Булах Б.В.
ННК "ІПСА" НТУУ "КПІ"

Грід-сервіси інтелектуального аналізу даних

У багатьох галузях науки, техніки, бізнесу нині накопичено величезні обсяги експериментальних чи статистичних даних. Але саме по собі нагромадження необроблених даних, вочевидь, не є ціллю – ці дані мають стати об'єктом аналізу, метою якого є вилучення з них корисної інформації. Практична цінність інтелектуального аналізу даних (ІАД), мета якого полягає у виявленні прихованих закономірностей у масивах необроблених даних, з часом лише зростає. І якщо причиною вражаючої динаміки збільшення обсягів накопичених даних став бурхливий розвиток інформаційних технологій, то цілком логічними є спроби подолати проблему ефективного аналізу даних саме за допомогою інформаційних технологій. Технологія Грід може стати надійним помічником у дослідженнях, які мають справу з обробкою надвеликих масивів даних.

Структурні елементи процесу ІАД загальновідомі. Алгоритмічне ядро ІАД складають різні за складністю процедури класифікації, кластеризації, асоціації, аналізу відхилень, прогнозування і т. д. Сам процес ІАД є багатокроковим та ітераційним, і включає такі етапи, як підготовка, очищення даних, вибір моделі та її параметрів, навчання моделі, аналіз якості проведеної обробки тощо. На сьогодні існує чимало програмних реалізацій алгоритмів та етапів ІАД: комерційних та з відкритим кодом; як у вигляді набору програмних модулей (напр. ADaM, Weka), так і бібліотек для мов чи середовищ програмування (таких як C/C++, Java); можливості з ІАД закладаються і у потужні СУБД (MS SQL Server, Oracle). Проте багато з них вимагають від користувача певних навичок програміста, особливо у випадку, коли надана функціональність недостатня для дослідників і потребує нарощення. З наявних продуктів більшість базується на власних підходах та рішеннях, і часто вони сумісні, у найкращому випадку, лише за форматами представлення даних. Болючим залишається питання обробки у прийнятний термін надмірно великих масивів даних. Як наслідок, користувачу часто бракує гнучкості наявного інструментарію, можливості легко організувати масштабні розподілені обчислення.

На подолання вищезгаданих недоліків спрямована пропозиція застосування Грід-технологій для ефективного рішення задач ІАД. Відомо, що обчислювальні задачі, рішення яких на окремому ресурсі було б неможливе, успішно вирішуються у Грід. Проміжне програмне забезпечення Грід пропонує набір служб, які беруть на себе задачі контролю доступу та безпеки, передачі даних, виділення обчислювальних ресурсів. Таким чином, інтегруючи інструментарій ІАД у Грід-інфраструктуру, можливо створити таке середовище, яке б дозволяло реалізовувати складні сценарії аналізу із залученням ресурсів Грід-мережі.

В рамках такої інтеграційної пропозиції можливим виглядає створення програмної платформи для ІАД, яка б задовольнила потребу аналітиків у гнучкому середовищі аналізу і надавала би можливість побудови складних розподілених сценаріїв обробки даних. Її основу складає екосистема Грід-орієнтованих сервісів. Ці сервіси реалізують окремі алгоритми та процедури обробки даних, доступні через стандартні інтерфейси (механізм веб-сервісів), сумісні з Грід-інфраструктурами (безпека, передача даних), можуть використовувати семантичні механізми для автоматизації їх виявлення та компонування у складніші сценарії. Платформа також має надавати прості засоби включення нових сервісів, створення сервісів на базі компонентів сторонніх розробників (принцип "обгорток"), зручний інтерфейс користувача, доступ через веб-портал як єдину точку входу до системи.

12th International conference on science and technology "System analysis and information technologies" SAIT 2010
Institute for Applied System Analysis of National Technical University of Ukraine "KPI", Kyiv, Ukraine, May 25–29, 2010
ISBN 978-966-2153-42-2, revision 1.1 (2010.05.30), available at http://sait.org.ua/books/sait2010.ebook.pdf