

ДОНБАСЬКА ДЕРЖАВНА МАШИНОБУДІВНА АКАДЕМІЯ

Богданова Надія Сергіївна

ДОСЛІДЖЕННЯ МЕТОДІВ, МОДЕЛЕЙ ТА ІНФОРМАЦІЙНИХ
ТЕХНОЛОГІЙ СИНТАКСИЧНОГО АНАЛІЗУ НАУКОВО-ТЕХНІЧНИХ
ТЕКСТІВ ДЛЯ ІНФОРМАЦІЙНОЇ ПІДТРИМКИ ДІЯЛЬНОСТІ
ІНЖЕНЕРІВ-КОНСТРУКТОРІВ

Спеціальність 8.05010102 - Інформаційні технології проектування

АВТОРЕФЕРАТ

На здобуття освітньо-кваліфікаційного рівня «магістр»

Краматорськ - 2014

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Сучасного інженера конструктора, як користувача науково-технічної інформації цікавить швидкий і якісний пошук даних по конкретній тематиці. Використовувані з цією метою програми морфологічного аналізу не дозволяють однозначно вирішити поставлене завдання швидкого пошуку, тому в російській мові часто можна міняти порядок слів у реченні, не змінюючи його сенс. Подібна обставина не дає можливість у реченні чітко виділити головні за змістом слова та встановити зв'язки між ними.

Один із способів вирішення цієї проблеми полягає у формалізації знань предметної області шляхом синтаксичного перетворення фрази в стандартизовані форми – триплети. Виконання такого аналізу вручну дуже важко, тому механізм перетворення фраз розвинений ще недостатньо. Тому тема дослідження є актуальною.

Мета і завдання дослідження.

Мета роботи. Автоматизувати синтаксичний аналіз науково-технічних текстів для інформаційної підтримки діяльності конструкторів.

Для досягнення поставленої мети необхідно вирішити такі завдання:

1. Виконати аналіз предметної області синтаксичний аналіз науково-технічних текстів.
2. Проаналізувати існуючі засоби синтаксичного аналізу текстів на природних мовах і їх придатність для автоматизації складання онтологій на основі науково-технічних текстів.
3. Розробити модель процесу синтаксичного аналізу науково-технічних текстів.
4. Розробити метод обробки тексту для інформаційної підтримки науково-технічних текстів стосовно діяльності інженера-конструктора. За участю фахівця з синтаксичного аналізу.

5. Спроектувати і реалізувати ПЗ для синтаксичного аналізу науково-технічних текстів і формування понятійної структури діяльності інженерів-конструкторів у форматі RDF.

6. З використанням розробленого ПЗ провести експеримент на основі даних отриманих для RDF триплетів (підмет і присудок, доповнення). Оцінити збіг триплетів початковому тексту. Оцінити навчання системи.

7. Розробити рекомендації з використання ПЗ для синтаксичного аналізу науково-технічних текстів, виявити економічний ефект.

Об'єкт дослідження. Процес перетворення науково-технічного тексту природною мовою в триплет (із застосуванням формату RDF).

Предмет дослідження. Алгоритми, методи, методики етапів формалізації знань, що містяться в науково-технічних текстах за допомогою синтаксичного аналізу.

Методи дослідження:

1 Використання методології та стандартів W3C в рамках концепції семантичної павутини.

2 Синтаксичний аналіз для виділення триплетів та подання їх у форматі RDF.

3 Морфологічний аналіз для виділення лексем (приведення слів до словникової форми).

4 Системні методи проектування та вищі мови програмування при розробці ПМК.

Наукова новизна роботи.

1. Створено модель синтаксичного аналізу науково-технічних текстів в основі, якої лежить виділення ключових слів і їх взаємозв'язків. Результати аналізу представляються у вигляді онтології із застосуванням rdf файлу.

2. Отримали подальший розвиток положення про те, що найменшою структурною одиницею rdf файлу може бути триплет.

