Кислий Р.В. *— рецензент Петренко А.І.*
ННК "Інститут прикладного системного аналізу" НТУУ «КПІ», Київ, Україна

До побудови грід-систем знань (Knowledge Grid)

Сучасна грід мережа для здобуття знань – Knowledge Grid. Існує багато моделей, методів, алгоритмів та задач Knowledge Grid, в даному випадку розглянуто загальну модель побудови та використання Knowledge Grid. На рис.1 зображена загальна схема грід-системи знань.

Knowledge Grid дозволяє використовувати базові можливості грід для побудови сервісів, які підтримують розподілене здобуття знань у базах даних (KDD) в грід.

Такі послуги дозволяють користувачам використовувати програми для здобуття знань, які працюють з даними, програмним забезпеченням та обчислювальними ресурсами з різних частин гріду. З цією метою Knowledge Grid має механізми для публікації і пошуку інформації більш високого рівня, що управляють KDD додатками, а також візуалізують результати їх роботи. Такий підхід може бути описаний через багаторівневу архітектуру, як показано на рисунку 1.

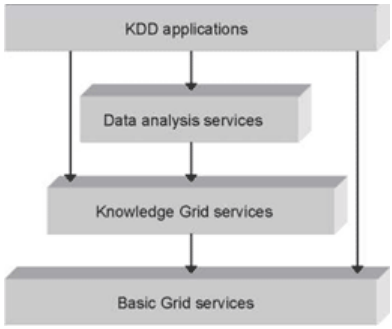

Рис. 1. Архітектура Knowledge Grid

Basic grid services. Основні функціональні можливості, що надаються стандартним середовищем грід, таким як Globus Toolkit, UNICORE та gLite.

Knowledge Grid services. Сервіси, спеціально розроблені для підтримки та реалізації data mining. Вони включають в себе керування ресурсами, які забезпечують механізми опису, публікації та отримання інформації про джерела даних, алгоритми data mining і обчислювальні ресурси, які дозволяють користувачам створювати і використовувати розподілені KDD програми.

Data analysis services. Спеціальні сервіси, які використовують знання грід-сервісів для забезпечення високого рівня аналізу даних. Служби аналізу даних може провести попередню обробку даних або запустити data mining (наприклад, класифікація, кластеризація тощо), або більш складне завдання виявлення знань (наприклад, паралельні класифікації, мета-навчання тощо).

KDD applications. Програми для здобуття знань можуть використовувати не тільки стандартні методи грід систем, а й інші моделі, мови й програми для створення розподілених KDD програм.

Основні компоненти Knowledge Grid розділяються на два типи: *Resource Management Services* і *Execution Management Services.*

На рис.2 показана архітектура Knowledge Grid. Усередині кожної групи є два рівні сервісів: сервіси високого рівня та рівня ядра. Ідея полягає в тому, що на рівні користувача програми безпосередньо взаємодіють з сервісами високого рівня, які для виконання запитів клієнтів викликають відповідні операції рівня ядра.

Resource Management Services. Ця група сервісів включає в себе стандартні сервіси та сервіси високого рівня для керування ресурсами Knowledge Grid. Серед таких ресурсів джерела даних і алгоритми мають основне значення. Тому архітектура Knowledge Grid передбачає спеціальні компоненти, а саме DAS і TAAS, для роботи з даними і алгоритми.

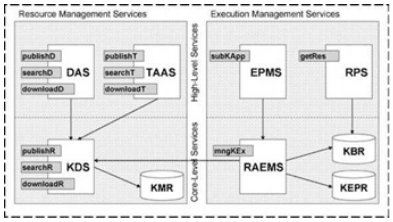

Рис. 2. Схема Knowledge Grid

Служби доступу до даних (DAS) пов'язана з публікацією, пошуком і передачею наборів даних, які будуть використовуватися в KDD додатках, а також пошуком висновків (в результаті роботи data mining). DAS виконує операції PublishData, SearchData і DownloadData. *PublishData* викликає на рівні користувача додаток для публікації метаданих про набір даних, як тільки операція публікації викликана, вона викликає PublishResource.[1]

