УДК 681.3

С.В. МІНУХІН, С.В. ЗНАХУР

МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО ПОШУКУ ІНФОРМАЦІЇ В GRID-МЕРЕЖІ

Розглядається підхід щодо кластеризації ресурсів GRID-мережі для розподіленого пошуку інформації на основі використання нейронної мережі. Механізм пошуку базується на архітектурі LDAP та асоціює вузли GRID-мережі з відповідними ресурсами та запитами за допомогою мережі Кохонена.

1. Вступ

Однією з сучасних парадигм побудови глобальних комп'ютерних мереж є парадигма GRID-системи, де реалізована ідея використання архітектури розподіленого комп'ютинга з колективною формою доступу до обчислювальних та інформаційних ресурсів. Ключовою проблемою використання інформаційних ресурсів GRID є експоненціальний ріст інформації та існуюча гетерогенність глобальної мережі [1,2,7,8]. Сучасні інформаційні пошукові системи (ІПС) не дозволяють адекватно та швидко обробити запити користувачів. Тому для пошуку ресурсів у GRID-мережах пропонується здійснити їх метаопис та визначити, на яких серверах (вузлах – сховищах даних) вони будуть зберігатися та оброблятися, тобто необхідно кластеризувати інформаційні ресурси згідно з атрибутами їх метаопису. Для рішення цієї задачі використовувалися статистичні методи кластеризації, які відносять образ ресурсу до відповідного класу (кластеру) із множини вузлів. Аналіз існуючих ієрархічних і неієрархічних методів математичної статистики показав, що існуючі неієрархічні методи виявляють більш високу стійкість по відношенню до шумів (викидів), некоректного вибору метрики, включення незначних змінних для кластеризації. Перевага ієрархічних методів – в наочності побудови дерева кластерів і можливості отримання детального уявлення про структуру даних. Останнім часом ведуться активні розробки нових алгоритмів кластеризації, здатних обробляти надвеликі бази даних. До таких алгоритмів відносяться: Birch, Cure, Clarans, DBScan [1,3]. Основним недоліком цих алгоритмів є те, що вони вимагають вибору деяких порогів щільності для спостережень, що є суб'єктивною апріорною інформацією, яка може бути недоступною. Загальним недоліком статистичних методів є достатньо висока часова затримка рішення задач кластеризації для об'ємів даних більш ніж пентабайт та необхідність апріорної інформації щодо характеристик спостережень або кластерів. Дослідження показали, що нейронні мережі (НМ) з успіхом можуть застосовуватися в різних галузях, особливо в задачах класифікації та кластеризації. Використовуючи відповідний клас НМ, можна реалізувати рішення задачі кластеризації метаінформації відповідно до вузлів мережі. Таким чином, метою роботи є побудова механізму пошукової машини GRID-мережі для ефективного пошуку інформаційних ресурсів, який використовує апарат штучних НМ [4,5]. Задачі дослідження такі:

1) визначення загальної архітектури ІПС для GRID-мережі на основі використання протоколу LDAP;

2) кластеризація метаінформації ресурсів за допомогою НМ Кохонена та асоціація їх з вузлами обробки (серверами);

3) визначення механізму розподілу запитів відповідно до вузлів оброблення (серверів).

2. Сутність

Розглянемо механізм пошуку інформації у Grid-системі [2, 6]. Як архітектура пошукової системи для Grid-системи пропонується кластерна модель із вбудованим механізмом пошуку LDAP (рис.1). LDAP (англ. Lightweight Directory Access Protocol - «полегшений протокол доступу до каталогів») - це мережевий протокол для доступу до служби каталогів X.500. Сервіс каталогів - це оснащений засобами пошуку репозиторій, в якому наділені відповідними повноваженнями користувачі та служби можуть знаходити інформацію про ресурси, обчислювальні вузли, мережеві пристрої і програми. На рис. 1 цифрами позначено порядок взаємодії компонентів при надходженні даних до системи, а буквами – порядок дій при обробленні запиту користувача. Дані зберігаються в файловій системі сервера (серверів), на якому розміщено сховище даних. Індексація даних виконується за допомогою спеціального індексатора, який повинен бути окремо розроблений для кожного типу даних. Його задача полягає в скануванні файлової системи для пошуку нових файлів з даними, виділенні з них метаінформації та генерації XML-документа за встановленою схемою, що містить структуровану інформацію про нові дані.