3. На основі моделі синтаксичного аналізу науково-технічних текстів вперше отримані мінімальні оцінки кількості проаналізованих текстів, необхідних для задовільного пошуку інформації.

Практичне значення отриманих результатів.

Розроблений підхід, методики і програмне забезпечення можуть бути використані для підвищення ефективності праці аналітика по створенню онтології для науково-технічних текстів. Надалі розроблена онтологія може використовуватися для забезпечення релевантного пошуку в інтернет. Поставлена лабораторна робота з дисципліни «Інтелектуальний аналіз даних».

Особистий внесок здобувача.

На основі аналізу літературних джерел автор диплома з'ясувала, що вирішення проблеми якісного пошуку інформації складається з етапів морфологічного, синтаксичного і семантичного аналізу. Відшукала і підключила словники для проведення морфологічного аналізу. Вивчила можливості програми Protégé і з'ясувала, що вона підтримує формат файлу rdf, на основі якого можливе формування онтології. Створила концептуальну і логічну модель ПО, а потім реалізувала її програмно.

Публікації.

1. Богданова Н.С. Представление модели предметной области в виде иерархической ER-диаграммы. Міжнародна наукова конференція «Атомарні функції: теорія та застосування»АФ-2012. Матеріали конференції. Харків, 2013. – С. 129-132.

2. Акимов В.И., Богданова Н.С. Исследование методов, моделей и информационных технологий синтаксического анализа научно-технических текстов для информационной поддержки деятельности инженеров конструкторов. Міжнародна наукова методична конференція «Сучасні інформаційно-комунікаційні технології в науці та освіті. Тези доповідей. 10-11 грудня 2013 р., Харків, 2013. – С. 7.

3. Акимов В.И., Богданова Н.С., Винницкая Я.А. Децентрализованное хранение информации в распределенных компьютерных системах: обзор технологии DHT // Научный вестник ДГМА. – 2014.

Структура та обсяг роботи. Дипломна робота складається зі вступу, п'яти розділів, висновків, переліку використаних джерел із 73 найменувань, 28 малюнків, 72 таблиць та 5 додатків. Загальний обсяг дипломної роботи складає 147 сторінок, включаючи 122 сторінок основної частини і 25 сторінок додатків.

ЗАГАЛЬНІ ВИСНОВКИ

1. При автоматизації роботи з текстами природною мовою синтаксичний аналіз є обов'язковим етапом, що передуює семантичній інтерпретації.

2. Вивчення методів синтаксичного аналізу показало, що відсутній необхідний корпус для навчання існуючих синтаксичних аналізаторів російською мовою. Існує також проблема, коли в корпусі пошук за запитом може видавати сотні і навіть тисячі результатів, які просто фізично неможливо переглянути.

3. Показано, що на основі масиву науково-технічних текстів даної предметної області може бути генерований масив метаданих у форматі RDF. Його потім можна використовувати для побудови онтології, що забезпечить інформаційну підтримку діяльності конструктора.

4. З використанням теорії множин визначена математична модель – формальна модель, для задіяння її в розробленому проекті. Принцип використання формальної моделі полягає в розбитті науково-технічного тексту на кінцеву множину речень і слів, облік відносин між словами і формування кінцевої множини триплетів в rdf файл.

5. Розглянуто методику бізнес-процесу «Побудова триплета» і виявлено, що він повинен складатися як мінімум з сегментації, токенизації,

морфологічного аналізу, синтаксичного аналізу. Бо саме ці етапи забезпечують перетворення тексту до формату RDF.

6. Запропоновано декілька варіантів оцінки триплетів, отриманих при аналізі науково-технічного тексту: аналіз залежності кількості оброблених текстів від кількості записаних триплетів, кількість оброблених текстів від кількості лексем.

7. На основі досліджень, проведених в даній роботі і розроблених математичної та інформаційної моделі відповідно до прийнятої методики, спроектовано програмне забезпечення, що дозволяє автоматизувати синтаксичний аналіз науково-технічних текстів для інформаційної підтримки діяльності конструкторів