МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЧОРНОМОРСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ПЕТРА МОГИЛИ

Кащенко Дмитро Олегович

УДК 004.89

ДОСЛІДЖЕННЯ ХАРАКТЕРИСТИК ВЕБ-ЗАСТОСУНКІВ ВНЗ ЗА ДОПОМОГОЮ РОБОТА-ПАРСЕРА

124 – Системний аналіз

Автореферат
магістерської наукової роботи на здобуття освітньої кваліфікації
«Магістр системного аналізу»

Миколаїв – 2020

Магістерська наукова робота є рукопис.

Робота виконана в Чорноморському національному університеті імені Петра Могили Міністерства освіти і науки України на кафедрі інтелектуальних інформаційних систем

Науковий керівник: д.п.н., професор
інтелектуальних інформаційних систем
Мещанінов Олександр Павлович

Рецензент: к.ф-м.н, доцент кафедри
інженерії програмного забезпечення
Пузирьов Сергій Володимирович

Захист відбудеться «27» лютого 2020 р. о 9^{00} год. на засіданні екзаменаційної комісії (ауд. 2-403) у Чорноморському національному університеті імені Петра Могили за адресою: 54003, м. Миколаїв, вул. 68-ми Десантників, 10.

З магістерською науковою роботою можна ознайомитися в бібліотеці Чорноморського національного університету імені Петра Могили за адресою: 54003, м. Миколаїв, вул. 68-ми Десантників, 10.

Автореферат представлений «23» лютого 2020 р.

Секретар
екзаменаційної комісії,
к.пед.н., доцент Н. М. Болюбаш

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність дослідження визначається відсутністю існуючих рішень по дослідженню аналогічних веб-застосунків вищих навчальних закладів

Метою роботи є аналіз веб-застосунків ВНЗ, дослідження їх відмінностей, і побудова рекомендацій для впровадження на цільовий веб-застосунок.

Об'єктом дослідження є процес аналізу веб-застосунків ВНЗ і інструменти для їх аналізу.

Предметом дослідження є розробка моделі робота-парсера для автоматичного збору інформації і аналізу веб-застосунків, та дослідження їх характеристик

Практичне значення даної магістерської наукової роботи полягає у можливості автоматичного дослідження змісту веб-застосунків і створенні рекомендацій для покращення роботи цільових веб-застосунків..

Магістерська наукова робота складається із вступу, 4 розділів, висновків, додатків. Загальний обсяг роботи складає 83 сторінки, , 3 таблиць та 35 посилань на літературні джерела.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі магістерської наукової роботи обґрунтовано актуальність обраної теми, сформульовано мету і задачі дослідження, визначено предмет та об’єкт дослідження.

У першому розділі були досліджені функції веб-застосунків вищих навчальних закладів, їх особливості. Були описані функції Webometrics, і особливості його рейтингової системи. Крім того були досліджені веб-застосунки в рейтингу Webometrics, і особливості декількох веб-застосунків в порівнянні з цільовим сайтом.

У другому розділі досліджене поняття роботів парсера. Для чого вони необхідні, сфери їх застосування. Були дослідженні основні представники роботів парсерів і бібліотеки парсингу. Були також досліджені бібліотеки обробки даних, насамперед кластеризації даних, серед яких були ELKI, Cytoscape, та Commons Math.

У загальному сенсі, парсинг - це лінійне зіставлення послідовності слів з правилами мови. Поняття «мова» розглядається в самому широкому контексті. Це може бути нормальною мовою (наприклад, український), який використовується для комунікації людей, а може бути і формалізована мова, зокрема, мова програмування.

Парсинг сайтів – це послідовний синтаксичний аналіз інформації, розміщеної на інтернет-сторінках. Текст веб-сторінок –це ієрархічний набір даних, структурований за допомогою людських і комп'ютерних мов. Людською мовою надана інформація, або знання, заради яких люди, власне, і користуються Інтернетом. Мови програмування веб-сторінок (HTML, JavaScript, CSS) визначають як інформація виглядає на моніторі.

Парсинг потрібнен тоді, коли необхідно швидко отримати і зберегти в структурованому вигляді будь-які дані з інтернету. Парсинг сайтів - це новий метод введення даних, який не вимагає повторного введення або копіпастінга.

