УДК 004.6

Збирання даних про користувачів віртуальної соціальної мережі за допомогою web-кроулера

Охотний С.М., студент 3 курсу,
Мелешко Є.В., канд. техн. наук, доцент
Центральноукраїнський національний технічний університет, м. Кропивницький

На сьогоднішній день соціальні мережі стали основою спілкування людей і відіграють ключову роль в інформаційній глобалізації суспільства. Соціальні мережі активно використовуються для розповсюдження реклами та маніпулювання свідомістю громадськості. Тому важливим питанням сучасності є захист людей від впливу негативної та шкідливої інформації, яка може нашкодити їхньому життю.

Соціальну мережу можна представити у вигляді графу, кожна вершина якого є її користувачем, а ребро – зв'язком між користувачами. Нині найбільшою соціальною мережею є Facebook, який має 1,7 млрд. користувачів, серед яких 5,4 млн. українців. Завантажити інформацію про користувачів соціальної мережі можна двома шляхами: використовуючи API, яке надається розробниками, або аналізуючи html-сторінки сайту мережі. За останній час багато соціальних мереж закрили можливість вільно збирати інформацію про користувачів (або ускладнили її), надаючи засоби в основному для розробки додатків. Крім того розробник стає залежним від змін в API соціальної мережі. При аналізі html-сторінок така залежність відпадає, але потрібно пам'ятати про розмітку сайтів, яка може змінюватися, а це у свою чергу вимагає постійної перевірки і внесення змін в розроблені для цього парсери html-сторінок. Слід зазначити, що можливість внесення змін розробниками сайту в html-розмітку можна попередньо закласти в архітектуру розроблюваного ПЗ, щоб зменшити та полегшити правки, які треба буде вносити у такому випадку. При аналізі html-сторінок не є практичним намагання отримати граф усієї соціальної мережі, адже для цього знадобиться завантажити близько 300 ТБ даних і обробити їх.

Для збору інформації з веб-сторінок використовують спеціалізоване ПЗ, яке можна узагальнити під термінами веб-кроулер, пошуковий робот або павук. Його застосовують для обходу сторінок Інтернету з метою збору необхідної інформації. Павуки є найважливішим елементом будь-якої пошукової Інтернет системи. Вони здійснюють загальний пошук інформації в Інтернеті, повідомляють про зміст знайденого документа, індексують його і добувають підсумкові дані. Вони також переглядають заголовки, деякі посилання і відправляють проіндексовану інформацію до бази даних пошукового механізму.

Найпростішим алгоритмом збирання даних з соціальної мережі є алгоритм Breadth-first-search (BFS), який відомий як пошук у ширину в графі. Алгоритм розпочинає роботу з стартової вершини, знаходить перших сусідів цієї вершини і поміщає їх у чергу з типом організації даних (First In, First Out) FIFO. Вершини черги відвідуються в порядку появи. BFS дає можливість обходити граф по рівням, а також можливість задати необхідну їх кількість, таким чином контролюючи глибину обходу.

Архітектура BFS-кроулера соціальної мережі базується на агенті, який завантажує дані про користувача з його веб-сторінки, та черги FIFO. Він розпочинає свою роботу з ідентифікатора стартового користувача, витягуючи необхідну інформацію про нього, і отримує список ідентифікаторів його друзів. Список ідентифікаторів поступово додається до черги FIFO. Після чого з неї вибирається наступний користувач і його друзі знову додаються в чергу. При використанні BFS-кроулера соціальної мережі важливим є рівень глибини обходу, оскільки черга заповнюється на порядок швидше, ніж вивільняється, а отже необхідні значні обсяги оперативної пам'яті (від 16 Гб). Беручи до уваги теорію «шести рукостискань» (яка говорить про те, що будь-які дві людини на Землі розділені між собою не більше ніж шістьма рівнями зв'язків) для обходу усіх користувачів соціальної мережі (за умови, якщо вони не входять до замкнутих кіл) вистачить п'яти або шести рівнів глибини обходу.

