

ПОРІВНЯННЯ ТЕКСТОВИХ ДОКУМЕНТІВ У ФОРМАТІ *.DOC

Томчук О.Ю.

Науковий керівник – доц., к.т.н. Гороховський О.І.

Розв'язання багатьох важливих задач вимагає використання високопродуктивних пошуково-експертних систем. Незважаючи на високу швидкодію сучасних систем, актуальною є розробка нових принципів організації пошуку та порівняння інформації. В останні роки швидко розвивається технологія пошуку, зокрема, в глобальній мережі Internet, тому через велику кількість дублювання інформації виникає необхідність розробки експертної системи для виявлення плагіату.

Даний програмний продукт розроблений для порівняння текстових документів в вищих навчальних закладах і запобігає копіювання дипломних та курсових проектів. В роботі виконаний аналіз основних методів та алгоритмів пошуку, що найбільш придатні для розв'язання задачі пошуку однакової інформації.

Для реалізації було обрано метод порівняння та аналізу текстової інформації на основі алгоритму «шинглів» для прискорення та поліпшення роботи і співставлення текстових документів.

Реалізація алгоритму порівняння двох текстових файлів розбивається на кілька головних етапів:

- канонізації тексту;
- розбиття тексту на шингли;
- знаходження контрольних сум;
- пошук однакових підпоследовностей.

В алгоритмі шинглів реалізовано порівняння контрольних сум текстів. В програмі використано CRC32.

Канонізація тексту, що використовується в методі шинглу полягає у очищенні тексту від стоп-символів і стоп-слів, призводить всі символи рядка до нижнього регістру і повертає список, що залишилися після чищення слів.

В стоп-символи включено розділові знаки, знаки переносу рядка і табуляції. В стоп-слова включено сполучники, приводи та інші слова, які при порівнянні не повинні впливати на результат.

«Плагіат» - програмне забезпечення, що є результатом застосування методу та алгоритму порівняння текстової інформації з використанням об'єктно-орієнтованої мови програмування C#. Програма створена в навчальних цілях з метою виявлення та запобігання копіювання дипломних і курсових робіт у вищих навчальних закладах.