Предметно-орієнтована мова аналітичного опрацювання даних

Андрій Болдак[1], Костянтин Єфремов[2]

Світовий центр даних з геоінформатики та сталого розвитку,
Національний технічний університет України «Київський політехнічний інститут»,
УКРАЇНА, 03056, м.Київ, пр. Перемоги 37, E-mail:
1. boldak@wdc.org.ua, 2. k.yefremov@wdc.org.ua

Abstract syntax, syntactic rules and transformation of abstract submission rules into that, which is performed by tool environment for declarative domain-specific language for data mining has been developed. Using of language that is proposed doesn't require special knowledge or programming skills that makes a process of scenario development close to experts.

Ключові слова – предметно-орієнтована мова, аналітичне опрацювання даних, статистична обробка даних, програмування, Світовий центр даних.

Однією з причин низької ефективності використання інструментарію аналітичного опрацювання даних в системах прийняття рішень є те, що, з одного боку, розробка нових та(або) модифікація існуючих сценаріїв обробки даних здійснюється засобами інструментальної мови з використанням імперативної та об'єктно-орієнтованої парадигми програмування. З іншого боку, експерти, які повинні визначати сутність аналітичного опрацювання даних, а, відповідно, і розробляти сценарії, не мають належного досвіду програмування та вимушені звертатися за допомогою до програмістів, що, в свою чергу, збільшує трудомісткість створення та модифікації сценаріїв.

З огляду на зазначене вище, мета роботи полягає у зниженні трудомісткості створення та модифікації сценаріїв аналітичного опрацювання даних за рахунок розробки, реалізації та використання предметно-орієнтованої мови, яка не потребує від експерта спеціальних знань та навичок програмування і достатня для опису процесу перетворення та аналізу даних.

На основі моделі обчислювального процесу, який розглядається як поглинаючий ланцюг Маркова з дискретними станами та автоматним часом [1], розроблено абстрактний синтаксис предметно-орієнтованої мови. Таку модель може бути описано за допомогою як послідовності подій, так і мережі взаємодіючих процесів. Використання обмеженого алфавіту подій з чітко визначеною в межах протоколу взаємодії процесів семантикою забезпечує підходу до опису сценаріїв, орієнтованого на процеси, переваги, пов'язані з тим, що така модель акцентує увагу розробника сценарію на структурі перетворень даних, які необхідно здійснити.

Зв'язок за даними між процесами здійснюється за допомогою потоків, які містять черги пакетів, що інкапсулюють як самі дані, так і метадані, необхідні для реалізації механізмів інтроспекції. Оскільки операції вибірки та поміщення пакетів в чергу ідентифікуються відносно процесів, потоки можуть подаватися у неявній формі, за допомогою поняття вхідних та

вихідних портів, які асоціюються з процесами. Таким чином, процеси з'єднуються в мережу за допомогою портів, а кожна пара «вихідний порт – вхідний порт» асоціюється з потоком пакетів. У відповідності до наявності вхідних і вихідних портів процеси можуть відноситись до трьох категорій: процеси-джерела даних, процеси-звіти, процеси-перетворення. Оскільки процеси отримують будь-які дані лише в спосіб, що передбачає вибірку пакетів з вхідних портів, виникає потреба у визначенні особливого типу процесів-джерел з постійним вихідним потоком пакетів – процесів-констант, які слугують для налаштування інших процесів.

Розроблено синтаксичні та графічні форми опису сценарію аналітичного опрацювання даних з використанням декларативної та імперативної парадигми програмування, в основі яких лежить модель графу, в якому вершини відповідають процесам, а дуги задають зв'язки між вихідними та вхідними портами.

Предметно-орієнтована мова аналітичного опрацювання даних та інструментальні засоби її реалізації впроваджуються у Світовому центрі даних з геоінформатики та сталого розвитку [2] і використовуються для розрахунку моделей оцінювання сталого розвитку в глобальному та регіональному контексті [3]. На основі бібліотеки MXGraph розроблено спеціалізований графічний редактор з набором процесів, що може розширюватись. В якості засобів реалізації процесу аналітичного опрацювання даних використано каркас, розроблений на основі бібліотеки FBP [4,5]. Досвід впровадження розроблених засобів показав, що можливість подання процесу аналітичного опрацювання даних у графічному вигляді, автоматична трансформація до виду, що може виконуватися в інструментальному середовищі, наявність інтерактивних засобів налагодження процесу аналітичного опрацювання даних дозволяють знизити трудомісткість розробки процедур попередньої обробки даних, моделей оцінювання, процедур формування аналітичних звітів з використанням результатів обробки даних великого обсягу.

Література

1. Кельберт М. Я., Сухов Ю. М. Вероятность и статистика в примерах и задачах. Т. II: Марковские цепи какотправная точка теориислучайныхпроцессов и их приложения.– М.: МЦНМО.– 2009. — 295 с.
2. Світовий центр даних з геоінформатики та сталого розвитку [Електронний ресурс] – Режим доступу:http://wdc.org.ua
3. Аналіз сталого розвитку – глобальний та регіональний контексти: моногр. /Міжнар.рада з науки (ICSU) [та ін.]; наук. кер. М.З.Згуровський. – К.:НТУУ «КПІ», 2010.– Ч. 2. Україна в індикаторах сталого розвитку.–220 с.
4. JavaImplementationof FBP Concepts (JavaFBP)[Електронний ресурс] – Режим доступу:http://www.jpaulmorrison.com/fbp/#JavaFBP
5. J.P.Morrison Flow-based programming : a new approach to application development.– Unionville, Ont.,CA : J.P. Morrison Enterprises.– 2010. – 336p.