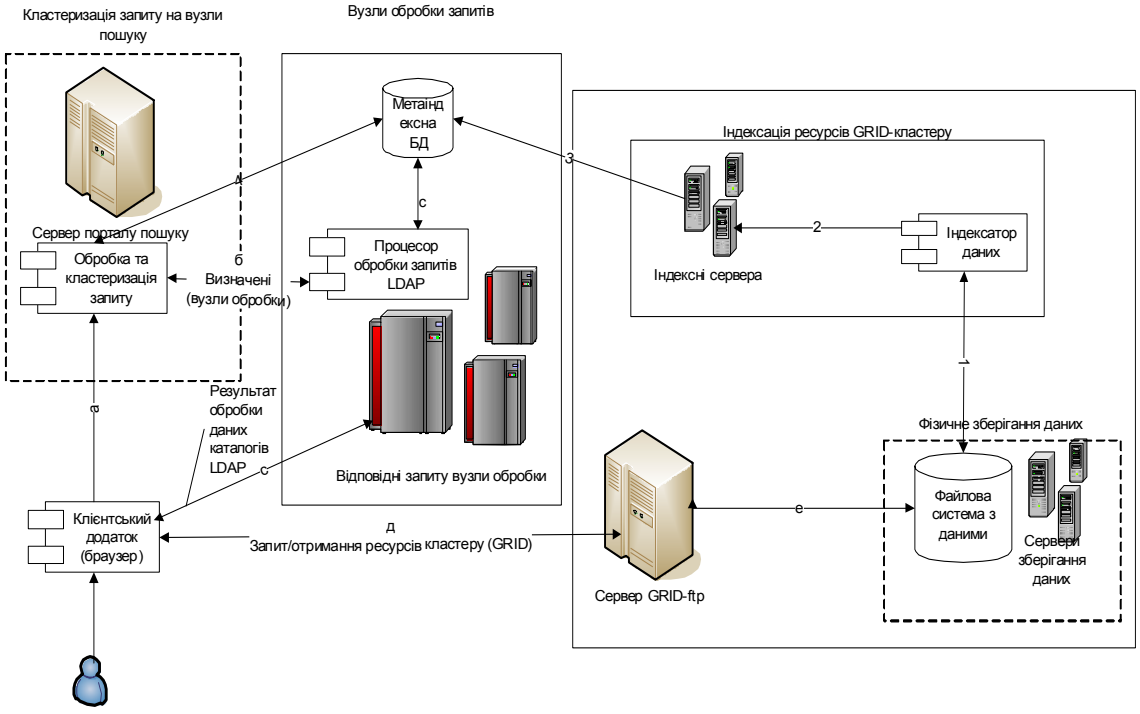

Рис. 1. Архітектура пошукової системи для GRID-мережі

У сучасних пошукових системах на основі індексування вихідні документи заносяться в базу даних без будь-якого додаткового перетворення, але при цьому смисловий зміст кожного документа відображається в деякий пошуковий простір [1, 2]. Процес відображення документа в пошуковому просторі називається індексуванням та полягає в присвоєнні кожному документу деякого індексу-координати в інформаційному просторі. Формалізоване представлення (опис) документа називається пошуковим індексом документа. Користувачеві досить сформувати пошуковий образ запиту для пошуку документа [1, 2].

Пошукова система на основі певних критеріїв і способів шукає документи, пошукові образи яких (ПОД) відповідають пошуковим образам запиту (ПОЗ) користувача, і видає релевантні запиту документи. Загальна схема пошуку на основі ПОЗ наведена на рис. 2.