Такого роду програмне забезпечення шукає інформацію під контролем користувача або автоматично, вибираючи нові або оновлені дані і зберігаючи їх в

такому вигляді, щоб у користувача був до них швидкий доступ. Наприклад за допомогою парсингу можна зібрати інформацію про продукти і їх вартості на сайті Amazon.

Серед основних функцій парсерів виділяють наступні:

1. Збір даних для дослідження ринку

Веб-сервіси для збору даних допомагають стежити за ситуацією в тому напрямку, куди буде прагнути компанія або галузь в наступні шість місяців, забезпечуючи потужний фундамент для дослідження ринку. Програмне забезпечення парсинга здатне отримувати дані від безлічі провайдерів, що спеціалізуються на аналітиці даних і у фірм з дослідження ринку, і потім зводити цю інформацію в одне місце для референції і аналізу.

- Великі обсяги.

В епоху бурхливого зростання Мережі і жорстокої конкуренції всім ясно, що успішний веб-проект немислимий без розміщення великої кількості інформації на сайті. Сучасні темпи життя призводять до того, що контенту має бути не просто багато, а дуже багато, в кількостях, які набагато перевищують межі, можливі при ручному заповненні.

- Часте оновлення.

Обслуговування величезного потоку динамічно мінливої інформації не в силах забезпечити одна людина або навіть злагоджена команда операторів. Часом інформація змінюється щохвилини і в ручному режимі оновлювати її навряд чи доцільно. Парсинг сайтів є ефективним рішенням для автоматизації збору і зміни інформації.

2. Отримання контактної інформації

Інструменти парсинга можна використовувати, щоб збирати і систематизувати такі дані, як поштові адреси, контактну інформацію з різних сайтів і соціальних мереж. Це дозволяє складати зручні списки контактів і всієї супутньої інформації для бізнесу - дані про клієнтів, постачальників або виробників.

3. Рішення стосовно завантаження із StackOverflow

З інструментами парсинга сайтів можна створювати рішення для офлайнового використання і зберігання, зібравши дані з великої кількості веб-ресурсів (включаючи StackOverflow). Таким чином можна уникнути залежності від активних інтернет з'єднань, так як дані будуть доступні незалежно від того, чи є можливість підключитися до інтернету.

4. Пошук роботи чи робітників

Для роботодавця, який активно шукає кандидатів для роботи в своїй компанії, або для здобувача, який шукає певну посаду, інструменти парсинга теж стануть незамінні: з їх допомогою можна налаштувати вибірку даних на основі різних доданих фільтрів і ефективно отримувати інформацію, без рутинного ручного пошуку.

5. Відслідковування цін в різних магазинах

Такі сервіси будуть корисні і для тих, хто активно користується послугами онлайн-шопінгу, відстежує ціни на продукти, шукає речі в декількох магазинах відразу.

В порівнянні з ручною роботою, робот-парсер виконує роботу швидше, обережніше, і упакує дані за шаблоном яким йому вкаже розробник.

