УДК 004.735

**К.т.н., доцент Орлова М.М., к.т.н., доцент Сапсай Т.Г.,
магістрант Спринсян Я.В.**

**Національний технічний університету України
«Київський політехнічний інститут»**

АЛГОРИТМИ ПОШУКУ ДАНИХ В PEER-TO-PEER МЕРЕЖАХ

Вступ

Незважаючи на швидке зростання продуктивності комп'ютерної техніки, з розвитком мережі Інтернет зростання обсягів даних, які необхідно розміщувати, передавати, а головне шукати, призвело до того, що системи збереження даних, які базуються на класичній клієнт-серверній архітектурі фактично досягли межі своєї масштабованості. Потужність сучасних персональних комп'ютерів, ширина пропускання каналу більшості користувачів Інтернету, а також можливість збереження великої кількості даних створили передумови для створення peer-to-peer (P2P, піринговых) мереж. P2P мережа – це мережа, яка ґрунтується на принципі рівноправності учасників. Тобто в чистій peer-to-peer мережі не існує поняття клієнтів та серверів, лише рівні вузли, які одночасно функціонують як клієнти та сервери по відношенню до інших вузлів мережі. Ця модель мережевої взаємодії відрізняється від клієнт-серверної архітектури, в якій зв'язок відбувається лише між клієнтами та центральним сервером. Така організація дозволяє зберігати працездатність мережі при будь-якій конфігурації доступних їй вузлів. Проте існують P2P мережі, які все ж мають сервери, але їх роль полягає вже не у наданні сервісів, а у підтримці інформації з приводу сервісів, що надаються клієнтами мережі.

Постановка задачі

Метою даної роботи є аналіз та дослідження особливостей функціонування peer-to-peer мереж та алгоритмів пошуку даних в таких мережах. Для цього вирішуються наступні задачі:

1. Аналіз особливостей пошуку даних в P2P мережах.
2. Дослідження алгоритмів пошуку даних в централізованих та децентралізованих P2P мережах.

Класифікація P2P систем

За функціями:

- Розподілені обчислення. Складна обчислювальна задача розділяється на невеликі незалежні частини. Обробка кожної з частин виконується на окремому комп'ютері і результати збираються на центральному сервері. Центральний сервер відповідає за розподілення елементів роботи серед окремих комп'ютерів в Інтернеті. Кожен із зареєстрованих користувачів має клієнтське програмне забезпечення. Воно використовує час простою комп'ютера для виконання обчислень, наданих сервером. Після того, як обчислення закінчені, результат надсилається до сервера, і інша задача передається клієнту.
- Файлообмін. До цієї категорії належать більшість P2P мереж. Вони включають системи та інфраструктури, що розроблені для розподілу цифрової мультимедійної інформації та інших даних між користувачами. Такі системи використовуються як для розподілу контента у відносно простих додатках для прямого розподілення файлів, так і у більш складних, які створюють розподілу середовища зберігання даних, що забезпечує їх безпеку, ефективну організацію, індексацію, пошук, оновлення і отримання. Прикладом таких мереж є Gnutella, Kazaa, Freenet, Groove та інші.
- Співпраця. Природа технології P2P робить її придатною для забезпечення співпраці між користувачами. Це може бути обмін повідомленнями, он-лайн ігри, сумісна робота над документами тощо.

За ступенем централізації:

- Чисті peer-to-peer мережі. Вузли є рівними. Кожен вузол виконує роль як сервера, так і клієнта. Не існує центрального сервера, що керує мережею. Прикладами таких систем є Gnutella та Freenet.
- Гібридні peer-to-peer мережі. Мають центральний сервер, що зберігає інформацію про вузли та відповідає на запити відносно цієї інформації. Вузли забезпечують мережу ресурсами (центральний сервер їх не має), повідомляють сервер про наявність цих ресурсів, надають ресурси іншим вузлам.

За способом з'єднання:

- Неструктуровані мережі. Розміщення контенту (файлів) в неструктурованій мережі ніяк не пов'язане з топологією оверлейної мережі. Механізми пошуку варіюють від таких методів як «breadth-first» або «depth-first», в яких кількість запитів, що передаються вузлам по мережі, росте експоненціально, до більш формалізованих

алгоритмів, таких як «Random Wolkers Algorithm». Головним недоліком таких мереж є те, що, оскільки немає ніякої кореляції між вузлами та даними, які вони зберігають, то й немає ніякої гарантії, що запит знайде вузол, який має бажані дані.

- Структуровані мережі. В структурованих мережах оверлейна топологія строго контролюється, і файли (або вказівники на них) розміщуються чітко у визначених місцях. Ці системи, по суті, забезпечують відповідність між контентом і його місцезнаходженням. Тож запити можуть бути ефективно направлені вузлу з контентом, що необхідно знайти. На даний момент найбільш популярним типом структурованої мережі P2P є розподілені хеш-таблиці, в яких хешування використовується для встановлення зв'язку між даними та конкретним вузлом, який за них відповідає.