Індексатор GRID реєструється у індексному сервісі Globus Toolkit, який використовується для розробки та функціонування проміжного програмного забезпечення GRID, і періодично ним виконується [6-8]. Дані, що генерує індексатор, надходять до загального дерева даних індексного сервісу ресурсу, і у випадку, коли цей ресурс не є вершиною ієрархії індексних сервісів, утворених за допомогою сервісу агрегації, передаються вище за ієрархією.

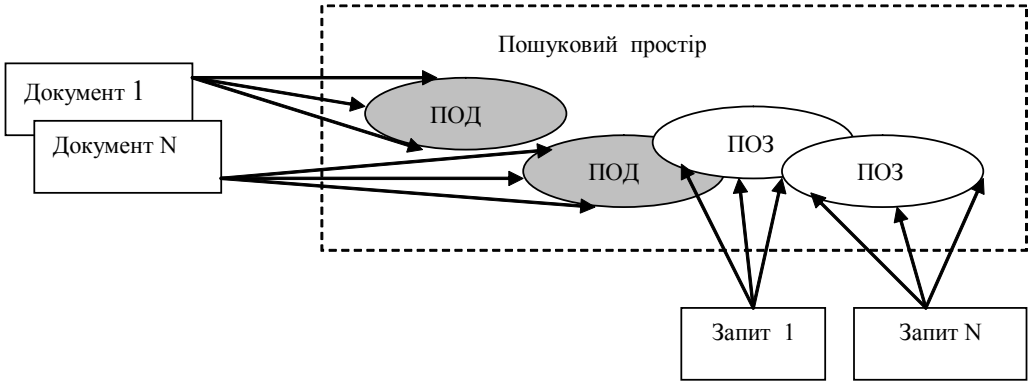

Рис. 2. Пошук інформації по ПОД

В результаті головний індексний сервіс системи завжди має детальну і актуальну інформацію про дані, які присутні в усіх архівах, документах GRID-системи. У тому випадку, коли додаток користувача має отримати певні дані зі сховища даних, він використовує засоби пошуку по індексному сервісу системи і отримує адресу URL файлу, що розташований в певному архіві. Якщо користувачеві відомо, в якому архіві знаходяться необхідні йому дані, він може проводити пошук не за головним індексним сервісом, а за індексним сервісом необхідного йому архіву, що сприяє більш рівномірному розподілу навантаження за ресурсами системи. Опис ресурсу згідно з протоколом LDAP дозволяє використовувати атрибути документа (згідно з ієрархією опису ресурсу).

Для підвищення швидкості пошуку в роботі пропонується використовувати механізм визначення серверів (вузлів) пошуку на основі метаінформації щодо ресурсу. Для цього необхідно вирішити такі завдання:

1) кластеризація ресурсів на основі їх атрибутів опису в LDAP (метаінформації);
2) асоціація кластеризованих ресурсів та вузлів їх оброблення;
3) асоціація атрибутів запиту з вузлами оброблення;
4) визначення достатньої кількості вузлів оброблення для оброблення потоку запитів.

В роботі перші три завдання пропонується вирішити на основі апарату інтелектуальної обробки даних (нейронної мережі).

Для завдання представлення ресурсу у вигляді системи знань вхідними даними є текст, а результатом – система знань у вигляді набору метаданих. Кластеризація метаданих дозволяє кластеризувати документ (ресурс).

Таким чином, індексатори формують метаопис ресурсу GRID-мережі, що має перелік атрибутів, найважливішим з яких є ключові слова й частотна характеристика ключових слів документа. Для подальшого розподілу метаінформації щодо серверів БД необхідно здійснити кластеризацію документів у режимі реального часу. В результаті метаінформація щодо ресурсу (документа) буде віднесена до відповідного сервера (серверів), що у подальшому дозволить на них здійснити й обробку запитів на пошук документа: ключові слова запиту можливо асоціювати з тими кластерами (серверами), де зберігається відповідна метаінформація щодо документа. Реалізувати поставлене завдання пропонується за допомогою мережі Кохонена. Мережа Кохонена – метод, розроблений для відображення багатовимірних даних на двовимірну площину. Ця мережа навчається без вчителя: на вхід поступають навчальні дані і відбувається корекція синаптичних ваг нейронів відповідно до нових спостережень. Швидкість навчання мережі залежить від порядку надходження навчальних даних на вхід мережі [3].

