Етапи розбору соціально-орієнтованого сайту

Пелещишин Андрій
Кафедра СКІД
НУ "Львівська політехніка"
Львів, Україна
apele@ridne.net

Мастикаш Олег
Кафедра СКІД
НУ "Львівська політехніка"
Львів, Україна
mastykash@itstep.org

Find social connections on the Internet is easy, however, to collect data and present them in a convenient form - a difficult task. The article considers the most popular methods of data collection from socially oriented websites. The analyzed analysis of unstructured information page user's social network. Presented scanning multilevel model of dynamic content social network.

Ключові слова: HTTP сервер, парсинг, неструктурована інформація, вихідний код сторінки

Вступ

Станом на початок 2016 року кількість зареєстрованих користувачів мережі Facebook становила 1.7 млрд, з яких 1.1 млрд кожного дня відвідують свою сторінку в Facebook. Кількість зареєстрованих користувачів Вконтакті на той самий час становила більше 392 млн. Результати дослідження компанії Gemius за січень 2016 року показують, що Інтернетом в Україні хоча б раз на місяць на ПК і/або мобільних телефонах, смартфонах і/або планшетах користуються 20,2 мільйона інтернет-відвідувачів, із яких 70% онлайн-аудиторії відвідують сайт www.vk.com. Тому взаємодія з соціальними мережами стала невід'ємною частиною життя сучасної людини.

Соціальні інтернет-мережі сьогодні займають досить важливе місце в житті активної людини. Коли люди знайомляться в реальному світі, вони обов'язково обмінюються своїми даними для того, щоб потім можна було знайти один одного в Фейсбуці або у Вконтакті. Загалом у соціальній мережі міжособистісний контакт, набагато ширший, ніж той, що ми можемо собі дозволити в реальному житті.

В соціально орієнтованих сайтах є велика кількість легкодоступної інформації, яку користувачі самі розміщають на своїх сторінках. Соціальні мережі, форуми, портали, блоги є корисним джерелом отримання даних, що містять і інформацію для ідентифікації особи, і додаткові дані про інтереси, сімейний стан, освіту, коло спілкування тощо [1].

Особливості аналізу соціальних мереж

Особливості аналізу соціальних мереж полягають в тому, що інформація у кожній соціальній мережі є динамічною та неструктурованою. І для отримання даних потрібно або відправляти специфічні запити на HTTP сервер, або використовувати API сайту [2]. Відповіді сервера, які є повідомленнями зі сторінок користувачів, можна зберегти і базу даних. Ще однією складністю при розробці програми є особливість поведінки користувачів, яка полягає в тому, що близько половини усіх користувачів соціальної мережі закривають свої сторінки для неавторизованих користувачів. Для обходу цієї проблеми потрібно виконувати додаток від імені зареєстрованого користувача соціальної мережі. Однак, при надмірній активності користувач блокується сервером соціальної мережі, тому перед переходом на наступну сторінку потрібно робити паузу, що помітно знижує швидкість роботи. Щоб прискорити роботу, можна використовувати багатопоточність, де кожен потік звертається до сервера як окремий зареєстрований користувач, а список сторінок розподіляється між потоками[3].

Під час сканування сторінки користувача можна отримати неструктуровану інформацію великого об'єму. Для структурування отриманої інформації потрібно розробити складну архітектуру збереження даних. Вирішення завдання аналізу динамічного стану соціальної мережі потрібно побудувати модель сканування, що складається із декількох рівнів[4]:

На першому рівні модель включає модель даних соціальної мережі і бібліотеку для отримання даних зі сторінок користувачів.

На другому рівні модель включає засоби, що дозволяють на базі даних першого рівня отримати аналіз характеристик користувачів і матеріалів з урахуванням їх динаміки.

На третьому рівні модель включає засоби, що дозволяють на базі даних другого рівня отримати дані про користувачів, спільнотах і матеріалах, узагальнені з точки зору масштабів всієї мережі.

Четвертий рівень включає моделі і відповідні алгоритми, що визначають шляхи і способи сканування мережі.

Кожна із наступних моделей доповнює іншу, та дозволяє скласти детальних портрет користувача. Побудова портеру поведінки користувача дасть змогу визначити уподобання користувача на основі його поведінки в соціальних мережах.

ПАРСИНГ СТОРІНКИ СОЦІАЛЬНОЇ МЕРЕЖІ

Роботу системи парсингу (розбору) веб-сторінки можна розбити на три універсальних етапи [5]:

- ***Отримання вихідного (HTML) коду сторінки***. У різних мовах для цього передбачені різні способи, наприклад, в PHP використовується бібліотека cURL (або функція file_get_contents), в C# реалізовані класи HttpWebRequest, HttpWebResponse, WebClient (належать до простору імен System.Net) [6].
- ***Фільтрація та обробка даних***. Отримавши сторінку, необхідно обробити її – очистити від тегів та невалідного коду, витягнути потрібний контент, вилучивши дані які нас не цікавлять, структурувати отримані дані. Можна, звичайно ж, використовувати для цього регулярні вирази, або скористатися простішим шляхом - використати спеціалізовані бібліотеки.
- ***Збереження результату***. Останній етап – це формування звіту. Результати переважно зберігаються в реляційній базі даних, текстових документах, на інших сайтах тощо.

Даний алгоритм можна застосувати тільки до одної веб-сторінки сайту. Для повного аналізу цілого сайту потрібно отримати посилання на усі сторінки. Це реалізується шляхом рекурсивного проходу по внутрішніх посиланнях ресурсу.

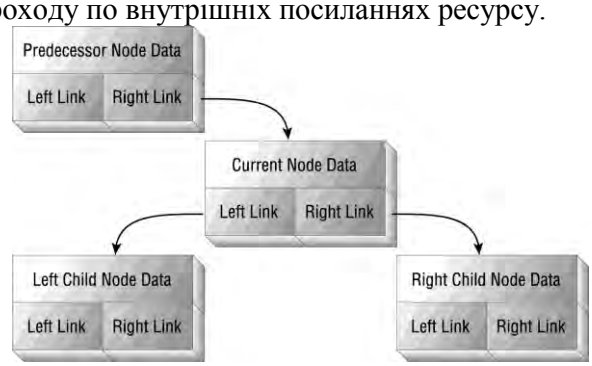

Рис. 1. Рекурсивний обхід вузлів ресурсу

ЛІТЕРАТУРА

[1] SERENA, F. David. Social network graph inference and aggregation with portability, protected shared content, and application programs spanning multiple social networks. U.S. Patent No 9,536,268, 2017.

[2] HOTHO, Andreas; JÄSCHKE, Robert; LERMAN, Kristina. Mining social semantics on the social web. Semantic Web, 2017, 8.5: 623-624.

[3] ASHA, K. N.; RAJKUMAR, R. Survey on Web Mining Techniques and Challenges of E-commerce in Online Social Networks. Indian Journal of Science and Technology, 2016, 9.13.

[4] NAIR, Rahul, et al. Computerized systems and methods for generating a dynamic web page based on retrieved content. U.S. Patent Application No 15/190,412, 2016.

[5] DESHPANDE, Mayur Venktesh, et al. *System for reconfiguring a web site or web page based on real-time analytics data*. U.S. Patent No 8,880,996, 2014.

[6] ХАНЕНКО, О. А. Огляд можливостей Python як засобу для створення Веб-парсера. *Міжнародний науковий журнал*, 2016, 6 (2):