УДК 004.021

АЛГОРИТМ ВИЗНАЧЕННЯ СХОЖОСТІ ТЕКСТІВ НОВИН НА ОСНОВІ ПОЛІНОМІАЛЬНОГО ХЕШУВАННЯ

Гранік Михайло, Месюра Володимир

Вінницький національний технічний університет

Анотація

Метою роботи є розробка алгоритму визначення схожості текстів новин. У роботі запропоновано алгоритм порівняння схожості текстів новин на основі поліноміального хешування. Алгоритм може бути використано для кластеризації текстів новин.

Abstract

The purpose of the paper is to develop the algorithm to determine the similarity between news texts. New algorithm based on polynomial hashing is proposed in this paper. This algorithm can be used for clusterization of the news texts

Вступ

Проблема визначення схожості текстів новин є дуже актуальною. Отримання чисельної міри схожості текстів новин може бути ефективно використане для задачі кластеризації новин. Кластеризація новин, у свою чергу, є важливою практичною проблемою, адже результати її розв'язання можуть бути використані у агрегаторах новин та у системах оцінювання правдопобдібності інформації текстів новин.

Визначення схожості текстів (не обов'язково текстів новин) також є важливою практичною проблемою. Таке визначення широко використовується у біоінформатиці (наприклад, у генній онтології). Переважно у цій області його використовують для визначення схожості генів, протеїнів тощо.

Визначення чисельної міри схожості текстів також використовується для пошуку інформації, класифікації текстів, автоматичного визначення теми текстів, автоматичної генерації запитань та відповідей на них, машинного перегляду, анотування тексту (визначення підмножини тексту, що передає його основну ідею) [1].

Існує декілька шляхів визначення схожості текстів. Серед них виділяють є визначення косинусного коефіцієнта, обрахування коефіцієнта Жаккара, коефіцієнта Соренсена, коефіцієнта Сімпсона, коефіцієнта Браун-Бланке, коефіцієнта Кульчинського [1, 2].

Недоліками перерахованих методів є те, що вони краще працюють для порівняння множин, ніж саме для порівняння текстів. Якщо використовувати їх без жодних модифікацій, то слова, вжиті у різних відмінках, числах, тощо будуть вважатись різними елементами множин. Метою даної роботи є розробка алгоритму, позбавленого цього недоліку.

Поліноміальне хешування

Хешування рядків — процес встановлення однозначної відповідності між рядком та чисельним значенням з певного фіксованого чисельного проміжку. Варто зазначити, що одному рядку завжди відповідає рівно одне чисельне значення, однак одному і тому самому чисельному значенню можуть віповідати декілька рядків.

При поліноміальному хешуванні значення хешу для заданного рядка довжини n обраховується за формулою:

(1)

де коефіцієнти a – чисельні значення, поставлені у відповідність кожному символу рядка (наприклад, номер у таблиці ASCII), p – показник многочлена, M – число, що визначає інтервал, до якого належитиме хеш (легко бачити, що можливі значення хешу належать чисельному діапазону [0; M – 1]). Тобто рядку ставиться у відповідність поліном із коефіцієнтами, рівними чисельним значенням кожного із символів рядка. Степінь полінома дорівнює довжині відповідного рядка.

Переваги поліноміального хешування є такими:

1) Простота визначення.

2) Можливість ефективного обрахунку поліноміального хешу заданого рядка (складність обчислення — O(N), де N — довжина рядка).

3) Можливість ефективного обчислення хешу будь-якого підрядка.

Поліноміальні хеші використовуються в багатьох алгоритмах, що працюють із рядками. Відомий приклад — алгоритм Рабіна-Карпа, мета якого — знайти всі позиції входження одного рядка в інший як підрядка.

Від вибору параметрів формули визначення поліноміального хешу залежить ймовірність колізії (тобто випадку, коли двом різним рядкам відповідають однакові значення хешу). Зазвичай, значення p обирають так, щоб це було просте число, більше за будь-яке значення коефіцієнтів a. Звичайно, збільшення значення M зменшує ймовірність колізії (однак збільшує обсяг пам'яті, необхідний для зберігання хешу рядка).