Рішення завдання

Для функціонування інтелектуального репозитарію метаінформації необхідні такі функціональні модулі:

– модуль індексації документів;

– модуль кластеризації документів.

На модуль індексації покладені завдання передоброблення документів і побудова частотних словників термінів, що зустрічаються. Далі, в рамках модуля кластеризації і класифікації, на основі значень відносних частот повинні створюватися наочно-орієнтовані кластери. В процесі класифікації виконується завдання зіставлення інформаційного ресурсу з певним кластером.

Для оцінки значущості слів в індексаторі використовуються методи визначення частот слів кожного документа і частот, розрахованих за формулою Шенона (відношення «сигнал-шум»)[1]:

$$w_i = \frac{S^k}{N^k},$$

де N^k – шум терміну,

$$N^k = \sum_{i=1}^{n} \frac{f_i^k}{F^k} * \mathrm{Log}\frac{F^k}{f_i^k},$$

тут f_i^k – частота k-го терміну в і-му документі; F^k – частота k-го терміну по всіх документах; S^k – сигнал терміну,

$$S^k = \mathrm{Log}F^k - N^k.$$

Для кластеризації застосовується нейронна мережа, що використовує метод навчання без вчителя (unsupervised learning), – самоорганізуючі карти Кохонена (Self-Organizing Map – SOM)[4].

Пропонується використовувати дві основні процедури настройки нейронної мережі: ініціалізація ваг нейронів випадковим чином і самонавчання мережі Кохонена (алгоритм SOM).

Алгоритм навчання мережі Кохонена

Крок 1. Ініціалізація параметрів мережі.

Крок 2. Цикл за числом ітерацій в мережі.

Крок 2.1. Визначення відстаней між вхідним вектором X і вектором ваг W кожного нейрона за формулою:

$$D_j = \sqrt{\sum_i (x_i - w_i)_j^2}.$$

Крок 2.2. Визначення нейрона-переможця з мінімальною відстанню.

Крок 2.3. Визначення області активації нейрона-переможця.

Крок 2.4. Визначення ваг нейронів усередині області активації за формулою [4]:

$$W_j(t+1) = W_j(t) + a[X - W_j(t)],$$

де a – крок навчання для мережі Кохонена.

Крок 3. Запис документа в масив кластерів.

Вихідний потік кластерів представляється у вигляді динамічного двомірного масиву. При попаданні документа в кластер на перетині «документ-кластер» в комірці ставиться одиниця. Навчання нейронної мережі відбувається на кожному документі. Таким чином, метаінформація щодо кожного документа буде зберігатися у відповідному кластері (серверах).

Аналіз роботи мережі Кохонена показав, що вона здатна розділяти спостереження лише за ступенем близькості їх ознак. При цьому номер вузла, до якого віднесено спостереження, та номер його класу, в загальному випадку, не збігаються, тобто мережа не наділяє кожен з вузлів конкретним змістом. Результати експериментів показали, що мережу доцільно застосовувати тільки для виділення центрів кластерів спостережень, а не для асоціювання серверів відповідним кластерам [4].

