УДК 681.51УДК 004.65+519.6

В. Б. Мокін, д-р. техн. наук, проф.;
Ю. М. Коновалюк, асп.

РОЗРОБКА МОДЕЛЕЙ ВХІДНИХ ДАНИХ ДЛЯ ІТЕРАТИВНОГО МЕТОДУ ПОШУКУ РІЗНОФОРМАТНОЇ ЕКОЛОГІЧНОЇ ІНФОРМАЦІЇ

Запропоновано нові моделі документів основних форматів ГІС, пошукових запитів та онтології з використанням апарату контекстно-вільних граматик. Використання цих моделей дозволяє алгоритмізувати та оптимізувати ітеративний метод пошуку різноформатної екологічної інформації з урахуванням семантичних та просторових відношень між параметрами об'єктів ГІС.

1. Постановка задачі

Пошук інформації є одним із ключових аспектів роботи з даними в електронній формі. Основними форматами, в яких зберігається інформація в геоінформаційних системах (ГІС), є: карти (просторова інформація), бази даних (детальні атрибутивні дані), текстові документи (будь-яка інформація, яка стосується природних об'єктів — опис їх стану, атрибутивні дані у вигляді таблиць, законодавчі акти тощо). Є багато відомих систем пошуку інформації для кожного вказаного типу джерел: для карт ГІС це — засоби просторового пошуку в пакетах програм ArcGIS, Mapinfo, ГІС «Панорама», Digitals тощо; для текстових документів це — пошук точних входжень, пошук за шаблонами, пошук за регулярними виразами, а також ефективні їх реалізації для пошуку інформації в мережі Інтернет; для баз даних це — мова конструювання запитів SQL та ін. Також існують комбіновані рішення для різноформатного пошуку, такі як Oracle Spatial. Однак, проблема багатоформатного пошуку в загальному випадку не розв'язана.

Поставлено та розв'язано важливу задачу ефективного пошуку різноформатної (ГІС, бази даних, текст) екологічної інформації з урахуванням взаємозв'язків між різними типами природних об'єктів. Зокрема, розроблено новий метод пошуку різноформатної екологічної інформації [1]. Удосконалення цього методу шляхом використання нових моделей вхідних даних, пошукових запитів і результатів пошуку дозволить здійснювати пошук за ітеративним підходом. Для цього, перш за все, слід провести формалізацію даних та алгоритмів їх обробки: формування проміжних результатів, синтезу нових запитів пошуку та визначення кінцевих результатів пошуку.

Таким чином, можна сформулювати задачу у такому вигляді: необхідно розробити нові інформаційні моделі вхідних даних, використовуючи один з відомих апаратів формалізації та нотації, що дозволить формалізувати, алгоритмізувати та, в подальшому, запрограмувати ітеративний різноформатний пошук екологічної інформації в документах основних форматів ГІС.

2. Вибір апарату формалізації та розробка моделей документів основних форматів

В загальному випадку вхідні дані S можна представити у вигляді множини документів всіх основних форматів:

$$S = \left[\{M\},\{B\},\{T\}\right],$$

де $\{M\}$ — множина карт ГІС; $\{B\}$ — множина баз даних, $\{T\}$ — множина тестових документів, які є основними форматами, в яких зберігається екологічна інформація.

Основою будь-якої карти M у більшості пакетів програм для роботи з ГІС (ГІС-пакетів) є класифікатор. Класифікатор L містить множину об'єктів $\{J\}$, які можуть бути відображені на карті і множину шарів $\{A\}$, які характеризуються назвою N_A і кодом I_A. Також класифікатор містить множину семантик $\{E\}$, які можуть бути притаманні об'єктам, і множину семантик $\{E_J\}$, які властиві кожному конкретному типу об'єктів із множини $\{J\}$. Кожна семантика, в свою чергу, характеризується кодом I_E, назвою N_E, типом значень T_E, які вона може містити, і розміром Z_E. Кожен об'єкт J характеризується назвою N_J, множиною семантик $\{E_J\}$, шаром A_J, до якого він належить, кодом I_J, типом об'єкта T_J (точковий, лінійний, площинний, тощо), а також позначенням V_J.

$$L = \left[\{A\},\{E\},\{J\}\right];$$

$$A = \left[I_A,\ N_A\right];$$

$$E = \left[I_E,\ N_E,\ T_E,\ Z_E\right];$$

$$J = \left[I_J,\ N_J,\ \{E_J\},\ A_J,\ T_J,\ V_J\right].$$

Кожна карта характеризується ім'ям файла N_M, класифікатором L, множиною примірників об'єктів $\{O\}$ і типом координатної системи T_C.

$$M = \left[N_M,\ L,\ \{O\},\ T_C\right].$$

У свою чергу, кожен примірник об'єкта характеризується ключем (унікальним кодом в межах окремої карти) I_O, типом об'єкта з класифікатора J_O, множиною семантик $\{E_O\}$, значення яких фактично наявні в даному об'єкті, і метрикою M_O, яка є множиною точок $\{P\}$, кожна з яких задана координатами X_P і Y_P (або і Z_p — координата висоти на рівнем моря, м). Кожна з семантик, у свою чергу, характеризується посиланням на опис семантики у класифікаторі E_L і фактичним значенням семантики C_E.