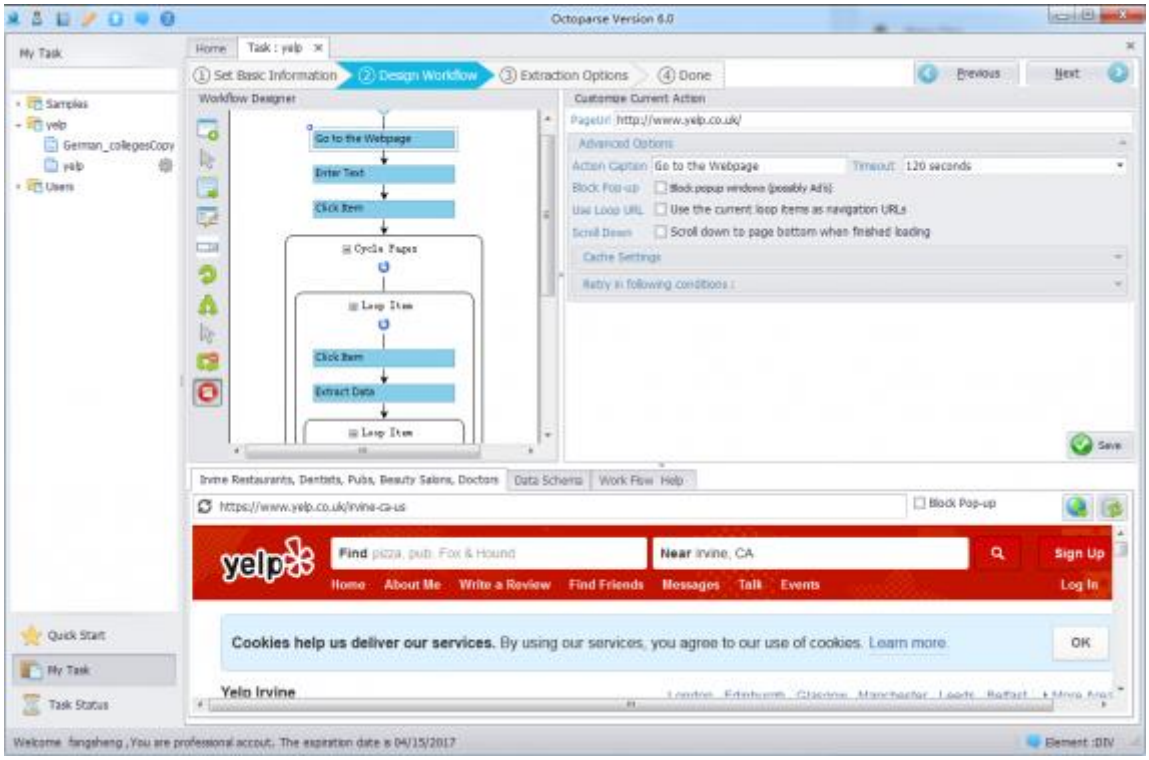
Результати (будь то база даних чи таблиця) потребуватиме обробки, але це вже інший процес.

Серед альтернативних продуктів на даний момент не існує такого продукту, що досліджує веб-застосунок цілком. Зазвичай парсери виділяють певні області для дослідження, будь то певне поле ціни, чи товарний продукт. Це необхідно в маркетингових цілях, для дослідження існуючого ринку і побудови власної маркетингової кампанії. Серед таких програмних продуктів найбільшою популярністю користуються наступні:

- **Octoparse**

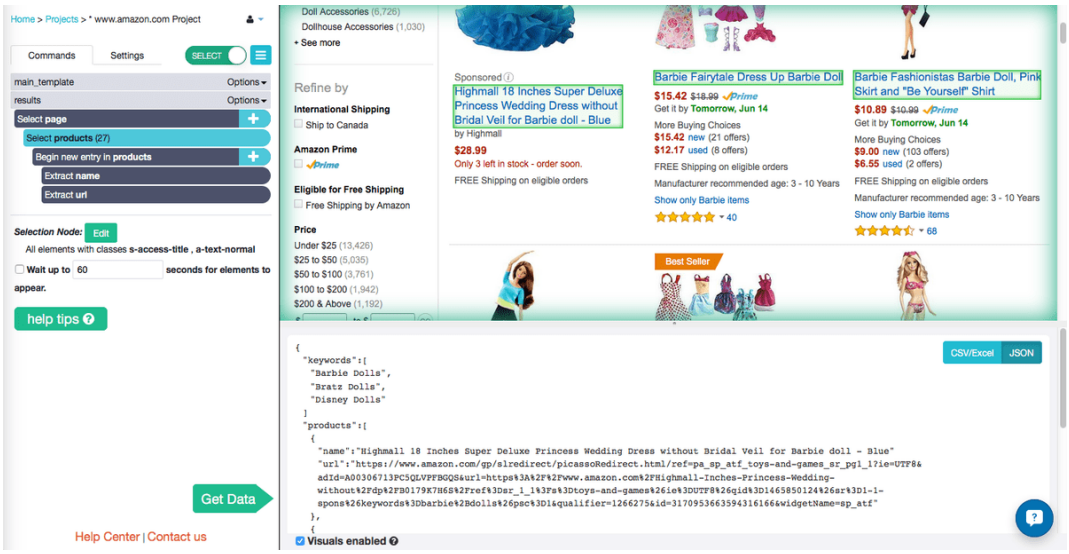
Даний парсер використовується у маркетингових цілях, і пропонується розробниками для відслідковування змін у веб-застосунках конкурентів. Серед своїх переваг він має: простий і інтуїтивний інтерфейс, можливість зберігати інформацію в хмарних сховищах, і зберегти дані на власному пристрої. Серед профільних

переваг він може змінювати власний IP-адресу, і налаштовувати періодичність парсингу. На рисунку нижче зображений інтерфейс програмного продукту.



