МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет «Запорізька політехніка»

Факультет комп'ютерних наук та технологій
Кафедра «Комп'ютерні системи та мережі»

ФЕДОРОВ ВІТАЛІЙ АНАТОЛІЙОВИЧ
Група КНТ-513м

СИСТЕМА АВТОМАТИЗАЦІЇ ПРОЦЕСІВ ЗБОРУ ТА АНАЛІЗУ ІНФОРМАЦІЇ З ВЕБСАЙТІВ

АВТОРЕФЕРАТ

магістерської роботи на здобуття освітньо-кваліфікаційного рівня «магістр» 123 «Комп'ютерна інженерія» освітньої програми «Комп'ютерні системи та мережі»

2024 р.

Магістерська робота є рукопис.

Робота виконана в Національному університеті «Запорізька політехніка», на кафедрі комп’ютерних систем та мереж

Керівник кандидат технічних наук, доцент **Ільяшенко Матвій Борисович,** Національний університет «Запорізька політехніка», доцент кафедри комп'ютерних систем та мереж

Офіційний рецензент: **Малий Олександр Юрійович,** к.т.н., доцент кафедри «Інформаційні технології електронних засобів» НУ «Запорізька політехніка»;»

Захист відбудеться "23" грудня 2024р.

Секретар екзаменаційної комісії, доцент кафедри комп’ютерних систем та мереж Т.В. Голуб

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. У сучасному світі, де інтернет став невід'ємною складовою нашого повсякденного життя, і майже кожна частина світу має доступ до глобальної мережі, виникає все більше нових викликів, які потребують вирішення. Одним із таких викликів є задача збору та аналізу інформації. Оскільки інтернет містить величезні обсяги даних, актуальним стає питання отримання впорядкованої та навіть базово проаналізованої інформації без безпосередньої участі людини в цьому процесі.

Рішення для автоматизованого збору та аналізу інформації дійсно існують, однак більшість програмних застосунків, що вирішують ці завдання, мають суттєвий недолік — відсутність гнучкості. Це означає, що коли структура веб-сторінки змінюється, процес автоматизації збору та аналізу інформації зазвичай припиняється. Але проблема негнучкості полягає не лише в цьому. Навіть якщо розробники стверджують, що їхня програма універсальна, часто це є помилковим твердженням, що вводить користувачів в оману.

Більшість таких програм являють собою велику кількість спеціалізованих синтаксичних аналізаторів, розроблених для конкретної групи веб-сайтів. Коли ж користувач потребує аналізу даних із менш відомого ресурсу, він стикається з тим, що цей ресурс не підтримується системою, і процес аналізу стає неможливим без втручання розробників. Вони можуть вирішити проблему шляхом створення нового парсера, хоча такий парсер лише частково інтегрується у "універсальну" систему.

У випадках, коли програмний застосунок дійсно є універсальним, він, як правило, представляє лише базову структуру парсера. Однак такий застосунок зазвичай не надає користувачу можливості створювати власні запити для збору інформації, що відповідають його потребам, без залучення фахівців або розробників. У більшості випадків для цього необхідно звертатися до спеціалістів, які створюють нові синтаксичні аналізатори або налаштовують запити в межах програми.

Таким чином, хоча на ринку дійсно існують рішення для автоматизації збору інформації з веб-сайтів, вони або не є по-справжньому універсальними, або не надають користувачу

можливості самостійно створювати й налаштовувати запити до ресурсів, які його цікавлять, без залучення додаткових спеціалістів.

У ситуаціях, коли програмний застосунок дійсно позиціонується як універсальний, насправді він зазвичай є лише основним каркасом для парсера. Однак такий застосунок часто не дозволяє користувачу самостійно формувати індивідуальні запити для збору даних, які б задовольнили його конкретні потреби, без залучення розробників або фахівців. У більшості випадків для виконання таких завдань необхідно звертатися до професіоналів, які створюють нові синтаксичні аналізатори або налаштовують необхідні запити в рамках наявної програми.

Отже, незважаючи на те, що на ринку існують рішення для автоматизації збору інформації з веб-сайтів, вони або не є повністю універсальними, або не пропонують користувачу можливості самостійного створення й налаштування запитів до ресурсів, які його цікавлять, без потреби у залученні додаткових фахівців.

