Розподілений інтелектуальний аналіз даних (Distributed Data Mining)

Старовойтенко Д.С.

ННК “ІПСА” НТУУ “КПІ”

Data Mining переводиться як "видобуток" або "розкопування даних". Нерідко поруч з Data Mining зустрічаються слова "виявлення знань в базах даних" (knowledge discovery in databases) і "інтелектуальний аналіз даних ". Сфера застосування Data Mining нічим не обмежена - вона скрізь, де є яка-небудь інформація.

Ідеальною платформою для DDM є кластер, створений з’єднанням між собою групи комп’ютерів, або кластери в так звані обчислювальні мережі, з’єднані через Інтернет . DDM є філією області інтелектуального аналізу, особлива увага якої приділяється розподіленим даним і обчислювальним ресурсам. Існує дві точки зору на то, яка інформація розподілена на сайті: однорідна (горизонтально розподілена) та гетерогенна(вертикально розподілена). Обидва підходи формулюють концептуально точку зору, яка полягає в тому, що таблиці даних на кожному сайті є розділи єдиної глобальної таблиці. В більшості методів та систем для DDM припускається, що ресурси розподілені по горизонталі і є однорідними. Кожний сайт має власні локальні дані і генерує дійсні локальні поняття. Згодом вони обмінюються з усіма іншими джерелами інформації, щоб отримати глобально діючі поняття. Кожен сайт може створити локальні набори певного рівня, які часто зустрічаються. Згодом всі місцеві результати об’єднуються і проводиться оцінка для отримання глобальних, частих наборів.

Існують такі Grid-сервіси для Distributed Data Mining :

1) Service Oriented Architecture (SOA). По суті це модель програмування для створення гнучких, модульних і сумісних програм. SOA дозволяє складання програм за допомогою частин незалежно від деталей реалізації, місця їх розміщення та початкової мети їх розроблення.

2) Open Grid Services Architecture (OGSA). Реалізація SOA моделі в контексті с Grid. OGSA забезпечує чітко визначений набір основних інтерфейсів для розвитку взаємодіючих систем і додатків Grid. Вона приймає web- служби в якості базової технології.

3) WS-Resource Framework (WSRF), який представляє собою набір з шести специфікацій, які підтримують Грід- сервіси та інші ресурси, які мають власний стан.

4) Open Service Framework for Grid-based Data Mining. Ця конструкція дозволяє розробникам створити дизайн Distributed Knowledge Data Discovery, як композицію з простих послуг, доступних в Grid. У той же час, ці послуги мають використовувати інші основні послуги Grid для передачі інформації та управління.

В інтелектуальному аналізі даних існують проблеми розподілення великих і складних наборів інформації. Основне вирішення - Distributed Data Mining. Так, як неефективно зберігати багато даних в одному місті, а інколи і зовсім неможливо, то вся інформація розбивається на декілька частин і відсилається на різні місця зберігання. Цей метод дозволяє оптимізувати аналіз інформації, оскільки навантаження ділиться між сайтами. Поєднання ефективних методів DDT і Грід-сервісів - визначає нові технології для роботи з великими і складними розподіленими даними.

1. Петренко А.І. Grid і інтелектуальна обробка даних Data Mining - «Системні дослідження і інформаційні технології», Київ, №4, 2008
2. Parallel and Distributed Data Mining: An Introduction http://dml.cs.byu.edu/~cgc/docs/atdm/Zaki.pdf
3. Grid-based Distributed Data Mining Systems, Algorithms and Services http://www.siam.org/meetings/sdm06/workproceed/HPDM/Domenico_talia_Invited_Session.pdf