УДК 004.4:414(045)

М.Ю. Сич

*Національний авіаційний університет*

АЛГОРИТМ СЕМАНТИЧНОЇ ОБРОБКИ ТЕКСТОВОЇ
ІНФОРМАЦІЇ

Алгоритм займається обробкою введеної послідовності слів. Він оперує словами, намагаючись знайти зв’язки між ними, встановити залежності та основні смислові лінії у введеній словарній конструкції.

У якості вхідних даних алгоритму виступає простий текстовий рядок. Цей рядок проходить набір послідовних обробок, зміст яких буде розкритий нижче.

У словосполученні є головне та залежне слово чи їх сукупність. Встановленим зв’язком будемо вважати пару «головне слово - залежна конструкція слів», у якій всі слова мають чіткий однозначний зміст та є зрозумілим характер їх взаємодії, тобто вид зв’язку.

Обробка речення приводить на кінцевому етапі до набору словосполучень, що мають перетини різними словами, тобто це просто група взаємопов’язаних пар слів, або точніше сказати змістів слів. Для кожного слова ми маємо чітко визначений зміст, а також всі зв’язки і їх характер для всіх слів, з яким це слово взаємодіє.

Для обробки слів необхідно відмовитись від простого кодування у вигляді простих послідовностей символів. Новий вид кодування слів буде зручним для зберігання інформації про слово, його структуру та основні властивості. Таким чином, першим кроком роботи алгоритму є приведення слів речення у новий формат, заснований на морфологічних та синтаксичних особливостях слів.

Алгоритм пошуку зв’язків між словами є ітераційним, тобто представляє собою набір послідовних повторюваних кроків, на кожному з яких обробляється певна група слів.