

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА КОМП'ЮТЕРНА ТЕХНІКА

УДК 681.51

В. Б. Мокін, д. т. н., проф.;

Ю. М. Коновалюк, асп.

НОВИЙ МЕТОД ПОШУКУ РІЗНОФОРМАТНОЇ ЕКОЛОГІЧНОЇ ІНФОРМАЦІЇ НА ОСНОВІ ОНТОЛОГІЧНОЇ БАЗИ ДАНИХ ТА ЇЇ XML-ПРЕДСТАВЛЕННЯ

Розроблено теоретичні та практичні основи нового методу пошуку різноформатної екологічної інформації на основі онтологічної бази даних та її XML-представлення, який дозволяє здійснювати комплексний пошук по всіх основних типах джерел (карти геоінформаційних систем, бази даних, текст) одночасно, не накладає суттєвих обмежень на формати джерел і враховує просторові відношення між об'єктами під час пошуку їх числових характеристик та текстового опису. Розроблено та успішно апробовано програмне забезпечення для реалізації методу.

Постановка задачі

Останнім часом накопичилось чимало різного роду екологічної інформації, тобто даних про об'єкти довкілля та пов'язані з ними процеси – це і дані про якість та кількість природних ресурсів, і відомості про забруднювачів навколишнього природного середовища, і документи законодавчого характеру, що регламентують діяльність у сфері охорони довкілля та управління і моніторингу природних ресурсів [1].

Екологічна інформація буває, як правило, чотирьох основних видів форматів: карти геоінформаційних систем (просторова інформація про розташування природних об'єктів), бази даних (атрибутивні дані – кількісні та якісні характеристики про стан природних об'єктів), текстові документи (інформація, яка стосується природних об'єктів – описи їх стану, законодавчі акти тощо), графічні файли (фотографії природних об'єктів). Існуючі засоби пошуку інформації дозволяють здійснювати ефективний пошук в документах окремих типів. Існують також розвинуті системи, які дозволяють здійснювати пошук в документах різних типів, але в таких системах, як правило, відсутня підтримка пошуку в картах геоінформаційних систем (ГІС). Ті системи, які ж все таки дозволяють здійснювати і пошук у базах даних, і пошук у картах ГІС (просторово-орієнтований пошук), найвідомішою з яких є система Oracle версії 11, мають суттєве обмеження — уся інформація (текст, бази даних, ГІС) повинна бути спочатку інтегрована в одну систему єдиного специфічного формату. Отже, актуальною є задача створення теоретичних та практичних основ швидкого пошуку екологічної інформації, яка зберігається у різних форматах традиційного типу – текст (doc, htm), електронні таблиці (xls), бази даних (mdb), карти геоінформаційних систем (map, sit, shr тощо). При цьому модель даних пошуку повинна забезпечувати режим асоціативного пошуку, який би ґрунтувався на взаємозв'язках між різними типами об'єктів на карті ГІС. Це дало б значно більші можливості пошуку. Наприклад, це дозволило б знайти усі водойми в радіусі 5 км від міста і усю інформацію про них, якість їх вод, постанови щодо риболовлі або знайти усі населені пункти (чи будинки – в залежності від деталізації карти) в одновідсотковій зоні ймовірного затоплення території біля річки Тиса та усі розпорядження і постанови щодо норм будівництва у цій зоні, а також усі відомості про знайдені населені пункти чи будинки тощо.

Розробка онтологічної бази даних

Для розробки методу пошуку просторово-орієнтованої екологічної інформації потрібно формалізувати шукану інформацію для кожного типу джерел, виходячи з того, що основним джерелом для пошуку є карти ГІС. З точки зору пошуку карта ГІС являє собою множину просторових об'єктів (G), текстовий документ – множину ключових слів (D), база даних – множину значень полів усіх записів усіх таблиць (B). Нехай, маємо множини карт ГІС (G^*), текстових документів (D^*), баз даних (B^*), а також множину слів пошукового запиту ($Q = \{Q_1, Q_2, \dots, Q_n\}$). Тоді результатом S комплексного пошуку буде

$$S = \bigcap_{i=1}^n R(G^*, Q_i) \cup \bigcap_{i=1}^n R(D^*, Q_i) \cup \bigcap_{i=1}^n R(B^*, Q_i),$$

де $R(A, B)$ – функція пошуку, яка знаходить місця знаходження B у множині даних A .

