Власюк В. В., магістрант

Кафедра інформаційних систем

Одеський національний політехнічний університет

АНАЛІЗ АЛГОРИТМУ СКАНУВАННЯ ВЕБ РЕСУРСІВ ДЛЯ ВЕРИФІКАЦІЇ РЕКЛАМНИХ ВІДЕО КАМПАНІЙ В ІНТЕРНЕТІ

В роботі проаналізовано та досліджено алгоритм сканування веб ресурсів для верифікації рекламних відео кампаній та запропоновані методи покращення алгоритму.

Ключові слова: верифікація рекламних відео кампаній, методи боротьби з підробкою доменів, алгоритм сканування веб ресурсів.

Постановка проблеми та мета роботи. Разом з розвитком інтернет ресурсів та рекламного бізнесу, великої популярності набирає шахрайство в мережі особливо пов’язаних з підробкою доменних імен. Запит може бути дуже легко підробленим на шляху від веб сайту до рекламного агентства і реклама насправді не є показаною цвілевій аудиторії. Метою роботи є дослідження та запропонування рекомендацій що до використання алгоритму веб сканування для верифікації рекламних відео кампаній.

Основна частина роботи. Для основи системи верифікації найбільше підходить схема багатопоточного сканування яка застосовується в більшості популярних рішень. Така система дозволяє обробляти домен с декількома запитами по різним вузлам, що збільшує швидкість обробки. Так як дана система має працювати з великою кількістю доменів, розподілення обробки доменів на серверні потужності зменшить час обробки та розподілить навантаження. Найбільш простий спосіб поліпшення продуктивності та ефективності сканерів - це паралельні обчислення [1]. Веб-сканери мають справу з декількома проблемами одночасно, деякі з яких суперечать одна одній, тому їхня реалізація має більш складну структуру з компонентами які вирішують більшість проблем стандартних рішень [2]. Нижче представлений алгоритм сканування веб ресурсів який пропонуються для основи системи верифікація відео реклами (Рис. 1).

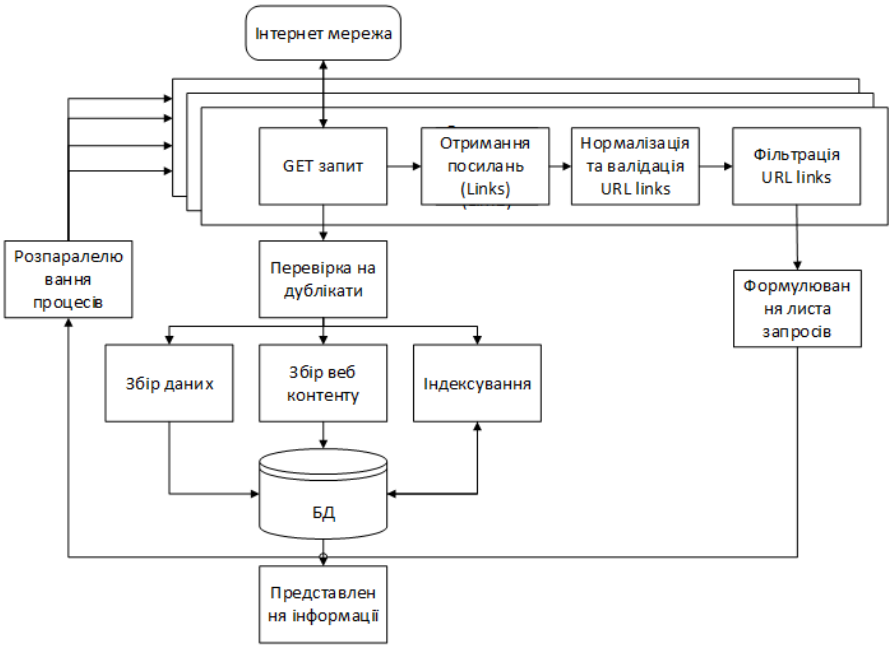

Рисунок 1 – Алгоритм сканування веб ресурсів

Для вирішення задач верифікації до даного алгоритму має бути доданий блок сканування сторінок а також покращені блоки паралелізації процесів та формулювання листів посилань.

Для блоку паралелізації процесів запропоновано використання покращеного алгоритму проходу графа в ширину на основі підходу Мунгала та Ранадей [3]. Застосування цього алгоритму дає змогу використання всіх потужностей апаратного забезпечення з рівномірним навантаженням.

Блок сканування сторінок та перевірки дублікатів запропоновано виконувати в паралельних потоках. Сканування буде відбуватись з застосуванням cookie сесій та автоматизованого забезпечення Selenium [4], що дасть змогу запускати всі мережеві запити на сторінці та контролювати сторінку.

Блок формування листів пропонується побудувати на основі фільтрування дублікатів [5] та принципів пріоритезації за ключовими словами пов’язаними з відео плеєрами, відео контентом, рекламою, інформацією про сайт та інші корисні сторінки. Даний підхід дасть змогу оптимізувати сканування та отримувати результати з меншої кількості сторінок.

Висновки. Проведений аналіз та огляд схеми алгоритму сканування веб ресурсів для верифікації рекламних відео кампаній. Запропонований метод

проходу графу в ширину який підтримує паралельні процеси, що зменшує кількість кроків алгоритму та час виконання. Винесені пропозиції що до оптимізації початкової схеми алгоритму з додаванням блоку сканування та використанням пріоритезації в блоці формування листів.

Керівник магістерського дослідження к.т.н., доцент кафедри ІС Галчонков О.М.

Література

1. С. Кастільо, Ефективне сканування в Інтернеті , ACM SIGIR Forum, vol. 39, ACM– 2011, – 55-56 с.

2. Е. Феррара., Видобування веб-даних, програми та методи: огляд Knowl.-Based Syst. / [Е. Феррара, П. де Meo, Г. Фіумара, Р. Баумгартнер].–2014, 301 с.

3. I/O-Complexity of Graph Algorithms, [Kameshwar Munagala, Abhiram Ranade] – Department of Computer Science and Engineering Indian Institute of Technology Bombay – 2014 – 8с.

4. Selenium WebDriver [Електронний ресурс] – документація. – Режим доступу: http://www.seleniumhq.org/docs/03_webdriver.jsp

5. Виявлення біля дублікатів для сканування в Інтернеті, [Г. С. Манку, А. Джайн, А. Дас Сарма]. – Праці 16 Міжнародної конференції "Всесвітня мережа", ACM, 2007. –141-150с.