SearchData викликає клієнтський інтерфейс, якому необхідно знайти набір даних на основі заданого набору критеріїв. DAS звертається з запитом до місцевих KDS, посилаючись на відповідні SearchResource і, як тільки пошук буде завершений, він отримує KDS. Такий результат полягає в наборі посилань на набори даних, що відповідають заданим критеріям пошуку. Варто відзначити, що пошукова операція не тільки обробляється на локальному комп'ютері, але і перенаправляється на інші хости. Операція DownloadData працює аналогічно попереднім: виконується введення набору даних, завантаження і перенаправлення запиту на DownloadResource місцевого компонента KDS.

Інструменти і алгоритми доступу (TAAS) пов'язані з публікацією, пошуком і передачею інструментів, які будуть використовуватися. Такі інструменти можуть бути інструментами data mining і засобами візуалізації. TAAS має ту ж базову структуру, що і DAS, та виконує основні завдання, взаємодіючи з місцевими KDS, які, у свою чергу, можуть викликати один або декілька інших віддалених примірників KDS. Операції, які експортуються TAAS: PublishTool, SearchTool і DownloadTool. Вони мають ті ж функціональні можливості, що і функції DAS, з тією різницею, що операції TAAS пов'язані з інструментами, а не з даними.[2]

Knowledge Directory Service (KDS) є єдиною службою на рівні ядра групи RMS. KDS керує метаданими, що описують ресурси Knowledge Grid. Такі ресурси включають вузли, сховища даних, інструменти і алгоритми, ZRS використовуються для вилучення, аналізу та керування даними, отриманими в результаті data mining. Така інформація зберігається в локальному сховищі, *Knowledge Metadata Repository* (KMR). PublishResource викликається для публікації інформації (метаданих) про ресурс; зберігаючи їх метадані в місцевій KMR. SearchResource операція викликається для отримання ресурсів на основі даного набору критеріїв представленого запиту. Важливим аспектом, який слід відзначити, є те, що KDS виконує такі завдання пошуку як на місцевому рівні, шляхом доступу до місцевих KMR, так і віддалено, за допомогою запитів інших віддалених KDS (що, в свою чергу, буде отримувати доступ до своїх місцевих KMR).

Execution Management Services. Послуги цієї групи дозволяють користувачеві створювати і запускати KDD додатки, а також для візуалізації результатів.

The Execution Plan Management Service (EPMS) дозволяє визначати структури програм, побудувавши відповідний графік виконання, і додавати набір обмежень за ресурсами. Цей сервіс на основі моделі, отриманої від клієнта, генерує відповідний абстрактний план виконання, який є формальним поданням структури програми. Як правило, вона не містить інформації про фізичні ресурси, які будуть використовуватися, а, скоріше, їх критерії. Тим не менш, EPMS може включати в себе як певні, так і абстрактні ресурси, тобто ресурси, які визначаються через логічні імена.

Results Presentation Service (RPS) надає можливості для представлення і візуалізації моделей знань (наприклад, правила асоціації, кластеризація моделей), а також зберігати їх у відповідний формат для подальшого використання.

The Resource Allocation and Execution Management Service (RAEMS) використовується, щоб знайти відповідність між абстрактним планом виконання (отриманому від EPMS) та наявних ресурсів, з метою врахування обмежень (CPU, пам'ять, бази даних, вимоги до пропускної спроможності), накладених на виконання плану. Результатом цього процесу є конкретний план виконання, який чітко визначає кількість ресурсів для кожного процесу data mining. Зокрема, він відповідає вимогам, зазначеним в абстрактному плані виконання з реальними іменами.

Література. **1.** Werner Dubitzky, Data Mining Techniques in Grid Computing Environments. *Univercity of Ulster*. UK, WILLEY-BLACKWELL, 2008, **2.** Haimonti Dutta, DISTRIBUTED DATA MINING ON A GRID INFRASTRUCTURE, A PROPOSAL FOR DOCTORAL RESEARCH, 2006.