Знахур С.В.
Харківський національний економічний університет

Концепція пошуку інформації у GRID-мережі

Концепція пошукової системи базується на передумові, що існують регіональні GRID-портали, до яких підключено кластерні структури. Запити користувачів GRID-мережі ставляться в чергу на виконання (керуюча програма перебуває на порталі) згідно умовам сертифікату користувача. Для кожного запиту керуюча програма визначає його рівень складності й пріоритет. Керуюча програма на основі кількості запитів і їхньої складності визначає мінімальну кількість серверів БД (Data Storage), які ці запити можуть обробити. Складність запиту визначається кількістю ключових слів і логічних операцій. Пріоритет запиту визначається на підставі черговості в стеці запитів, вартості запиту. Одним із завдань управління процесом пошуку є надходження мінімальної кількості серверів, що забезпечують обслуговування всіх запитів з необхідною ефективністю. Це можливо вирішити на основі рішення завдання про найменше покриття (ЗНП). Тобто проблему розподіленого пошуку можливо вирішити на основі використання апарату теорії графів. Але головне завдання інформаційно-пошукової системи (ІПС) – пошук інформації, релевантної інформаційним потребам користувача. Виникають два взаємозалежні завдання: подання інформації в системі і її пошук відповідно до інформаційних потреб користувача. Переважну більшість пошукових алгоритмів засновано на векторній моделі тексту. У роботі векторною будемо називати модель опису інформації області, а лінійною – модель пошуку інформації. Такий поділ обумовлено тим, що документи (або їх ідентифікатори у випадку пошуку за назвою) записуються у вигляді двійкових векторів, у той час як пошукові запити – це лінійні перетворення над цими векторами. У векторній моделі інформаційного потоку можна виділити кілька основних понять: словник, документ, потік і процедури пошуку й корекції запитів. Під словником розуміють упорядкована безліч термінів, потужність якого позначають як M. Документ – це двійковий вектор розмірності M. Якщо термін входить у документ, то у відповідному розряді цього двійкового вектора проставляється 1, у противному ж випадку – 0. Всі операції в лінійній моделі індексування й пошуку документів виконуються над пошуковими образами документів. В роботі пропонується використання алгоритму Солтона, який базується на стверджені, що для індексування використовують тільки ті терміни (слова), які мають високу частоту зустрічальності усередині документа й низьку у всьому інформаційному масиві. Ця характеристика обчислюється як відношення частоти зустрічальності терміна в документі до частоти зустрічальності терміна в БД (масиві). Використовуючи цю міру системи індексування, документу асоціюють перші 20 – 40 символів, які й становлять його пошуковий образ. Складність процедури релевантного пошуку у GRID-мережі зумовлено тим, що відомі алгоритми орієнтовано на пошук у локальній інформаційній системі. В глобальних мережах існують наступні проблеми: по-перше, немає єдиного інформаційного масиву, який можна було б одразу завантажити, побудувавши індекс; по-друге, через відсутність єдиної інформаційної служби не можливо організувати систему з контрольованим словником. У такий спосіб в ІПС глобальної мережі відбуваються два процеси: постійний ріст інформаційного масиву (об'єму БД), з одного боку, і постійне збільшення словника системи – з іншого. У таких умовах рішення задачі пошуку інформації у GRID-системах пропонується реалізувати наступним чином. Розміщення словника на вузлах GRID мережі, які виконують функції Data Storage. Оновлення словника відбувається синхронно на основі рішення задачі про найменше покриття (ЗНП), тобто організується пошук тих термінів та асоційованих з ними документів, які були змінені (або додані), для подальшої їх передачі в словники (та індекси) на усі вузли кластеру (кластерів) GRID мережі. У цьому випадку не здійснюється передача фізичних документів, а передаються зміни у словнику та відповідні асоціації з фізичними документами.