Проблеми пошуку в P2P мережах

Задача пошуку в піринговых мережах зводиться до швидкого та ефективного знаходження найбільш релевантних відповідей на запит, що передається вузлом у всю мережу. Зокрема, актуальна задача — зменшення мережевого трафіку, що з’являється під час обробки запиту (наприклад, пересилання запиту великій кількості вузлів), і у той же час отримання швидких і якомога якісніших результатів.

Алгоритм Random Walkers

Алгоритм Random Walkers (RWA) – це алгоритм, що найчастіше застосовують для пошуку в неструктурованих мережах. Ключова ідея RWA полягає в тому, що кожний вузол випадково надсилає повідомлення із запитом (так звана «посилка») одному із своїх сусідніх вузлів. Щоб скоротити час, необхідний для отримання результатів, ідея однієї «посилки» розширена до «k посилок», де k — кількість незалежних посилок, послідовно запущених від вихідного вузла. Очікується, що k «посилок» після T кроків досягнуть тих самих результатів, що й одна посилка за kT кроків, тобто в методі Random Walkers передбачається лінійне збільшення повідомлень, що розсилаються.

Розподілена хеш-таблиця

Розподілена хеш-таблиця (Distributed hash table, DHT) — протокол передачі даних та механізм децентралізованого (без виділеного сервера) збереження інформації про ресурси та вузли файлообмінної мережі типу

peer-to-peer. Однією з реалізацій DHT є протокол Kademlia. В такій мережі кожен вузол мережі при першому підключенні до мережі отримує унікальний номер (ID), що вибирається з певної множини (в деяких реалізаціях, наприклад в Kademlia, − це 160-бітове число), яке генерується випадковим чином. Для порівняння двох ID вводиться поняття метрики або відстані. У випадку Kademlia воно обчислюється як виключне "або" двох чисел (XOR). Чим менше значення такої відстані, тим два вузли мережі вважаються ближчими один до одного. Метрика, введена таким чином, не відображає географічної близькості учасників мережі.

Коли вузол розміщує у мережі деякий ресурс (файл), він обробляє його зміст та обчислює значення хеш-функції, яка буде ідентифікувати ресурс у мережі. Хеш-функція обирається таким чином, щоб унікальні номери учасників та обчислені ключі ресурсів набували значень з однієї множини. Обчисливши значення хеш-функції, вузол шукає інший модуль, ID якого близький до знайденого ключа. Знайшовши такий вузол, розміщувач ресурсу передає свою IP-адресу та ключ, які знайдений модуль зберігає у себе.

Таким чином, клієнт мережі, який потім хоче завантажити ресурс, знаючи з деяких джерел його ключ, намагається отримати інформацію про знаходження цього ресурсу в тих учасників мережі, унікальний номер яких близький до ключа ресурсу, який необхідно знайти.

Пошук ресурсів в структурованій P2P мережі

Пошук ресурсів за назвами файлів (або за його описами, тегами) в децентралізованій структурованій P2P мережі може бути організовано у такий спосіб. Ім'я файлу розбивається на ключові слова, які при розміщенні ресурсу хешуються та зберігаються у мережі разом із назвою файлу та його хешем. Номер вузла, на якому ці відомості зберігаються, знаходиться аналогічним чином — він має бути якомога ближче до значення хешу відповідного ключового слова. Пошук за ключовими словами відбувається так: обчислюється ключ слів запиту, потім в учасників мережі, які мають ID близькі до цього ключа, відшукується повна назва файлу разом зі значенням хеш-функції.

Висновки

Проаналізовані алгоритми пошуку файлів дозволяють користувачу, знаючи їх назву, частину назви або опис, швидко знайти потрібну інформацію, не перевантажуючи мережу численним «сліпими» пересиланнями.

Слід зазначити, що, на відміну від мереж з клієнт-серверною архітектурою, задача ефективного пошуку в піринговых мережах — відкрита дослідницька проблема. В подальшому можлива розробка структури пірингової мережі із врахуванням географічного положення вузлів. Це дозволить побудувати систему пріоритетів, що, перш за все, призведе до швидшого завантаження потрібної інформації, а для тувачів Інтернету з розподіленим трафіком – й до економії коштів.

Література

1. *Ландэ Д.В.* P2P - по секрету всему свету. О пиринговых сетях // Сети и бизнес, 2008. - № 2 (39). – С. 104-110.
2. *М.Ю. Звягин, П.Ю. Шамин, В.Г. Прокошев.* Анализ эффективности алгоритмов поиска информации в децентрализованной сети // Информационные технологии и телекоммуникации в науке и образовании (IT&T ES'2007). Материалы международной научной конференции/ Редкол.: А.Н. Тихонов (пред.) и др.; ФГУ ГНИИ ИТТ «Информика».- М.: ЭГРИ, 2007. – С. 27-28.
3. *Kiyohide Nakauchi, Yuichi Ishikawa, Tomonori Aoyama.* Peer-to-peer Keyword Search Using Keyword Relation // The University of Tokyo, 2008. – P. 38 – 45.
4. *P. Maymounkov, D. Mazieres.* Kademlia: A Peer-to-Peer informatic system based on the XOR metric // Proc. 1st IPTPS, March 2002. – P. 53 – 65.