Якщо припустити, що значення полноміального хешу — рівномірно розподілена випадкова величина, то відповідно до парадоксу днів народження, ймовірність колізії для відносно невеликих значень M є достатньо великою. Так, для M= 1000000000 достатнь лише близько 30000 рядків для того, щоб колізія відбулась з ймовірністю 0.5, і приблизно 67000 рядків — для того, щоб вона відбулась з ймовірністю 0.9. Для уникнення колізій часто для одного рядка обраховують декілька значень поліноміального хешу — наприклад, по двом різним модулям. Це зменшує ймовірність колізії. Інколи один або декілька з цих модулів обирають випадковим чином.

Алгоритм визначення схожості новинних текстів на основі поліноміального хешування

Пропонується наступний алгоритм визначення схожості текстів новин на основі хешування:

1) Проведення операції стемінгу над усіма словами усіх текстів, що розглядаються.

Стемінг – операція скорочення слів шляхом видалення із них неважливих частин, таких як префікс, суфікс чи закінчення (проте вважати, що в результаті застосування операції стемінгу кожне слово замінюється на його корінь, некоректно). Застосування алгоритмів стемінгу є поширеним у пошукових системах. Саме стемінг допомагає із проблемою, описаною вище – після його застосування різні словоформи вважатимуться одним і тим же словом.

2) Видалення стоп-слів

Стоп-слова (або шумові слова) – це такі слова у тексті, що не несуть змістовного навантаження. Під стоп словами зазвичай мають на увазі прийменники, частки, деякі інші окремі слова інших частин мови. Так як стоп-слова не несуть змістовного навантаження, їх врахування при обрахунку схожості текстів можуть суттєво спотворювати отримані результати.

3) Прибирання із тексту усіх пробільних символів та пунктуаційних знаків.

4) Зведення отриманого тексту до нижнього регістру.

5) Розбиття тексту на синтаксичні n-грами.

Синтаксичним n-грамом називається підрядок отриманого тексту із n символів. Мета цього етапу алгоритму – отримати вектор, кожен елемент якого є унікальним n-грамом (будемо вважати два n-грами різними, якщо відрізняються їх стартові позиції у тексті і не звертатимемо уваги на їхнє значення)

6) Обрахування поліноміального хешу кожного із отриманих n-грамів.

Таким чином, після цього кроку для кожного тексту буде отримано вектор чисел, кожне з яких є обрахованним значенням хешу для деякого n-граму із цього тексту.

7) Визначення схожості текстів на основі коефіцієнту Жаккара для відповідних їм векторів, що зберігають хеші n-грамів.

Коефіцієнт Жакара обраховується наступним чином. У відповідність кожному із векторів ставиться відповідна множина хешів. Коефіцієнт Жаккара визначається як частка від ділення потужності множини перетину двох отриманих множин на потужність множини об'єднання даних множин. Чим ближчим до одиниці є значення коефіцієнту Жаккара, тим більш схожими вважаються тексти. Цей метод є узагальненим методом порівняння двох множин на схожість. Саме отримане значення коефіцієнту і вважається мірою схожості текстів новин.

Висновки

Розроблено алгоритм визначення схожості текстів новин на основі поліноміального хешування. Цей алгоритм є кращим у порівнянні із деякими класичиними алгоритмами (такими, як, наприклад, визначення косинусного коефіцієнту, визначення коефцієнту Жаккара у його класичному вигляді) за рахунок того, що він спеціалізується саме на порівнянні текстів, а не на порівнянні звичайних множин. Такий ефект досягається завдяки використанню операції стемінгу, прибиранню стоп-слів, використанню синтаксичних n-грамів. Також використання поліноміального хешування дозволило пришвидшити машинний час порівняння текстів (адже порівнювались числа, а не рядки).

На основі даного алгоритму розроблено та реалізовано відповідне програмне забезпечення. Отримані результати засвідчують коректність розробленого алгоритму.

Алгоритм може бути використаний і для порівняння більших текстів. В такому випадку доцільно використовувати не весь вектор хешів n-грамів, а якусь його підмножину.

Розроблений алгоритм може бути вдосконалено шляхом визначення оптимальних значень довжини n-грамів.

Список використаних джерел:

1. Singhal Amit. Modern Information Retrieval: A Brief Overview / Singhal Amit // Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4) .– 2001.– P. 35-43.

2. Матеріали курсу Data Mining, що викладався у University of Utah [Електронний ресурс] .– Режим доступу до матеріалів: http://www.cs.utah.edu/~jeffp/teaching/cs5955/L4-Jaccard+Shingle.pdf.