ЗАСОБИ РОЗПІЗНАВАННЯ ТИПІВ ДОКУМЕНТІВ

Леонов О.О.
Науковий керівник - проф. каф. СПЗ, д.т.н. Крісілов В.А.

При створенні сховищ даних велике значення має введення даних у систему, тому що ручна обробка великих обсягів інформації неефективна й недоцільна. Спеціально створювані модулі витягу й перетворення даних оперують потоками вхідних документів, що поступають, переважно в електронному виді. При роботі із вхідними документами ключовими завданнями є їхня формалізація й класифікація.

Завдання формалізації полягає у виділенні значеннєвих і структурних елементів з вихідних даних з метою подання документа у формі списку полей та їхніх значень, а також інформації про їхнє взаємне розташування.

Завдання класифікації документів складається у визначенні, до якого класу віднести документ, що надійшов на вхід системи, з метою передачі його конкретному оброблювачеві. Від точності класифікації залежить якість роботи системи витягу й перетворення даних.

У процесі роботи був виконаний огляд і аналіз методів класифікації: методи, засновані на інформації про структуру документа; методи, засновані на формуванні дайджестів і концепцій; імовірнісні й частотні методи; класифікація по найбільш близькому зразку; семантичний аналіз і обробка природних мов. Було прийняте рішення про створення комплексного методу, що використає одночасно кілька механізмів, тому що в чистому виді розглянуті методи економічно неефективні й невиправдано технологічно складні.