Після кластеризації метаінформації щодо ресурсів та асоціації її відповідним вузлам GRID- мережі необхідно вирішити наступне завдання пошукової системи – асоціювати запити користувачів з вузлами, які містять відповідну до запитів метаінформацію. В роботі навчену мережу Кохонена пропонується використати для кластеризації запитів користувачів згідно з існуючими кластерами (серверами) обробки БД метаінформації. Для цього запит користувача за допомогою процедури парсінгу (parsing) [2] розділяється на окремі ключові слова, яким необхідно дати вагу (аналог частотної характеристики) [1]. Отриманий вектор атрибутів подається на вхід мережі Кохонена для його асоціювання з кластером (серверами обробки). Послідовність рішення задачі кластеризації запитів на отримання інформаційних ресурсів представлено за допомогою нотації діаграми IDEF0 (рис. 3), що дозволяє структуровано уявити механізм асоціації запитів до серверів, які зберігають метаінформацію щодо кластеризованих ресурсів. На діаграмі рис. 3 слід виділити такі процеси: розміщення інформаційного ресурсу (документа), індексування документа і формування БД пошукової машини, тестування й експлуатація пошукової системи, попередній статистичний аналіз результатів виконання запитів, кластеризація інформаційних ресурсів на основі статистичної обробки даних з використанням апарату штучного інтелекту.

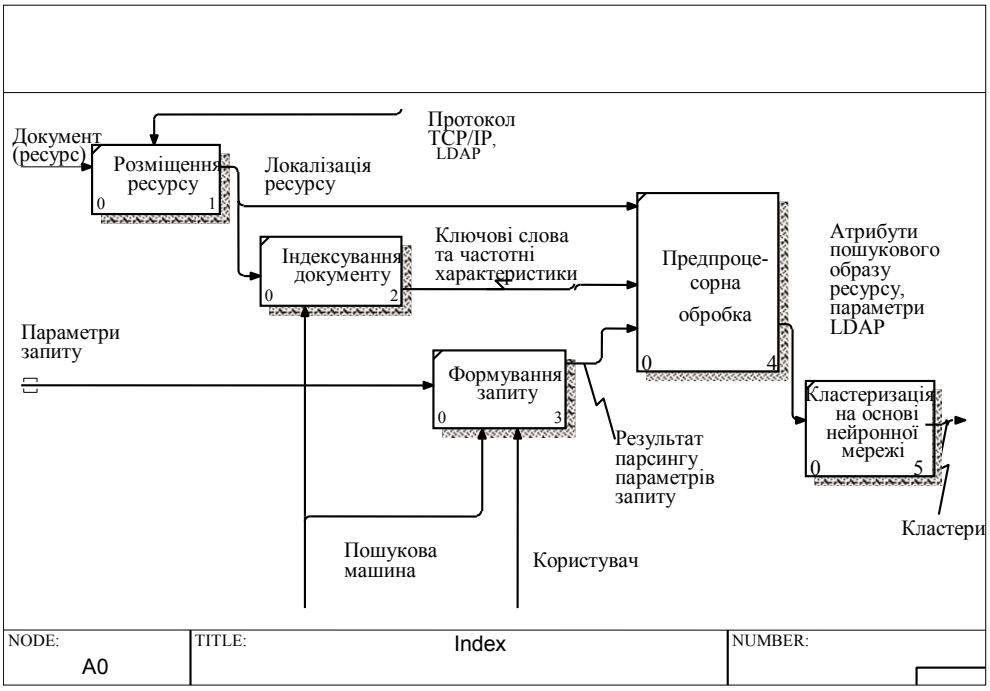

Рис. 3. Основні процеси організації інтелектуального пошуку в GRID- мережі

Сервер-диспетчер (сервер порталу рис. 1), отримавши запит від користувача, на основі кластеризації його параметрів асоціативно зв’язує запит з вузлами (серверами) GRID-мережі. Мережа Кохонена дозволяє визначити декілька серверів, які містять близьку за складом метаінформацію (але не тотожну). У випадку, коли сервери вільні, запит можливо обробити на них одночасно. Сервери, які виділені в асоційовані таксономічні кластери, містять близьку за змістом інформацію й відповідають однаковим параметрам запиту користувачів. Тому для збільшення швидкості отримання результатів пошуку доцільно організувати запит до БД кожного з них.

