ПОБУДОВА ВЕБ-ГРАФА МЕТОДОМ КРАУЛІНГА

Матющенко О.Д., студент групи ПА-16-2,

Гук Н.А., доктор фізико-математичних наук, професор

Дніпровський національний університет ім. Олеся Гончара

Abstract. A way of constructing the internal structure of a web site in the form of a web graph was considered, software was developed to perform the procedure of crawling the site and establishing semantic links between its pages.

Keywords: WEBSITE, HTML-PAGE, WEB-GRAPH, URLS, WEBSITE, CRAWLER, INTERNAL STRUCTURE OF THE WEBSITE, HYPERLINKS, ALGORITHM.

Сьогодні практично всі організації мають своє представництво у мережі Інтернет у вигляді web-сайтів. Зростання обсягів інформації, розташованої на сайті, відбувається дуже швидко, тому з часом логічна побудова ресурсу може порушуватися. Зараз аналіз web-сайтів з метою контролю логічного зв'язку їх сторінок відбувається із застосуванням графової моделі. Вершинами web-графу являються html-сторінки сайту, а ребрами – гіперпосилання між сторінками. Із використанням такої структури можливо відстежувати логічність зв'язків між сторінками, виконувати кластерізацію сторінок за типами, будувати траєкторії пошукових запитів. Тому розробка методів побудови web-графів є актуальною задачею.

Одним з інструментів для побудови web-графу сайту є краулер. Краулер (англ. Crawler) – програма, яка виконує обхід сайту, здійснюючи перехід з однієї веб-сторінки на іншу, та використовує метод пошуку в ширину. Відповідний процес обходу web-сайту будемо називати краулінг (англ. Crawling). Обхід сайту починається з головної сторінки, яка ідентифікується за доменним ім'ям, граф має багаторівневу структуру, рівень сторінки визначається кількістю гіперпосилань, які необхідно пройти на шляху до неї, починаючи від головної сторінки. Якщо записати шлях переходу по сторінках веб-ресурсу, то вийде орієнтований граф ресурсу.

При побудові програми для краулінгу сайтів застосовано наступні правила:

1. URL-адреси сторінок сайту повинні бути нормалізовані.

2. Програма повинна слідувати правилам, зазначеним у robots.txt.

3. Обхід сторінок відбувається з головної сторінки, яка стає вершиною графа.

4. Програма повинна ігнорувати зовнішні гіперпосилання та посилання на протоколи.

5. Сторінки з однаковим змістом, але різними URL потрібно вважати різними сторінками.

6. Необхідно ставити невеликі паузи між відправленням пакетів даних.

Після виконання обходу всіх сторінок сайту можна побудувати граф для візуалізації структури web-додатку.

Відповідне програмне забезпечення розроблено із використанням мови програмування Python та бібліотеки Scrapy, яка є високорівневим фреймворком для перегляду web-сторінок з метою пошуку структурованих даних.

Із застосуванням розробленого програмного забезпечення виконано процедуру краулінгу сайту факультету прикладної математики Дніпровського національного університету ім. О. Гончара (fpm.dnu.dp.ua).

Для візуалізації отриманої інформації про структуру сайту використовується додаток з графічним інтерфейсом Gephi, за допомогою якого можливо виконати укладку графу для спрощення його візуального сприйняття.

Запропонований підхід для побудови графової моделі web-сайту можна використовувати для вивчення структури з метою її поліпшення, пошуку найбільш затребуваної інформації, опису моделі поведінки користувача сайту, кластерізації сторінок за типами, будувати власні рекомендаційні системи для користувачів.

References

1. Brin S., Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine // Computer Science Department, Stanford University, Stanford — 1998. - С. 107-117.

2. Kobayashi, M. & Takeda, K. (2000). "Information retrieval on the web". ACM Computing Surveys. 32 (2): 144–173 .

3. A. Gulli; A. Signorini (2005). "The indexable web is more than 11.5 billion pages". Special interest tracks and posters of the 14th international conference on World Wide Web. ACM Press. pp. 902–903.