- **ParseHub**

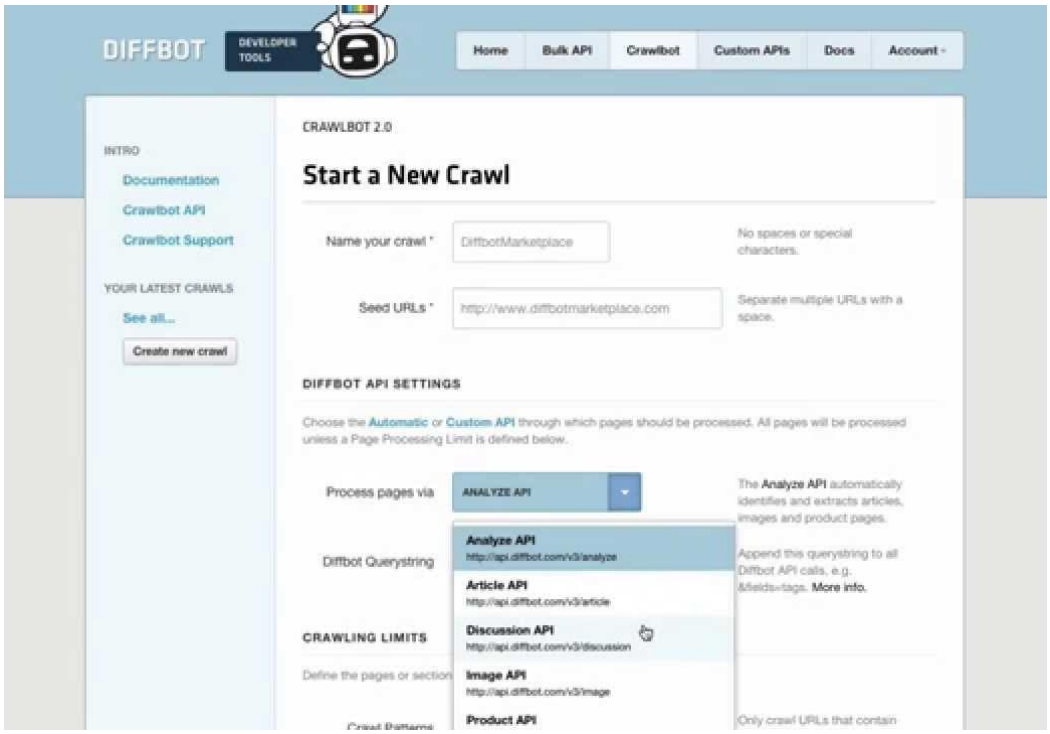
Даний парсер також використовується частіше усього у маркетингу, він має налаштування автоматичної зміни IP-адреси, періодичності запитів, він також може працювати із скриптами, і регулярними виразами. Він може парсити декілька веб-застосунків і отримувати дані текстового формату чи формату HTML з атрибутами від сторінки, що прибликує до необхідного нам функціоналу. На рисунку нижче зображений інтерфейс програмного продукту.



- **DiffBot**

Diffbot - є багатопрофільним сервісом, за допомогою якого можна навчати штучний інтелект, чи збирати інформацію з новин чи конкурентних підприємств. За допомогою нього можно відслідковувати зміни в цінах, або згадування в медіа чи інших джерелах. Нажаль він не підходить нашій цілі по функціоналу.

Отже, як можна побачити, існуючі альтернативні рішення в кращому випадку лише частково. Для виконання поставлених задач краще всього буде створити власний продукт. На рисунку нижче зображений інтерфейс парсеру Crawl що входить в програмний продукт DiffBot.



Для реалізації програмного продукту і оптимізації робочого часу було вирішено використовувати бібліотеки обробки HTML, а саме бібліотеку JSoup.

