

УДК 004.8

Куликовська Н.А.¹, Азаров В.В.²

¹ асист. НУ «Запорізька політехніка»

² студ. гр. КНТ-537сп НУ «Запорізька політехніка»

ПОШУК ЗАПОЗИЧЕНОЇ ІНФОРМАЦІЇ В ІНТЕРНЕТІ

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. [1] Задача пошуку чітких дублікатів є тривіальною на відміну від пошуку нечітких.

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.

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. 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).

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