УДК 519.766.2

МОДЕЛЬ АНАЛІЗУ НЕСТРУКТУРОВАНОЇ ІНФОРМАЦІЇ ДЛЯ ПОБУДОВИ БАЗИ ЗНАНЬ КОРПОРАТИВНОЇ ІНФОРМАЦІЙНОЇ СИСТЕМИ

Лисенко О.О.[1)], Кокітко Р.І. [2)]
Тернопільський національний економічний університет
[1)] старший викладач; [2)]магістрант

I. Постановка проблеми

В сучасних корпоративних інформаційних системах (КІС) з точки зору управління знаннями основний акцент робиться не на збережені різноманітної інформації, а на виявлені закономірностей та принципів, які дозволяють вирішувати бізнес задачі, тобто здійснювати накопичення знань. Джерелом знань є різноманітні документи, які поступають в систему на опрацювання. Основним завданням підвищення ефективності КІС стає розробка системи трансформації інформації, яка для обробки слабко структурованої та неструктурованої інформації буде використовувати конкретну модель знань предметної області роботи КІС [1].

II. Мета роботи

Метою дослідження є розробка семантично-орієнтованого лінгвістичного процесора для перетворення різноманітної електронної інформації в інтелектуальний актив компанії.

III. Опис моделі

Пропонована модель аналізу та перетворення неструктурованої інформації включає сім етапів:

- попередній лінгвістичний аналіз (На вхід поступають документи широко спектру форматів, а також виділяються документи, які в полях ієрархічних специфікацій включають списки ключових слів та та словосполучень документа);

- графемний аналіз (На вхід поступає множина текстів $T = \{t_1, t_2 ... t_n\}$ документів у формі повнотекстової бази даних. Задача графемного аналізу полягає у виділені текстових одиниць, які мають графемне значення та виділення прагматичних відношень між ними, які дозволяють класифікувати їх з деяким формалізованим класом);

- морфологічний аналіз (Основною функцією даного аналізу є перетворення словозмінних форм та представлення слова в канонічній формі);

- контекстний аналіз (З множини лексем тексту L формуються словосполучення. Для визначення словосполучення вибирається послідовність слів, які володіють еквівалентною морфологічної інформацією);

- статистичний аналіз (Система отримує інформаційне представлення кожного документа у вигляді ключових слів та словосполучень тексту $L_m = \{t_m^i\}, 1 \le i \le n$, UDK - тексту u_m та множини значень рубрикатора тексту $R_m = \{r_m^i\}, 1 \le i \le n$, де m - номер документа, який поступає в КІС);

- логіко-алгебраїчний аналіз навчальної вибірки (Використовуючи отримані на попередніх етапах інформаційні представлення кожного документу описуємо зв'язок предметної області діяльності менеджера, який працює з документами та предметних змінних, які об'єктивно визначають глибину представлених знань документа. Для цього будуємо бінарні предикати $P_l(l,m), P_r(r,m), P_u(u,m)$. Предикати представляємо у вигляді таблиць, в ячейках яких ставимо 0 або 1 в залежності від того чи рівний предикат 0 чи 1 для даних значень предметних змінних l, u, r, m);

- динамічна класифікація інформації (В результаті виконання усіх попередніх етапів кожному документу динамічно присвоюється його інформаційне представлення, яке порівнюється з еталонним представленням діяльності того або іншого менеджера).

Висновок

Використання описаного процесора дозволяє виділяти із неструктурованих потоків тексту на природній мові семантичну інформацію, а також формувати базу знань інформаційної системи.

Список використаних джерел

1. Тарловский В. А. Использование семантико-ориентированного лингвистического процессора для добывания новых знаний из потока документов корпоративной информационной системы / В.А. Тарловский, Н.Ф. Хайрова // Вестник Национального технического университета «ХПИ», – Харьков: НТУ «ХПИ», 2010.– № 67,– С. 132-138.