Приложение на Алгоритми за Изчисление на Сходство и Вариация на Текстови Низове

Борис Банков, Иван Куюмджиев

Икономически университет – Варна,
бул. "Княз Борис I" 77, Варна 9002, България

Abstract. In the field of computational statistics exists a problem of how to measure and calculate the similarity between two or more pieces of text. The issue becomes more prominent when we try to apply the rules of transliteration to names of people or places and try to match them with their actual names. During our work on a software system for finding and matching scientific publications, written by our academic staff we conducted a study on algorithms for calculating similarity and variance in text strings.

***Keywords**: variance, algorithms, calculating text similarity*

Въведение

Част от мисията на Икономически университет – Варна е да следва развитието на високите технологии и да въвежда иновации в областта на отчитането на публикационната дейност на академичния състав. За тази цел се разработи проект, който да спомогне дейностите по отчитане на публикациите на научната общност в ИУ – Варна и процедурите, свързани с акредитация и рейтингови класации.

В тази връзка е реализирана електронна база от данни за публикациите на преподавателите в ИУ-Варна, която е базирана на старата информационна система на библиотеката на университета. Процесът е по-подробно описан в доклада „Миграция на неструктурирани данни в релационна схема: проблеми, решения, алгоритъм“. За управление на информацията в електронната база от данни се изгради уеб базирана система, която позволява на преподаватели от ИУ – Варна да имат собствен потребителски профил с достъп до публикациите си. Основният модул, залегнал в миграцията между двете информационни системи, представлява алгоритъм за изчисление на сходство между **имената на автори** в старата информационна среда и **имената на регистриралите се преподаватели** в новата уеб система.

Целта на текущия доклад е да представи алгоритмите за изчисление на сходство и вариация в текстови низове и да аргументира избора на подход, които се използва в бизнес логиката на уеб базираната система за регистриране на публикационна дейност.

Дефиниране на проблема

Основната функция на разработената уеб система е да позволи на академичната общност в ИУ – Варна да поддържа архив с актуална информация за своята публикационна дейност. В системата са дефинирани три основни градивни единици – **публикация**, **автор** и **онлайн профил на преподавател**. В тази връзка е необходимо всеки преподавател да се регистрира в системата, да открие публикациите си, които се пазят в базата от данни и да ги синхронизира с онлайн профила си. В момента на изграждане на системата всяка публикация представлява единица, която може да е свързана с няколко автора, както и за всеки автор може да има множество публикации.

Основното предизвикателство пред нас е, че авторите в старата информационна система не могат да бъдат обвързани с единицата онлайн профил, поради две причини:

1. Възможно е да съществуват записи за автори, които няма да се регистрират или да имат достъп до системата. Това са напр. съавтори на публикации, които не са служители в ИУ – Варна.
2. Възможно е един преподавател да присъства с грешно изписани имена, или името му да бъде съкратено, според библиографското описание на публикацията.

За тази цел системата е необходимо да провери името на преподавателя от онлайн профила му и името на автора в библиографския запис, да отчете приликите и при достигане на определена висока степен на сходство да позволи синхронизация. За да се реализира тази функционалност сме приложили алгоритмите на т.н. размито търсене (fuzzy search algorithms).

Изследване на алгоритми

Изборът на подход за откриване на съответствие започва от възможностите под формата на методи, които се използват в програмната среда. В случая PHP предоставя три основни метода:

- similar_text() метод – използва броя съвпадащи символи между два текстови низа * 200 / (дължината на текст 1 + дължината на текст 2), както е описано в (Oliver, 1994);
- levenshtein() метод – изчислява броя символи за размяна, вмъкване или премахване, за да превърне текст 1 в текст 2 (Levenshtein, 1965);

- soundex() метод – чрез фонетична схема оценява транслитерирани текстови низове с еднакъв първи символ и специален индекс от 3 цифри (Knuth, 1973).

Сложността на levenshtein метода може да се определи като O(m*n), където m и n са дължините на съответните низове, което дава по-точни резултати спрямо O(max(n,m)**3) на similar_text(), но изисква повече време за изчисление.

Проучванията ни продължават с изследване на алгоритми с по-сложна реализация, както следва:

- cosine similarity – текстовите низове се превръщат във вектори, използвайки модела на векторно пространство (Smith, Danielsson & Jonsson, 2012) и се измерва косинуса на ъгъла между тях;
- Hamming distance – измерва броя различни символи на съвпадащи позиции в два текстови низа с еднаква дължина (Hamming, 1950);
- Jaro-Winkler distance – използва метрика за проверка на текстови низове с обща представка и е подходящ за проверка на грешно изписани имена, преведени от друг език или съдържащи ASCII символи (Winkler, 1999).

Приложение на Jaro-Winkler distance

Най-подходящият алгоритъм за изпълнение на задачите по синхронизация на публикационна дейност е Jaro-Winkler distance, поради факта, че отчита възможността за грешно изписване на имена, както и малката дължина на сравняваните текстови низове. В по-къси думи обикновено срещането на вариация дава по-високо размиване. За реализация на програмната логика при сравнение на имената се разглеждат два варианта.

При първият се взимат и се сравняват по двойки името, презимето и фамилията на единицата автор и на единицата онлайн профил на преподавател. Тъй като резултата от алгоритъма е в областта {0,1}, като 1 е пълно съвпадение, при сумиране на резултатите се търси средна стойност по-голяма от 0,9. На следваща стъпка се проверяват по двойки фамилията и първи символ от малкото име (съкратено библиографски) на автора и онлайн профила и се търси среден коефициент на съвпадение над 0.95.

Заключение

Концепцията за алгоритмите за размито търсене е част от по-голяма област на изследване на проблеми свързани с изчисление на сходство, а именно извличане на информация и откриване на знания. Въпреки, че при сравнение на имена няма възможност да се отчете семантична прилика между два текстови низа, в така представения доклад се разглежда един от класическите проблеми на компютърните системи, а именно откриването и премахването на дублиращата се информация.

Използвана литература

1. Oliver, I. (1994). Programming Classics: Implementing the World's Best Algorithms, Prentice-Hall, Inc.;

2. Levenshtein, V. (1965). Binary codes capable of correcting deletions, insertions, and reversals, Doklady Akademii Nauk SSSR, 163(4) p. 845-848.

3. Knuth, D. (1973). The Art Of Computer Programming, vol. 3: Sorting And Searching, Addison-Wesley, p. 391-392.

4. Smith, C., Danielsson, H., & Jönsson, A. (2012). A good space: Lexical predictors in vector space evaluation. In Lrec 2012–Eighth International Conference on Language Resources and Evaluation p. 2530-2535.

5. Hamming, R. (1950). Error detecting and error correcting codes. Bell Labs Technical Journal, 29(2), p. 147-160.

6. Winkler, W. (1999). The state of record linkage and current research problems. In Statistical Research Division, US Census Bureau.