УДК 004.912

О.М. Дишлюк, О.П. Нечипорук

*Національний авіаційний університет*

РОЗРОБКА МЕТОДІВ ТА АЛГОРИТМІВ ІНДЕКСАЦІЇ  
ЕЛЕКТРОНИХ ТЕКСТІВ

Розглянуто загальну постановку задачі класифікації і базовий підхід до індексації електронних текстів.

Виконано аналітичний огляд сучасних алгоритмів індексації, розглянуті оцінки їх ефективності. Окремо проаналізовано алгоритми індексації для документів, що зберігаються в графічних форматах TIFF, JPEG и PDF.

Особлива увага була приділена методу кореляційної індексації текстової і графічної інформації, який дозволяє забезпечити високу релевантність пошуку та істотно скоротити його часові витрати. Метод кореляційної індексації при створенні індексу архіву електронних документів враховує кореляції між словами або ASCII символами. Облік кореляцій дозволяє при індексації автоматично виявляти термінологію (як слова, так і послідовності слів), характерну для документів, записаних в архів. Такий підхід дозволив розробити алгоритм пошуку схожих документів, при якому як пошуковий запит може використовуватися повна сторінка тексту. Пошук схожих документів дає можливість не лише шукати необхідну інформацію, а й аналізувати та класифікувати цю інформацію. В разі індексації текстової інформації з врахуванням кореляцій між ASCII символами пошук схожих документів є мовно-незалежним.

Створено декілька алгоритмів, на підставі яких буде виконана програмна реалізація індексації електронних текстів.

Спроектовано структуру програмного модуля індексації електронних текстів.

Запропоновано тести для аналізу ефективності розроблених алгоритмів.