Мета і завдання дослідження. Мета та завдання дослідження полягають у створенні системи автоматизованого збору та аналізу інформації. Основними цілями є пошук аналогів на ринку та виявлення їх недоліків. Також передбачено дослідження варіантів використання синтаксичних аналізаторів для аналізу веб-сайтів та оцінка релевантності їх застосування. Планується вивчити етапи роботи подібних систем на ринку та досягти універсальності, пропонуючи не готовий застосунок, а конструктор запитів, який орієнтується на звичайних користувачів без професійних навичок.

У процесі дослідження було розглянуто питання переваг і недоліків програмних продуктів, що реалізують аналогічний функціонал. В результаті було розроблено програмний застосунок, основна ідея якого полягає в максимальній універсальності та багатофункціональності, щоб задовольнити якомога більше запитів звичайних користувачів. Цей програмний продукт призначений для надання можливості користувачам поетапно створювати запити, які в подальшому будуть автоматично виконувати пошук, аналізувати та зберігати дані в заданому користувачем форматі з обраного веб-сайту.

Об’єктом дослідження – є система автоматизованого збору та аналізу інформації з веб-сайтів. Зокрема, це стосується процесу розробки такої системи, механізмів взаємодії програмного

застосунку з веб-сайтами, а також принципів обміну даними між веб-сайтом і веб-сервером. У ході дослідження розглядаються різні методи представлення інформації користувачеві, які використовують веб-сайти, а також вимоги до користувача, необхідні для забезпечення коректної роботи веб-сайту та надання послуг. Окрему увагу приділяється сценаріям, у яких веб-сайт може не надати потрібну користувачеві інформацію через певні технічні або програмні обмеження.

Предмет дослідження – є визначення характеристик програмного забезпечення, створеного за допомогою автоматизованого конструктора запитів, а також аналіз особливостей типової програми для автоматизованого збору та аналізу даних з веб-сайтів.

Наукова новизна отриманих результатів. Наукова новизна результатів цього дослідження полягає в тому, що під час розробки системи автоматизованого збору та аналізу інформації було досягнуто реальної універсальності застосунку. Основна особливість запропонованої системи полягає в її здатності гнучко адаптуватися до різних веб-сайтів, незалежно від їхньої структури чи змін у HTML-розмітці. Це забезпечується алгоритмом, який дозволяє програмному застосунку автоматично підлаштовуватись під різноманітні формати та структури веб-сторінок, усуваючи необхідність постійного залучення розробників для оновлення парсерів або налаштувань. Такий підхід надає можливість більш ефективно й автоматизовано отримувати та аналізувати дані з різних веб-сайтів, що раніше вимагало втручання фахівців.

Практичне значення отриманих результатів:

Практичне значення одержаних результатів полягає в розробці програмного застосунку для автоматизованого збору та аналізу інформації з веб-сайтів, який вирішує основні проблеми існуючих на ринку рішень. На відміну від існуючих систем, які є вузькоспеціалізованими та не мають належної гнучкості, запропонований програмний застосунок надає користувачам можливість продовжувати роботу навіть у випадку змін у структурі веб-сайтів. Це означає, що користувач не втрачає доступ до аналізованої інформації при зміні HTML-розмітки веб-сторінок і не потребує залучення розробників для оновлення чи налаштування програмного забезпечення. Це значно зменшує час і витрати на

підтримку роботи таких систем, роблячи їх більш універсальними та зручними у використанні.

Апробація результатів магістерської роботи. Основні положення магістерської роботи та результати досліджень подано до участі на конференції:

- Система управління потоками подій в режимі реального часу за допомогою інтернету речей / Федоров В.А., Шевченко Д.О., Ільяшенко М.Б., Куликовська Н.А. // Міжнародній науково-практичній конференції студентів, аспірантів та молодих вчених «Молодіжна наука: інновації та глобальні виклики» Національного університету «Полтавська політехніка імені Юрія Кондратюка» , 6 листопада 2024 року.

Структура та обсяг роботи. Магістерська робота складається зі вступу, чотирьох розділів, висновків, списку використаних джерел, додатку. Основна частина містить 75 сторінок, 31 рисунок, 7 лістингів, список використаних джерел зі 24 найменування.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У першому розділі проведено аналіз проблеми створення комп'ютерної системи для автоматизованого збору та аналізу даних. Досліджено, як користувач взаємодіє з веб-браузером та веб-сайтом під час збору інформації. Основну увагу приділено ключовим етапам, поняттям і принципам, що використовуються для цієї взаємодії. Визначено важливі фактори, які впливають на функціонування програмного забезпечення, що забезпечує автоматизований збір і аналіз інформації з веб-сайтів. Також проведено аналіз існуючих рішень.

