

НОВИЙ ПІДХІД ДО ФОРМАЛІЗАЦІЇ ТА ОБРОБКИ ЕЛЕКТРОННИХ ЕКОЛОГІЧНИХ РЕСУРСІВ

Запропоновано новий підхід до формалізації та обробки електронних екологічних ресурсів, який дозволяє вирішити задачу видобування знань із текстів природною мовою та побудови на основі видобутих даних онтологічних баз знань, що дозволить більш ефективно здійснювати пошук та обробку інформації про геосемантичні характеристики різних просторових об'єктів.

Предложен новый подход к формализации и обработке электронных экологических ресурсов, позволяющий решить задачу извлечения знаний из текстов на естественном языке и построения на основе добытых данных онтологических баз знаний, что позволит более эффективно осуществлять поиск и обработку информации о геосемантических характеристиках разных пространственных объектов.

Проблемі видобування знань присвячено безліч зарубіжних праць, що об'єднуються в єдиний клас задач добування інформації з текстів (text mining). Інформація, що видобувається із природномовних текстів, подається у вигляді структур даних, поля яких заповнюються текстовими фрагментами (цитатами) [1, 2]. Недоліком зарубіжних розробок є сильна залежність від конкретної граматики мови. Серед російськомовних розробок відомі тільки дві системи компаній RCO і Yandex, які мають вкрай обмежене застосування, оскільки не існує простого способу їх адаптації до довільної предметної області [3–5]. Тому актуальним є задача автоматизованої обробки текстів природною мовою та побудова онтологічної бази знань на основі видобутих даних. Отож, необхідно розробити підхід до формалізації та обробки електронних екологічних ресурсів, оснований на видобуванні знань із природномовних текстів та автоматизованій побудові бази онтологій, що дозволить більш ефективно здійснювати пошук та обробку інформації про геосемантичні характеристики різних просторових об'єктів.

Як відомо, до джерел інформації, що використовуються як основа для побудови онтологічної бази знань, висуваються наступні вимоги: достовірність, несуперечність, точність формулювань. Крім того, до інформації, представленої низкою документів, додатково висувається вимога щодо усталеної структури. Такими характеристиками володіють щорічні Національні доповіді Міністерства охорони навколишнього природного середовища України про стан довкілля. Згідно Наказу від 11.06.2013 №258 «Щодо підготовки Національної доповіді про стан навколишнього природного середовища в Україні у 2012 році» для Національних доповідей устанавлюється визначена структура документу [6].

Для вирішення поставленої задачі необхідно вирішити наступні підзадачі:

1. Розробити модель видобування знань із текстів українською мовою з урахуванням її лінгвістичних особливостей.

2. На основі попередньої моделі розробити модель видобування знань про об'єкти із геосемантикою з екологічних звітів державних органів влади українською мовою.

3. Розробити інформаційну технологію автоматизованої обробки веб-документів та формування бази онтологій із урахуванням можливостей пошуку.

Автори пропонують створити новий комплекс засобів для автоматизованої обробки текстів природною мовою та формування бази онтологій, який включає наступні компоненти:

1. Модуль обробки текстів природною мовою, який виділяє із тексту набір речень згідно заздалегідь розробленого шаблону та виконує семантичний аналіз таких речень.

2. Модуль автоматизованого або автоматичного виділення геосемантичних характеристик об'єктів із проаналізованих речень та побудови онтологічної бази даних на основі видобутих характеристик.

Технологія може бути застосована для швидкої автоматизованої прив'язки документів до карт ГІС для створення географічно орієнтованих каталогів електронних бібліотек [7].

1. Kao, A., and Poteet, S. Natural Language Processing and Text Mining. / A. Kao, S. Poteet — Springer, 2006. — 277 с.

2. Ian H. Witten. Text mining // [Електронне джерело] — Режим доступу до статті: http://www.cos.ufrj.br/~rick/gc2010/_papers/aula13/04-IHW-Textmining.pdf

3. Симаков К. В. Модели и методы извлечения знаний из текстов на естественном языке: Автореферат дис. канд. техн. наук / К. В. Симаков; Московский государственный технический университет имени Н. Э. Баумана – М., 2008 — 16 с. — рус.

4. Марченко О. О., Анісімов А. В., Никоненко А. О. Алгоритмічна модель асоціативно-семантичного контекстного аналізу текстів природною мовою / А. В. Анісімов, О. О. Марченко, А. О. Никоненко — Проблеми програмування, 2008, № 2-3. — С. 379-384

5. Валенда А. Н. Методи та моделі функціонально-семантичної обробки текстів природної мови у системах штучного інтелекту: Автореф. дис. канд. техн. наук / Н. А. Валенда; Харківський національний університет радіоелектроніки — Х., 2006. — 19 с. — укр.

6. Наказ від 11.06.2013 № 258 «Щодо підготовки Національної доповіді про стан навколишнього природного середовища в Україні у 2012 році» // [Електронне джерело] — Режим доступу до статті:

<http://www.menr.gov.ua/index.php/normakty/60-9/acts/521-nakaz-vid-11-06-2013-258-shchodo-pidhotovky-natsionalnoi-dopovidi-pro-stan-navkolyshnoho-pryrodnoho-seredovyshcha-v-ukraini-u-2012-rotsi>

7. Мокін В. Б. Електронна екологічна бібліотека: нові підходи, технології та можливості / Мокін В. Б. // [Наукові праці Вінницького національного технічного університету. Електронне видання]. — 2009.— №3. — Режим доступу до журн.: http://nbuv.gov.ua/e-journals/VNTU/2009-3/2009-3.files/uk/09vbmtap_ua.pdf