ВЕБ-КРОУЛІНГ ЯК ЕТАП РЕАЛІЗАЦІЇ ПРОЦЕСУ ЗБОРУ ДАНИХ В МЕРЕЖІ ІНТЕРНЕТ

Чередніченко О.Ю., Янголенко О.В., Матвєєв О.М., Мозгін В.В.
Національний технічний університет «Харківський політехнічний інститут», м. Харків

Для отримання кількісних значень показників діяльності, використовують засоби веб-моніторингу. Першим етапом реалізації процесу веб-моніторингу є пошук джерел даних, а саме пошук веб-сторінок, на яких знаходяться дані, необхідні для вимірювання показників.

Пошук і збір веб-сторінок здійснюється засобами веб-кроулінгу (webcrawling), для їхнього подальшого індексування та підтримки функціонування пошукової системи. Метою обходу є швидкий та ефективний збір якомога більшої кількості корисних веб-сторінок разом з посиланнями, що їх об’єднують. Обхід здійснюється пошуковим роботом (веб-кроулером). Отримуючи стартову URL адресу, з якої розпочинається обхід, пошуковий робот завантажує відповідну веб-сторінку, видобуває із неї усі вихідні посилання та додає їх у чергу для подальшого обходу. Цей процес продовжується, поки черга не залишається порожньою. Ключовим елементом функціонування пошукового роботу, який впливає на його ефективність, є стратегія обходу посилань, які зберігаються у черзі. Найчастіше застосовуються два підходи: сліпий пошук та евристичний підхід. Під час сліпого пошуку при виборі наступного URL з черги на завантаження не використовується ніякий критерій: посилання для обходу вибираються у порядку їхнього розташування в черзі. Евристичний підхід поданий алгоритмами, основаними на певному критерії вибору наступного посилання із черги для обходу.

У цьому дослідженні пропонується удосконалений за рахунок використання формальної архітектури агента алгоритм тематичного направленого пошуку веб-сторінок, які є джерелами даних для моніторингу Тематичний направлений пошук являє собою інформаційний пошук, якому притаманні такі характеристики: простір пошуку невідомий заздалегідь; не визначено чітко інформаційну потребу користувача; існує велика кількість нерелевантних документів; неконтрольована якість відібраних документів. Оскільки значна частина посилань є нерелевантною, тобто не містить даних, необхідних для вимірювання показників моніторингу, в роботі запропоновано здійснювати оцінку перспективності веб-сторінки для подальшого пошуку, а потім на її основі приймати рішення щодо подальшого обходу посилань, які містяться на цій сторінці. Пошук продовжується, якщо веб-сторінка оцінена позитивно. Якщо веб-сторінка отримала негативну оцінку, то пошук у напрямку посилань, які на ній містяться, не здійснюється.

Таким чином, запропоновано алгоритм тематичного інформаційного пошуку джерел даних веб-моніторингу, який базується на використанні формальної архітектури агента та надає можливість здійснювати пошук відповідно до запиту системи управління.