Для формалізації та алгоритмізації такої функції пошуку потрібно уніфікувати механізм обробки запитів для різних типів джерел інформації. Розглянемо зв'язок між документами різних типів з точки зору пошуку просторово-орієнтованої інформації (рис. 1).



Рис. 1. Модель взаємозв'язків джерел даних (- - -> конкретизація)

Виходячи з того, що пошук орієнтований на карти ГІС, основою наведеної схеми є множина об'єктів карт ГІС. Кожен з цих об'єктів (місто, річка, ліс тощо) може характеризуватись одним або більше ключовими словами, які зустрічаються в текстових документах та текстових полях баз даних. Крім того, ці об'єкти можуть характеризуватись певними полями таблиць баз даних (довжина, площа басей-

ну річок; площа, кількість населення у містах тощо). Пошук цієї атрибутивної інформації здійснюється шляхом виконання SQL-запитів.

Основним підходом до пошуку є використання ключових слів [2]. Деяке удосконалення цього механізму дозволяє використовувати його для пошуку просторово-орієнтованої інформації. Це вдосконалення полягає у зв'язуванні ключових слів з об'єктами карт ГІС. Таким чином було розроблено об'єктно-орієнтовану структуру онтологічної бази даних [2], яка пов'яже об'єкти карт ГІС з ключовими словами, запитами до баз даних, а через них і, власне, з текстовими документами та інформацією у базах даних. При цьому використане XML-представлення інформації, яке найбільше підходить для представлення структурованих у запропонований спосіб даних (рис. 2).

