УДК 004.891:004.82 (045)

Вавіленкова А.І., д.т.н.
Національний авіаційний університет, Київ

МАТЕМАТИЧНИЙ АПАРАТ СИСТЕМИ ЛІНГВІСТИЧНОГО АНАЛІЗУ ТЕКСТОВИХ ДОКУМЕНТІВ

Дослідження методів інтелектуального аналізу даних за такими важливими критеріями для обробки природної мови, як швидкість, масштабованість, сфера використання, основні алгоритми, порядок слів, відстань між словами, використання тезаурусу, синтаксичні та семантичні зв'язки, дали змогу зробити висновок про те, що жоден із існуючих методів не забезпечує вирішення всього спектру задач змістовного аналізу текстової інформації. Тому було запропоновано новий математичний апарат, що дозволить здійснювати автоматичний лінгвістичний аналіз текстових документів:

- загальна форма логіко-лінгвістичної моделі речення природної мови;
- умови тотожності логіко-лінгвістичних моделей речень природної мови та способи їх виявлення;
- основні принципи та правила синтезу логіко-лінгвістичних моделей речень природної мови, що базуються на виявленні засобів змістовного зв'язку (семантичного та дейктичного повторення, вживання однакових граматичних форм, синтаксичної або транспозиційної деривації) у текстових документах;
- абстрактні моделі для формалізації опису логічних зв'язків між частинами текстових документів та їх геометричні інтерпретації, що дозволяють простежити тип розгортання думки в електронному текстовому документі;
- загальна форма логіко-лінгвістичної моделі електронного текстового документу, яка містить лінгвістичну та семантико-синтаксичну складову;
- алгоритм автоматичного формування логіко-лінгвістичних моделей електронних текстових документів.

Використання запропонованого математичного апарату системи лінгвістичного аналізу текстових документів можна застосовувати в методах порівняльного аналізу логіко-лінгвістичних моделей електронних текстових документів, який, на відміну від методу "шинглів", що використовується відкритими системами порівняльного аналізу, аналізуватиме текстові документи за змістом.