УДК 004.912

М.В. Лупандін, к.т.н.,  
І.А. Безкровна, І.М. Юр’єва

*Національний авіаційний університет*

РОЗРОБКА АЛГОРИТМІВ СЕМАНТИЧНОЇ  
СТРУКТУРИЗАЦІЇ ДАНИХ

Проблеми попередньої обробки електронних текстів в програмних системах відповідного призначення, в тому числі у пошукових системах, не втрачають актуальності протягом останніх років. Аналіз технологій попередньої обробки електронних текстів показав, що сучасні механізми структуризації даних мають передбачати можливість аналізу даних, що представлені в будь-якому форматі – у вигляді Web-сторінок, документів Microsoft Word, файлів Adobe Acrobat тощо. Розв’язання цієї задачі суттєво полегшує пошук та доступ до інформаційного змісту даних.

Задача пошуку текстової інформації полягає в знаходженні мінімальних смислових одиниць тексту, які є релевантними запиту. Знайдені одиниці мають відповідати вимогам повноти і точності. Під релевантністю розуміється деяка бінарна функція, вхідними параметрами якої є запит користувача і чергова одиниця тексту, що підлягає аналізу. За інформаційну одиницю тексту може братися документ, абзац, речення або інші фрагменти тексту. Проблема обчислення релевантності є центральною в задачі пошуку текстової інформації.

Виконано порівняльний аналіз лінгвістичного і статистичного підходів до попередньої обробки електронних текстів. З порівняння сучасних статистичних та нестатистичних підходів до аналізу текстів на природній мові можна зробити висновки, що найбільш ефективними будуть алгоритми, які поєднають в собі швидкість і незалежність від мови статистичних алгоритмів, з високою якістю обробки семантичного підходу.

В роботі розроблено ефективні алгоритми семантичної струк-туризації даних. Запропоновано тести для аналізу ефективності алгоритмів семантичної структуризації електронних текстів.