Таким чином, досягається оптимізація пошуку за рахунок одночасного використання декількох серверів та розподільного зберігання метаінформації щодо документів у кластерах.

3. Висновки

В статті описано модель пошукової системи, яка дозволяє одночасно обробляти запити користувачів до ресурсів GRID-систем на асоційованих вузлах GRID (серверах обробки

метаінформації). Рішення задачі розподілу запитів між вузлами (серверами) дозволяє суттєво збільшити ефективність обробки запитів за рахунок одночасного використання достатньої кількості вільних серверів. Ефективність визначається зменшенням часу обслуговування запиту відповідно до варіанту централізованої його обробки. Запропонований в роботі апарат інтелектуального пошуку базується на вирішенні задач кластерного аналізу метаінформації щодо ресурсів GRID, що дозволило класифікувати образ ресурсу до відповідного вузла мережі. Вирішення задачі кластеризації здійснюється за допомогою штучної нейронної мережі Кохонена, яка визначає у оперативному режимі приналежність ресурсу до відповідного класу (множини) вузлів обчислювального кластера (GRID-мережі). Мережа Кохонена дозволяє вирішувати й зворотне завдання – визначення відповідних вузлів кластера для оброблення запитів користувачів, які містять метаінформацію щодо необхідних ресурсів для формування переліку відповідей щодо запитів. Практичне значення дослідження полягає у можливості побудови механізму пошуку інформації, який дозволяє в результаті кластеризації та асоціації запитів у нейронній мережі отримати підмножину вузлів мережі, де запити будуть оброблюватися одночасно. Подальші дослідження можливо проводити у напрямку розробки алгоритмів для паралельного рішення завдання кластеризації на декількох процесорах (вузлах) GRID-мережі в MPI.

Список літератури: 1. *Игумнов Е.* Основные концепции и подходы при создании контекстно-поисковых систем на основе реляционных баз данных // http://www.citforum.ru/database/articles/search_sys.shtml. **2.** *Пономаренко В.С.* Методы и модели планирования ресурсов в Grid-системах / В.С. Пономаренко, С.В. Листровой, С.В. Минухин, С.В. Знахур: Монография. Х.: ВД «ІНЖЕК», 2008. 408 с. **3.** *Уиллиамс У.Т., Ланс Д.Н.* Методы иерархической классификации // Статистические методы для ЭВМ / Под ред. М. Б. Малютов. М.: Наука, 1986. С. 269–301. **4.** *Круглов В.В., Борисов В.В.* Искусственные нейронные сети. Теория и практика. М.: Горячая линия –Телеком, 2002. 382 с. **5.** *Барсегян А.А., Куприянов М.С,. Степаненко В.В, Холод И.И.* Методы и модели анализа данных: OLAP и Data Mining. Спб.: БХВ-Петербург, 2004. 336с. **6.** *Валиев М.К, Китаев Е.Л., Слепенков М.И.* Использование службы директорий LDAP для представления метаинформации в глобальных вычислительных системах (Using LDAP directory service for representation of metainformation in global computing systems): http://www.keldysh.ru/metacomputing/ism99.html. **7.** *Globus* Toolkit: http://www.globus.org. **8.** http://www.gridclub.ru/activity/kiam/documents.html.

Надійшла до редколегії 12.04.2009

Мінухін Сергій Володимирович, канд. техн. наук, доцент кафедри інформаційних систем ХНЕУ. Наукові интереси: інтелектуальна обробка інформації. Адреса: Україна, 61145, Харків, вул. Новгородська, 6-а, кв. 77, тел. 702-18-31, e-mail: ms_vl@mail.ru.

Знахур Сергій Вікторович, канд. економ. наук, доцент кафедри інформаційних систем ХНЕУ. Наукові інтереси: інтелектуальна обробка інформації. Адреса: Україна, Харків, пр. 50 -річчя ВЛКСМ, 32/186, тел. 702-18-31, e-mail: sergznakhur@mail.ru.