Jsoup - бібліотека з багаторічною історією, але сучасним ставленням:

- Він може обробляти старий і поганий HTML, але він також обладнаний для HTML5
- Має потужну підтримку маніпуляцій із підтримкою селекторів CSS, обходу DOM та легким додаванням чи видаленням HTML
- Він може очищати HTML, як для захисту від атак XSS, так і в тому сенсі, що покращує структуру та форматування

Для реалізації кластерного аналізу були досліджені різні бібліотеки кластеризації, і серед них були виділені наступні бібліотеки:

- *ELKI (Environment for deveLoping KDD-Applications supported by Index-structures) – ELKI* - це бібліотека з відкритим кодом (AGPLv3), яке написана на мові Java. Основний пріорітет ELKI - дослідження алгоритмів, з акцентом на непідконтрольні методи кластерного аналізу та виявлення зовнішньої структури. Щоб досягти високої продуктивності та масштабованості, ELKI пропонує структури індексів даних, такі як R*-дерево, що може забезпечити значне підвищення продуктивності. ELKI розроблений таким чином, щоб легко розширюватися для дослідників та студентів у галузі та підтримує внесок додаткових методів. ELKI має на меті створення великої колекції алгоритмів, які добре параметризуються, з метою легкої та справедливої оцінки та порівняльної оцінки алгоритмів.
- *Spark MLlib - MLlib* - це бібліотека машинного навчання Spark (ML). Його мета - зробити на практиці машинне навчання масштабованим та простим. Бібліотека пропонує такі інструменти, як: Алгоритми ML (загальні алгоритми навчання, такі як класифікація, регресія, кластеризація та спільна фільтрація), а також різні службові функції (лінійної алгебри, статистики, обробки даних), тощо.
- *Cytoscape* - це вільний, відкритий, візуальний інтерфейс для імпорту, візуального дослідження та аналізу графічних даних. Cytoscape активно підтримують розробники, які контролюють технічне обслуговування та вдосконалення основних функціональних можливостей. Cytoscape також дозволяє самостійно розробляти розширені функціональні можливості сторонніми розробниками як «плагіни». Список плагінів Cytoscape охоплює все, починаючи від пошуку даних із джерел баз даних, інтеграції та аналізу даних експресії генів, до аналізу надмірно представленої анотації гено-онтології у підграфі взаємодії. На жаль

через спеціалізацію на візуальне дослідження бібліотека Cytoscape не підходить для наших задач.

- *Commons Math* - це бібліотека полегшених, автономних компонентів з математики та статистики, які вирішують найбільш поширені проблеми, недоступні в мові програмування Java або Commons Lang.

В третьому розділі було запропоновано модель роботи робота-парсера для отримання і обробки даних з вебз-астосунків вищих навчальних закладів.

Для реалізації своїх технічних задач, веб-парсер має виконувати наступні дії:

1. Зайти на сайт Webometrics, в розділ ВНЗ України.
2. Отримати зі сторінки таблицю, зображену нижче на рис. 2.1.
   a. Отримати з кожного рядку дані (назва ВНЗ, посилання, оцінки, місце в рейтингу).
   b. Зберегти їх у массив даних застосунку.
3. Отримати посилання на наступну сторінку.
4. Повторити етапи 2-3, доки залишаються сторінки і таблиці.
5. Отримати посилання з масиву даних, і перейти на веб-застосунки ВНЗ.
6. Отримати дані від веб-застосунку і зберегти їх у окремий масив.
7. Зібрати данні, і провести їх обробку.
   a. Кластеризація даних за їх класовою назвою зі зберіганням рахунку
   b. Сортування за кількістю повторень
8. Презентувати вихідні дані.

Для виконання роботи, я вирішив створити парсер на мові програмування Java, так як вона добре підтримується, і має чималу кількість бібліотек.

Моя програма має складатися з декількох частин:

- Стартовий екран – на ньому користувач буде вказувати свій сайт (цільовий сайт), який буде порівнюватися з іншими, а також обирає континент і країну для пошуку інформації. З цього екрану буде ініційований пошук.
- Клас для парсингу таблиці Webometrics і аналізу отриманих даних – даний клас ініціює парсинг посилань на існуючі веб-застосунки ВНЗ і

парсить власне веб-застосунки і аналізує отримані дані, які згодом і представляє.

- Клас для зберігання даних веб-застосунків за посиланням – тут зберігається структура веб-застосунків, і інформація по класам застосованих на сторінках веб-застосунків навчальних закладів.

Для виконання роботи були обрані класи бібліотеки Jsoup. Jsoup оптимально підходить для парсингу великої кількості веб-застосунків різної якості, і орієнтована на гнучкість і простоту використання. Він був використаний для вилучення окремих даних з HTML-сторінок. Щоправда Jsoup не підтримує JavaScript, тому різноманітний пост-контент є недоступним для використання, але він нам не потрібен, так как ми досліджуємо власне структуру веб-застосунків.