$$O = \left[I_O,\ J_O,\ \{E_O\},\ M_O\right];$$

$$M_O = \{P_M\};$$

$$P_M = \left[X_P,\ Y_P\right];$$

$$E_O = \left[E_L,\ C_E\right].$$

Кожний текстовий документ характеризується ім'ям файла N_T і його вмістом C_T:

$$T = \left[N_T,\ C_T\right].$$

З точки зору доступу бази даних можуть характеризуватись ім'ям файла або назвою БД в серверній СУБД. В будь-якому випадку ця характеристика є текстовим рядком і доступ до бази даних для обох випадків здійснюється однаково, тому розрізняти їх не доцільно. Називатимемо цю характеристику рядком доступу N_B. Також кожна база даних характеризується множиною таблиць $\{T_B\}$. Таким чином,

$$B = \left[N_B,\{T_B\}\right].$$

У свою чергу, кожна таблиця має назву N_{TB} та складається з множини полів $\{F\}$ і множини записів $\{R\}$:

$$T_B = \left[N_{TB},\ \{F\},\ \{R\}\right].$$

Поля характеризуються назвою N_F, типом даних T_F і коментарем M_F:

$$F = \left[N_F,\ T_F,\ M_F\right].$$

Кожен запис множини записів має порядковий номер і є множиною значень полів таблиці.

$$R = \left[I_R,\ \{C_{NF}\}\right],$$

де C_{NF} — це значення поля з назвою N_F у записі R.

Вміст текстового документа формалізується послідовністю слів, чисел, знаків пунктуації, а також інших об'єктів, таких як зображення, графіки, формули тощо. Позначимо нетекстові елементи текстового документу через ζ. Для формалізації інформаційних елементів файла та контекстних зв'язків між ними пропонуємо використати контекстно-вільну граматику [2, 3], оскільки вона дозволить формалізувати входження ключових слів у документ з урахуванням зв'язків між об'єктами (словами).

Тоді текстовий документ може бути формалізований контекстно-вільною граматикою (на прикладі файлів українською мовою) [2, 3]

$C_T \rightarrow$ text,

text → text text_element|text_element, text_element → K^W|word|number|charset|delimiter|ζ|ε;

word → word-word| word'word|word alpha|alpha, number → digitset|digitset decimal__delimiter digitset;

digitset → digitset digit|digit, charset → charset character|character;

character → digit|alpha, digit → 0|1|2|3|4|5|6|7|8|9, alpha → a|б|..|я|ь|A|Б|..|Я|Ь, decimal_- _delimiter → .| .

В цій граматиці термінальним символом «delimiter» позначено будь-який символ, крім алфавітно-цифрових. А також, згідно з загальноприйнятою практикою [2], термінальним символом ε позначено порожній рядок. Символом K^W в граматиці позначено ключові слова. Для інших мов просто розширяється нетермінал alpha.

Таким чином, модель вхідних даних можна записати так:

$$S = \left[\begin{array}{l} \left\{ \left[\begin{array}{l} N_M, \left[\left\{ [I_A,\ N_A] \right\}, \left\{ [I_E,\ N_E,\ T_E,\ Z_E] \right\}, \left\{ \left[I_J,\ N_J,\ \{E_J\},\ A_J,\ T_J,\ V_J \right] \right\} \right], \\ \left\{ \left[I_O,\ J_O,\ \left\{ [E_L,\ C_E] \right\}, \left\{ [X_P,\ Y_P] \right\} \right] \right\}, T_C \end{array} \right] \right\}, \\ \left\{ \left[N_B, \left\{ \left[N_{TB}, \left\{ [N_F,\ T_F,\ M_F] \right\}, \left\{ \left[I_R, \{C_{NF}\} \right] \right\} \right] \right\} \right] \right\}, \{N_T, C_T\} \end{array} \right].$$

3. Розробка нової моделі онтології

В [1] для пошуку різноформатної екологічної інформації запропоновано використовувати онтології. Побудуємо модель такої онтології на основі моделі вхідних даних основних типів документів.

Онтологія D формалізується у вигляді множини об'єктів $\{O^D\}$ та зв'язків між ними $\{R^D\}$.

$$D = \left[\left\{ O^D \right\}, \left\{ R^D \right\} \right].$$

Кожний об'єкт онтології містить назву N_{OD} та інформацію про характер входження даних про цей тип об'єктів в документи усіх типів:

– для карт ГІС — це код об'єкта в картах I_J і його тип T_J;

– для баз даних — це множина таблиць $\{T_B\}$, в яких містяться атрибутивні дані про об'єкти цього типу;

– для текстових документів — це множина ключових слів $\{K^W\}$.