Завантажену веб-кроулером інформацію про користувачів та їх зв'язки необхідно зберігати в БД, для їх подальшого аналізу. Оскільки природною репрезентацією соціальної мережі є граф – доцільно використати графову систему керування базами даних (ГСКБД), яка має графову модель збереження та обробки даних. Такою є NoSQL (NotonlySQL) ГСКБД Neo4j. На даний момент це найпоширеніша ГСКБД з відкритим програмним кодом, реалізована на Java американською компанією Neo Technology (розробка ведеться з 2003 року). Neo4j не вимагає розміщення всіх даних в оперативній пам'яті, що дозволяє її використовувати для обробки значних за розміром графів. Крім того Neo4j у повній мірі підтримує ACID (Atomicity, Consistency, Isolation, Durability – властивості, що гарантують надійну роботу транзакцій бази даних – атомарність, узгодженість, ізольованість, довговічність). Neo4j має свою орієнтовану на роботу з графами декларативну мову запитів Cypher. Запити до БД також можна виконувати використовуючи JavaAPIабо мову Gremlin.

При інсталяції Neo4j вказується робоча директорія, в якій розміщуватимуться бази даних. Neo4j спроектована працювати лише з однією базою даних. Щоб переключитися на інший граф необхідно зупинити сервер Neo4j перемкнути БД і знову запустити його.

Для створення нового вузла необхідно виконати такий запит:

```
CREATE (root:Person {name:"Bob"})
RETURN root
```

Якщо потрібно створити нові вузли із зв'язками можна виконати наступний запит:

```
MATCH (root:Person {name:"Bob"})
FOREACH (name in ["Leonardo", "Raphael", "Michelangelo",
"Donatello","Splinter"] |
CREATE (root)-[:FRIEND]->(:Person {name:name}))
```

Додавання нових зв'язків між вже створеними вузлами (результат виконання запитів див. на рис. 1):

```
MATCH (sensei:Person {name:"Splinter"}),
(ninja:Person)
WHERE ninja.name in ["Leonardo", "Raphael",
"Michelangelo", "Donatello"]
MERGE (ninja)-[:FOLLOWER]->(sensei)
```

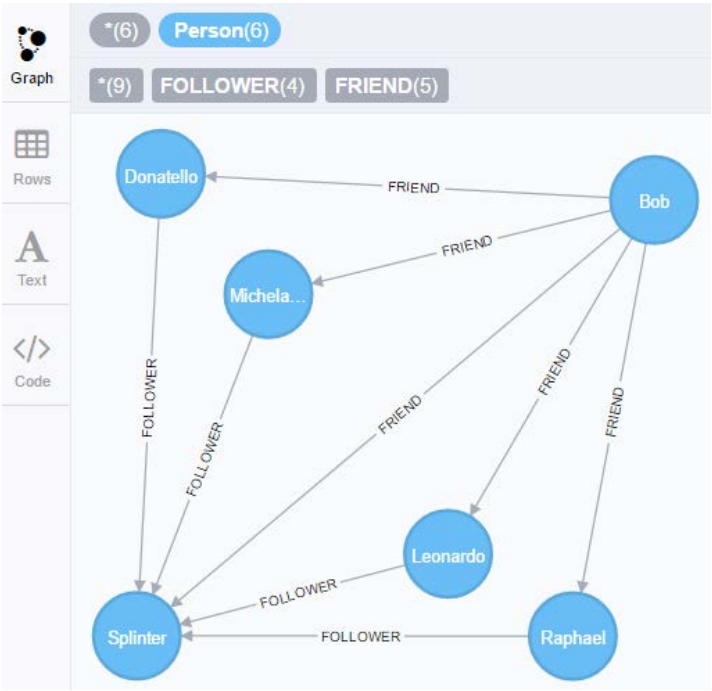

Рисунок 1 – Результати виконання запитів

Інтерфейс програмування додатків для Neo4j реалізовано для багатьох мов програмування, включаючи Java, .Net, Python, Clojure, Ruby, PHP. На великих об'ємах даних Neo4j працює набагато швидше, ніж реляційні БД і є зручною у використанні, тому вона ідеально підходить для побудови системи аналізу соціальних мереж.