УДК 004.056.53

# Алгоритми комп'ютерного аналізу текстів на природній мові

Мелешко Є.В., к.т.н., доцент
Науковий керівник – Смірнов О.А., д.т.н., професор
*Центральноукраїнський національний технічний університет, м. Кропивницький*

Комп'ютерний аналіз текстів використовується в багатьох інформаційних системах, наприклад, в пошукових роботах, пошуковій оптимізації, реферуванні текстів, системах питання-відповідь, машинному перекладі, представленні знань, експертних системах і т.д.

Типові задачі аналізу текстів: класифікація, кластеризація, автоматичне анотування, визначення ключових понять тощо.

Рівні подання лінгвістичної та екстралінгвістичної інформації [1]: 1) акустико-фонетичний рівень; 2) морфологічний рівень; 3) лексичний рівень; 4) синтаксичний рівень; 5) семантичний рівень. При обробці текстів зазвичай розглядаються 2-5 рівні обробки інформації.

**Алгоритми морфологічного аналізу** приводять кожне слово до нормальної форми та знаходять його морфологічні характеристики [2]. До них відносяться стемінг та лематизація [3]. Стемінг – ототожнення основи семантично схожих словоформ. Лематизація також здійснює ототожнення основ слів, але враховує при цьому частини мови, до яких відносяться словоформи, що підвищує точність ототожнення.

**Алгоритми лексичного аналізу** дозволяють розпізнавати лексичні одиниці тексту. Одним з фундаментальних алгоритмів лексичного аналізу є лексична декомпозиція, яка передбачає розбивку тексту на токени [3]. Токен – найчастіше слово, але може бути окремою морфемою або словосполученням. Для проведення стемінга слід спочатку розбити текст на токени; на основі списку токенів як правило виконується синтаксична декомпозиція.

**Алгоритми синтаксичного аналізу** здійснюють синтаксичну декомпозицію – розпізнавання речень на основі символів форматування тексту [3], та генерацію дерева синтаксичного виводу речень [2]. З цією метою використовуються парсери (синтаксичні аналізатори) – програми, що перетворюють вхідний текст в деревовидну структуру даних, яка відображає синтаксичну структуру вхідної послідовності.

**Алгоритми семантичного аналізу** виявляють взаємозв'язки між термінами в документах та взаємозв'язки між різними документами [2]. На рівні семантичного аналізу визначають ключові слова у тексті, зв'язки між ключовими словами, будують семантичні мережі та онтології [1, 2, 4].

Алгоритми семантичного аналізу розбирають синтаксичні дерева тексту за допомогою онтологічної бази знань та генерують семантичну

структуру речень [2]. Потім семантичні структури речень інтегруються в семантичний граф тексту.

На даному рівні вирішуються такі задачі як визначення тематики тексту, генерація реферату, смисловий переклад з однієї мови на іншу, підтримка діалогу з користувачем на природній мові тощо.

Однозначне визначення семантичної мережі на даний час відсутнє. В інженерії знань під нею мається на увазі граф, що відображає зміст цілісного образу. Вузли графа відповідають поняттям і об'єктам, а дуги – відносинам між об'єктами. Формально семантичну мережу можна задати в наступному вигляді:

$$H = <I, C, G>,$$

де I – множина понять; C – множина типів зв'язків між поняттями; G – відображення, що задає конкретні відносини з наявних типів C між елементами I.

Існує два типи семантичних мереж: однорідні мережі з асоціативними зв'язками та неоднорідні мережі, що містять зв'язки різних типів.

Об'єктами семантичної мережі можуть бути: сутності, властивості, дії, величини [4]. Можна виділити декілька часто використовуваних класів відносин в неоднорідних семантичних мережах: ієрархії, агрегації, функціональні, семіотичні, тотожності, кореляції [4].

Розробка алгоритмів комп'ютерного аналізу текстів передбачає співпрацю лінгвістів та програмістів. Перед комп'ютерною лінгвістикою стоять, перш за все, завдання лінгвістичного забезпечення процесів збору, накопичення, обробки та пошуку інформації. Центральними науковими проблемами комп'ютерної лінгвістики є проблема моделювання процесу розуміння змісту текстів і проблема генерації природної мови.

**Список літератури**

1. Харламов А. А. Когнитивный подход к смысловому анализу текстов // М.: Вестник МГЛУ. – 2013. – Вып. №13 (673). – С. 196-205

2. Марченко О.О. Алгоритми семантичного аналізу природномовних текстів : автореф. дис. на здобуття наук. ступеня канд. фіз.-мат. наук : спец. 01.05.01 "Теоретичні основи інформатики та кібернетики" / Марченко Олександр Олександрович. – Київ, 2005. – 13 с.

3. Яцко В. А. Алгоритмы и программы автоматической обработки текста // Иркутск: Вестник ИГЛУ. – 2012. – Вып. №1 (17). – С.150-160.

4. Найханова Л.В. Основные типы семантических отношений между терминами предметной области // Известия ВУЗов. Поволжский регион. Технические науки. 2008. №1. – С.62-71.