Таким чином, можна записати:

$$O^D = \left[N_{OD},\ I_J,\ T_J,\ \{T_B\},\ \left\{ K^W \right\} \right].$$

В свою чергу, кожний зв'язок R^D між об'єктами O^{DM} і O^{DR} характеризується ключовим словом K^W для ідентифікації такого зв'язку в текстових документах і парами полів $[T_B,\ F]$ таблиць бази даних для ідентифікації зв'язку в БД.

$$R^D = \left[O^{DM} \overset{K^W, \{[T_{BDM}, F_{DM}], [T_{BDR}, F_{DR}]\}}{\Rightarrow} O^{DR} \right],$$

де $[[T_{BDM},\ F_{DM}],\ [T_{BDR},\ F_{DR}]]$ — пара полів (F_{DM} і F_{DR}), по яких зв'язуються таблиці T_{BDM} і T_{BDR}, що відповідають об'єктам O^{DM} і O^{DR}, відповідно.

Таким чином, модель онтології можна записати так:

$$D = \left[\left\{ \left[N_{OD}, I_J, T_J, \{T_B\}, \left\{ K^W \right\} \right] \right\}, \left\{ \left[O^{DM} \overset{K^W, \{[T_{BDM}, F_{DM}], [T_{BDR}, F_{DR}]\}}{\Rightarrow} O^{DR} \right] \right\} \right]. \qquad (1)$$

4. Розробка моделі пошукових запитів

Модель онтології (1), завдяки присвоєнню кожному об'єкту і зв'язку між об'єктами ключових слів, дозволяє здійснювати пошук в документах усіх основних форматів одночасно, за допомогою

текстових пошукових запитів. Будь-який текстовий пошуковий запит може бути формалізований такою граматикою:

Q → Q query_element|query_element,

query_element → $K^W|N^W|\varepsilon$, N^W → charset,

charset → charset character|character, character → digit|alpha,

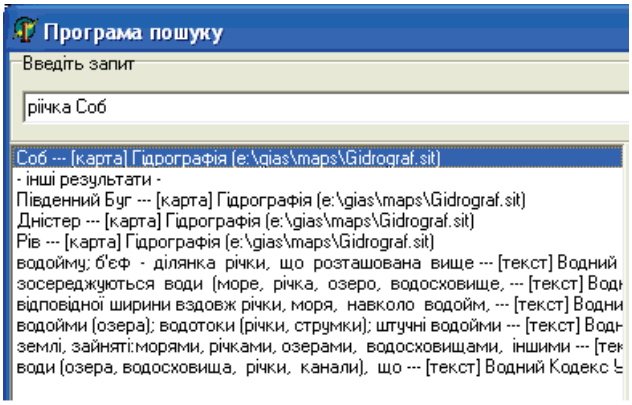
digit → 0|1|2|3|4|5|6|7|8|9, alpha → а|б|..|я|ь|А|Б|..|Я|Ь, decimal_delimiter → .|,

де N^W є будь-якою послідовністю символів, що не є ключовим словом. Таким чином, модель запиту можна спростити:

$$Q = \left[\left\{K^W\right\}, \left\{N^W\right\}\right].$$

5. Приклад пошуку

Для демонстрації пошуку застосуємо пошуковий запит «річка Соб» (рис.) до ГІС державного моніторингу довкілля Вінницької області та банку даних Держуправління охорони навколишнього природного середовища у Вінницькій області. Після здійснення пошуку отримано результати, що були знайдені на картах ГІС та в текстових документах. Для ітеративного пошуку можна автоматизовано ітеративно згенерувати запит на пошук водосховищ чи ставків на р. Соб згідно з визначенням ставка у чинному Водному Кодексі України та інформації банку даних водного кадастру Басейнового управління водними ресурсами річки Південний Буг Держводгоспу.



Приклад пошуку

Висновки

Запропоновано нові моделі документів основних форматів ГІС, пошукових запитів та онтології з використанням апарату контекстно-вільних граматик. Використання цих моделей дозволяє алгоритмізувати та оптимізувати ітеративний метод пошуку різноформатної екологічної інформації з урахуванням семантичних та просторових відношень між параметрами об'єктів ГІС.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Мокін В. Б. Новий метод пошуку різноформатної екологічної інформації на основі онтологічної бази даних та її XML-представлення / В. Б. Мокін, Ю. М. Коновалюк // Вісник Вінницького політехнічного інституту. — 2009. — № 2. — С. 66—69.

2. Alfred V. Aho (Author) The Theory of Parsing, Translation, and Compiling (Volume I: Parsing) / Alfred V. Aho, Jeffrey D. Ullman — Prentice Hall, 1972. — 542 p.

3. Stefano Crespi Reghizzi. Formal Languages and Compilation. — Springer, 2009. — 368 p.

Рекомендована кафедрою моделювання та моніторингу складних систем

Стаття надійшла до редакції 25.02.11
Рекомендована до друку 17.03.11

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

Кафедра моделювання та моніторингу складних систем, Вінницький національний технічний університет, Вінниця