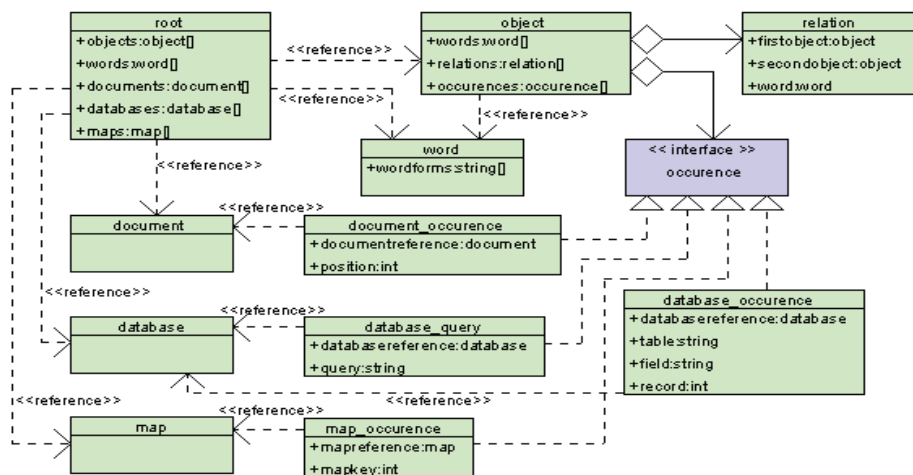


Рис. 2. Модель XML-представлення онтологічної бази даних

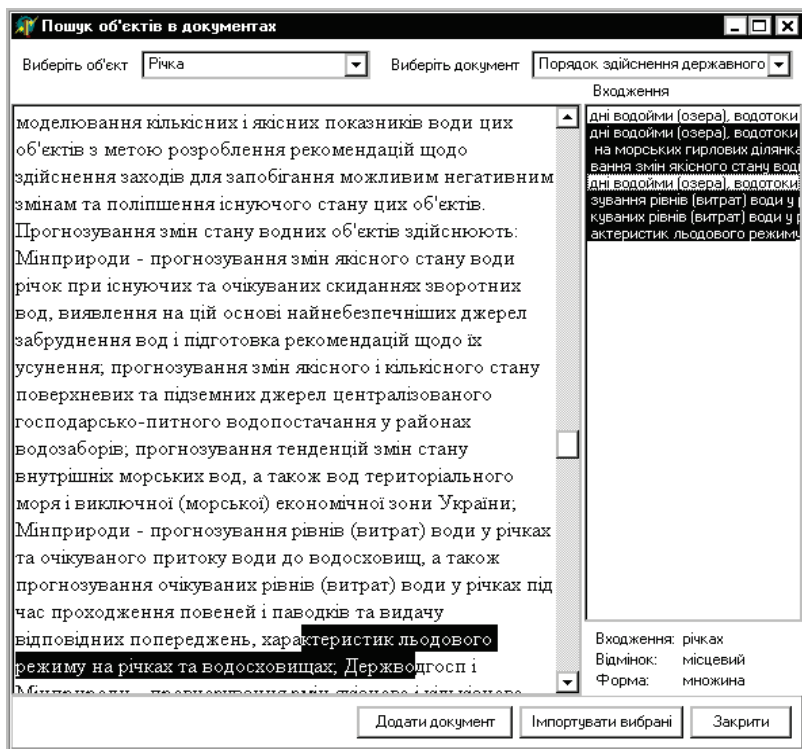


Рис. 5. Індексція текстових документів

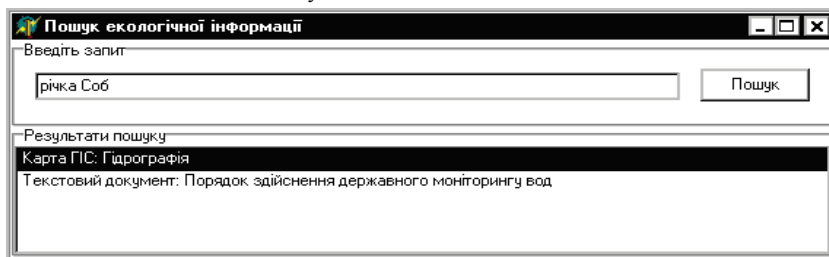


Рис. 6. Програма пошуку різноформатної екологічної інформації

Програма для наповнення онтологічної бази даних дозволяє автоматизовано імпортувати об'єкти з класифікатора ГІС (рис. 4). Під час імпорту є можливість додати нові ключові слова до онтологічної бази даних і одразу ж пов'язати їх з імпортованими об'єктами.

Також програма дозволяє автоматизовано індексувати текстові документи та бази даних (рис. 5).

Програма для пошуку різноформатної екологічної інформації має типовий для пошукових сервісів інтерфейс. Результат пошуку на прикладі ГІС «Панорама» (РФ, <http://www.gisinfo.ru>), відкритий в типовій програмі-оболонці, подано на рис. 6.

Висновки

Розроблено теоретичні та практичні основи нового методу пошуку різноформатної екологічної інформації на основі онтологічної бази даних та її XML-представлення, який на відміну від існуючих дозволяє здійснювати комплексний пошук по всіх основних типах джерел (карти геоінформаційних систем, бази даних, текст) одночасно, не накладає суттєвих обмежень на формати джерел і враховує просторові відношення між об'єктами під час пошуку їх числових характеристик та текстового опису. Створено та успішно випробувано спеціалізоване програмне забезпечення для наповнення онтологічної бази даних та пошуку екологічної інформації в джерелах різних форматів на прикладі ГІС «Панорама» (РФ, <http://www.gisinfo.ru>).

СПИСОК ЛІТЕРАТУРИ

1. Комп'ютеризовані регіональні системи державного моніторингу поверхневих вод: моделі алгоритми програми : [монографія] / Мокін В. Б., Боцула М. П., Горячев Г. В. та ін.; під ред. В. Б. Мокіна. — Вінниця: УНІВЕРСУМ–Вінниця, 2005. — 315 с. — ISBN 966-641-132-6.
2. Организация эффективного поиска на основе онтологий [Электронный ресурс] / О. И. Россеева, Ю. А. Загорюлько. — Режим доступа: http://www.dialog-21.ru/Archive/2001/volume2/2_49.htm

Рекомендована кафедрою моделювання та моніторингу складних систем

Надійшла до редакції 21.10.08
Рекомендована до друку 20.11.08

Мокін Віталій Борисович — завідувач кафедри, **Коновалюк Юрій Михайлович** — аспірант.

Кафедра моделювання та моніторингу складних систем Вінницького національного технічного університету