Для виконання роботи були обрані класи бібліотеки Jsoup. Jsoup оптимально підходить для парсингу великої кількості веб-застосунків різної якості, і орієнтована на гнучкість і простоту використання. Він був використаний для вилучення окремих даних з HTML-сторінок. Щоправда Jsoup не підтримує JavaScript, тому різноманітний пост-контент є недоступним для використання, але він нам не потрібен, так как ми досліджуємо власне структуру веб-застосунків.

Спочатку було бажано обрати вже готову бібліотеку кластеризації, і серед існуючих кандидатур було обрано бібліотеку ELKI, за її довге існування, підтримку і наявність навчальних матеріалів. Згодом же, на практичному створенні програмного продукту було вирішено створити власний спрощений алгоритм, для реалізації базових задач.

Для того щоб розпочати роботу з JSoup необхідно під’єднати бібліотеки до проекту. Це можна зробити використав Maven, або підключивши jar-файл напряму.

Стартовий екран проекту призначений для того щоб зібрати усі необхідні стартові параметри. Йому потрібні:

- Посилання на цільовий веб-застосунок– в нашому випадку за замовченням використовується веб-застосунок ЧНУ ім. П. Могили.
- Континент – так як веб-застосунки ВНЗ структуруються за континентом і країною. За замовченням вказана Європа.

- Країна – для пошуку в межах однієї країни. За замовченням вказана Україна.

При зчитуванні даних вони заносяться у дані програми, і згодом обробляються алгоритмами кластеризації і обробки інформації.

Коли зчитується строка рейтингу, вона поелементно зберігає інформацію про ранг ВНЗ у рейтингзі, його назву, і оцінки по критеріям.

Коли ці данні заносяться в елемент класу для зберігання, в конструкторі цього класу викликається метод для паралельного парсингу власне веб-застосунку цього ВНЗ. При роботі цього методу досліджується головна сторінка веб-застосунку, з усіма його елементами. Парсер досліджує елементи цієї сторінки, які класи використовуються і як.

Під час парсингу сторінки зберігається інформація про зміст веб-сторінки і підраховуються основні елементи і їх кількість. При цьому якщо певні елементи повторюються на сторінці вони не будуть багаторазово зберігатися в даних, натомість їх кількісна оцінка буде збільшена.

Після збереження даних по сайтам, буде ініціалізована кластеризація даних, під час якого отримані дані по елементам сайту будуть розбиті на кластери за їх призначенням. Результати цього процесу будуть представлені користувачу, де вони будуть взважені по їх ненаявності на цільовому веб-застосунку, і кількістю повторень на інших сайтах

У четвертому розділі наводиться опис розробленого програмно-алгоритмічного забезпечення і перспективи подальшої роботи над ним.

У спеціальній частині магістерської наукової роботи з «Охорони праці та безпеки життєдіяльності» розглянуто мікрокліматичні умови праці на робочих місцях на предмет виробничого освітлення та дотримання вимог експлуатації ПК в офісі компанії «SmartCAR». В результаті розрахунків встановлено, що передбаченої кількості вікон, тобто їх загальної площі вистачає для забезпечення вимог санітарних норм щодо природного освітлення приміщення для якого проводився аналіз умов праці. Визначено, що для забезпечення штучного освітлення слід використовувати 10 світильників, які необхідно комплектувати 2 люмінесцентними

лампами типу ЛДЦ потужністю 30 Вт та довжиною 0,9 м кожна. Загальна потужність освітлення складає 600 Вт. Розроблено інструктаж з техніки безпеки під час типових надзвичайних ситуацій.

У методичній частині розроблено лекцію і практичну роботу на темі «Парсинг веб-сторінок».

ЗАГАЛЬНІ ВИСНОВКИ

У даній магістерській роботі були досліджені веб-застосунки вищих навчальних закладів, їх структура і особливості.

Були досліджені сервіси-парсери, їх функції і особливості роботи. Під час роботи було відмічено, що частіше всього парсери використовуються для окремих елементів чи блоків для отримання актуальної інформації стосовно певних продуктів, чи для побудови маркетингової кампанії.

Під час роботи був досліджений рейтинг Webometrics, організації що оцінює веб-застосунки за наступними критеріями: Вплив, Присутність, Відкритість, і Цінність. Ці дані рейтингу було вирішено використати як стартову точку для дослідження веб-застосунків, так як рейтинг містить у собі оцінки веб-застосунків, їх повну назву, і посилання на власне веб-застосунки.

