УДК 004.912

АВТОМАТИЗОВАНА ОБРОБКА ВЕБ-ДОКУМЕНТІВ ІЗ ФОРМУВАННЯМ МАТЕРІАЛІВ ІЗ ПРИВ'ЯЗКОЮ ДО ОБ'ЄКТІВ ІНТЕРНЕТ-ГІС

Мокін В. Б., Богомолов Ю. С.

Вінницький Національний Технічний Університет, Україна

Анотація

Існує необхідність добувати спеціалізовані знання щодо геосемантичних характеристик об'єктів з текстів на природній мові. Отримані характеристики доцільно прив'язати до об'єктів Інтернет-ГІС заради спрощення доступу до цієї інформації як науковців, так і студентів.

There is a necessity to elicit specialized knowledge as for geosemantic features of objects from the natural language texts. The given characteristics are appropriate to be connected to the Internet-GIS objects in order to simplify access to this information for both researchers and students.

Вступ

Існує необхідність добувати спеціалізовані знання щодо геосемантичних характеристик об'єктів із природномовних текстів. Отримані характеристики доцільно прив'язати до об'єктів популярних Інтернет-ГІС заради спрощення доступу до цієї інформації як в наукових, так і в освітніх цілях. 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 закінчені системи компаній RCO і Yandex, які мають вкрай обмежене застосування, оскільки не існує простого способу їх адаптації до довільної предметної області [3–5]. Отож, задача дослідження формулюється наступним чином: необхідно розробити методи видобування знань із природномовних текстів (наприклад, із щорічних Національних доповідей Міністерства охорони навколишнього природного середовища України про стан довкілля) та наносити отримані у результаті аналізу видобутих знань характеристики об'єктів на карту або логічно приєднувати до об'єктів карти. 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 семантично розмічених матеріалів для популярних Інтернет-ГІС.

Автори пропонують створення нового комплексу засобів для автоматизованої обробки веб-документів та формування матеріалів із прив'язкою до Інтернет-ГІС, який включає наступні компоненти:

1. Онтологічна база даних із предметної галузі «Екологія та природокористування».
2. Модуль обробки текстів природною мовою, який виділяє із тексту набір речень, що містять терміни із онтологічної бази даних, а також виконує семантичний аналіз таких речень.
3. Модуль автоматизованого або автоматичного виділення геосемантичних характеристик об'єктів із проаналізованих речень та співвіднесення цих характеристик із об'єктами карт Інтернет-ГІС.

Швидкість роботи комплексу засобів забезпечується особливостями типового представлення екологічної інформації та її прив'язки до карт. Технологія може бути застосована і для швидкої автоматизованої прив'язки документів до карт ГІС для створення географічно орієнтованих каталогів електронних бібліотек [6].

Список використаних джерел:

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. 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