передбачає відбір лише тих контрольних сум, які будуть використовуватися при порівнянні двох документів. Звичайно це фіксована за розміром вибірка, яка містить деякі, відібрані за певним критерієм, шингли(наприклад задану кількість мінімальних по значенню контрольних сум). У випадку використання алгоритму шинглів для класифікації текстів, кожний шингл має брати участь у порівнянні, тому в образ документу потрапляють всі без виключення шингли.

5. Пошук однакових під-послідовностей

На останньому етапі відбувається порівняння двох образів шляхом визначення ступеню їх схожості і входження.

Результати експериментів

Для експериментів були сформовані чотири категорії текстів (енергетика, медицина, політика, транспорт), по 13 текстів у кожній. В якості текстів були відібрані автореферати дисертацій з електронного ресурсу Національної бібліотеки України імені В.І. Вернадського.

Процентна частка тексту в категорії визначалося як відношення кількості шинглів з тексту, що аналізується, які було знайдено у всіх текстах даної категорії, до суми аналогічних значень для всіх категорій:

$$R = \frac{S(c_i)}{\sum_{j=1}^{N} S(c_j)},$$

де $S(c_j)$ – функція, що визначає кількість шинглів, що були знайдені в усіх текстах категорії c_j, N – кількість доступних категорій.

Перший експеримент: перевірка алгоритму в залежності від довжини шинглу (від 3 до 6 слів). Досліджувалися вісім текстів, по два з кожної категорії. Результати експерименту показали, що в більшості випадків (65,63 %) алгоритм спрацював правильно. Але результати виявилися досить рівномірними, наприклад текст з категорії "енергетика" увійшов до визначених категорій з такими показниками: "енергетика" - 32,85 %, "транспорт" - 32,12 %, "медицина" - 17,52 % і "політика" - 17,52 %. Невдалі спроби віднесення текстів до певних категорій можна частково пояснити близькістю категорій, наприклад текст, в якому йдеться про газотранспортну систему, досить важко чітко віднести до однієї з категорій: "транспорт" або "енергетика".

Другий експеримент частково пояснює рівномірність результатів першого. Був проаналізований один і той самий текст з розбивкою на шингли довжиною від 2 до 6 слів, з метою визначити, які саме шингли "відбирає" алгоритм і які формують питому частку тексту в кожній категорії. Результати показали, алгоритм працює правильно, знаходить точний перетин досліджуваного тексту з іншими текстами визначених категорій, але значна частина шинглів, які вплинули на результат – це загальні фрази, які притаманні науковому тексту, і зустрічаються в усіх категоріях. Тому отримані результати такі рівномірні.

Висновки

Запропонований спосіб класифікації текстів за допомогою алгоритму шинглів показав у більшості випадків позитивний результат. Невдалі спроби класифікації частково пояснюються значним перетином множин шинглів з визначених категорій, а частково впливом на фінальний результат входжень в тексти загальних фраз, які не несуть смислового навантаження. У подальшому описаний метод може бути вдосконалено завдяки покращенню відбору шинглів, які приймають участь в порівнянні образів документів. Цього можна досягти поліпшенням якості канонізації, та ігноруванням шинглів, які входять в перетин певної, досить великої, кількості категорій.

Література

1. *Горный Е., Вигурский К.* Развитие электронных библиотек: мировой и российский опыт, проблемы, перспективы // Интернет и российское общество - Москва: Центр Карнеги, 2002. - С. - 279.

2. *Ланде Д.В., Фурашев В.М.* Питання класифікації *та* розпізнавання інформації при побудові інформаційно-аналітичних систем. Інформація і право, 2011. - N 3. - С. 142-155.

3. *J Prasanna Kumar, P Govindarajulu* Duplicate and Near Duplicate Documents Detection: A Review - European Journal of Scientific Research, Inc. 2009, pp.514-527.

URL: http://www.eurojournals.com/ejsr_32_4_08.pdf