На основі цього були сформульовані наступні висновки:

- визначення ролі програмного забезпечення для збору та аналізу є критично важливим для його розробки, адже цей аспект впливає на успішність процесу створення програмного продукту;

- досліджено, як користувач взаємодіє з веб-сайтом, з детальним розглядом усіх аспектів, які будуть використовуватися при розробці програмного застосунку;

- проведено аналіз етапів взаємодії між веб-браузером і веб-сайтом;

- сформовано етапи роботи програмного застосунку, що здійснює автоматизований збір і аналіз даних, імітуючи взаємодію між веб-браузером і веб-сервером.

- розглянуто, яку інформацію веб-сайт надає у відповідь на запити, а також можливі варіанти відповідей і їх вміст.

- проаналізовано різні типи веб-сайтів, такі як динамічні і статичні, з урахуванням їхнього впливу на роботу програмного застосунку.

- оцінено готові рішення, що реалізують автоматизований збір та аналіз даних з веб-сайтів.

У другому розділі було проведено аналіз інструментів, технологій і бібліотек, що використовуються для створення комп'ютерних систем автоматизованого збору та аналізу інформації з веб-сайтів. На основі проведеного аналізу можна зробити наступні висновки:

- Python є оптимальним вибором для швидкої та зручної розробки систем автоматизованого збору і аналізу, оскільки бібліотеки надають готові реалізації процесів, що спрощує створення застосунків;

- для розробки десктопних додатків, що здійснюють автоматизований збір та аналіз даних, доцільно використовувати мову C#, оскільки вона забезпечує необхідні засоби для створення користувацького інтерфейсу та містить функціонал для базових процесів взаємодії з веб-сайтами;

- для полегшення розробки програмного забезпечення рекомендується використовувати готові бібліотеки та інструменти;

- ознайомлено і вивчено основні інструменти, які пропонує .NET, для створення базового функціоналу програмного застосунку збору та аналізу даних з веб-сайтів.

У третьому розділі У цьому розділі були реалізовані програмні застосунки як для статичного, так і для динамічного автоматизованого збору та аналізу даних з веб-сайтів. Проведено аналіз функціональних і нефункціональних вимог до програмних застосунків, а також побудовано діаграми класів, діаграми використання, діаграми послідовностей тощо. На основі виконаних дій можна зробити наступні висновки:

- реалізація програмних застосунків - успішно реалізовано програмні застосунки для статичного та динамічного автоматизованого збору і аналізу даних;

- діаграми класів - створено та описано діаграми класів для застосунків статичного і динамічного автоматизованого збору та аналізу інформації;

- використання HtmlAgilityPack - для побудови DOM-об'єкта та взаємодії з ним був застосований пакет HtmlAgilityPack;

- функціональність побудови запитів - програмний застосунок забезпечує можливість формування запитів;

- користувацький інтерфейс - інтерфейс програмного застосунку є простим та зрозумілим для користувачів;

- зберігання даних у форматі JSON - програмний застосунок використовує JSON-файл для зберігання даних.

У четвертому розділі було проведено дослідження різних факторів, що впливають на ефективність роботи програмного застосунку. Для покращення результатів його функціонування були запропоновані кілька варіантів подальшого розвитку.

ВИСНОВКИ

В ході виконання дипломної роботи був спроектований і реалізований програмний застосунок для автоматизованого збору та аналізу інформації з веб-сайтів. На основі проведених досліджень можна зробити такі висновки:

- досліджено проблеми, пов'язані з розробкою програмних застосунків для автоматизованого збору та аналізу інформації;

- проведено теоретичні дослідження в галузі систем автоматизованого збору та аналізу даних;

- вивчені засоби, бібліотеки та інструменти, що можуть бути використані для вирішення поставленої задачі;

- реалізовано та проведено перевірку працездатності програмного застосунку для автоматизованого збору та аналізу інформації з веб-сайтів;

- визначено переваги та недоліки розробленого рішення для універсального автоматизованого збору та аналізу даних;

- сформульовано варіанти покращення програмного застосунку.