На основі цих даних було вирішено створити робот-парсер, який зможе автоматично отримувати зміст. Для створення парсера було досліджені варіанти використання програмних бібліотек для покращення якості отриманних даних. Серед доступних альтернатив було обрано програмну бібліотеку Jsoup. За допомогою проведених досліджень було створено прототип парсеру, який отримує дані з веб-застосунків. Отримані дані парсеру були опрацьовані, і виведені на окреме вікно у вигляді таблиці.

Розроблений продукт є прототипом і не є фінальним. Серед можливих доповнень у майбутньому можлива реалізація більш дружелюбного інтерфейсу і більш просунутого алгоритму обробки даних.

У методичній частині магістерської роботи розроблено лекцію і практичну роботу на темі «Парсинг веб-сторінок».

У спеціальній частині магістерської роботи з «Охорони праці та безпеки в надзвичайних ситуаціях» здійснено аналіз умов праці у Асоціації Університетів України, що є громадською організацією, що поєднує в собі академічну спільноту на демократичних засадах. Під час роботи було досліджено рівень природного освітлення, і підрахований необхідний рівень штучного освітлення в приміщені за допомогою люмінісцентних ламп. Крім того було розроблено інструктаж для дій працівників компанії в разі типових надзвичайних ситуацій, а саме: при пожежі, землетрусі, урагані, грозі.

АНОТАЦІЯ

до магістерської наукової роботи

«Дослідження характеристик веб-застосунків ВНЗ з використанням робота-парсера»

Студент :Кащенко Дмитро Олегович

Керівник: Мещанінов Олександр Павлович

В магістерській роботі був проведений аналіз веб-застосунків вищих навчальних закладів України, дослідження їх змісту і інструментів якими вони були створені, а також порівняння їх з «цільовим» сайтом, який обирається користувачем. Для полегшення роботи і автоматизації процессу був створений робот-парсер, який збирає інформацію з відкритих джерел, досліджує веб-застосунки, опрацьовує дані і представляє їх для побудови рекомендацій.

Метою роботи є аналіз веб-застосунків ВНЗ, дослідження їх відмінностей, і побудова рекомендацій для впровадження на цільовий веб-застосунок.

Предметом дослідження є розробка моделі робота-парсера для автоматичного збору інформації і аналізу веб-застосунків, та дослідження їх характеристик.

Об’єктом дослідження є процес аналізу веб-застосунків ВНЗ і інструменти для їх аналізу.

У першому розділі магістерської роботи розглянуті сучасні тенденції створення веб-застосунків. У другому розділі були досліджені технології для створення робота-парсеру. У третьому розділі були описано створення робота-парсера. У четвертому розділі було протестовано програмний продукт, і досліджені можливості його подальшого удосконалення. Крім того також наявна спеціальна частина «Охорона праці в громадській організації». А також методична частина, в якій описана лекція, і практична робота на тему парсингу веб-застосунків.

Сторінок – 83, Таблиць – 3 , Посилань – 35, Додатків – 2.

Ключові слова: веб-застосунок, аналіз, кластерізація, парсер, робот, автоматизація, алгоритм, програмний продукт, програмна бібліотека.

ABSTRACT

to Master's work

«Analyzing web-site features of universities using parser-robot»

Student : Kashchenko Dmyto Olehovych

Head: Meshchanynov Oleksandr Pavlovych

In this scientific work, websites of Ukraine's universities were analyzed among with their contents, and structural instruments they build with, as well as comparing them with target-web-site, selected by user. To ease and automate the process, robot-parser is made, and it gathers the data from web-sites, processes it and presents the data to make a recommendations on improving the target-web-site.

The purpose of the work is the analysis of web-sites of different universities, analyzing their differences and features, and making recommendations on improving target-web-site.

Subject of this scientific work is the developing the prototype of robot-parser for automatic data-gathering and analysis of web-sites and their features.

Object of the analysis made in this scientific work is the process of analysis of web-sites of universities, and instruments of analysis.

First section of the master's work shows modern tendencies in creating and developing web-sites. Second section describes the technologies that are needed to use parser-robot. Third section describes the algorithm used to make robot-parser, and methods that were generally used. Forth section describes the testing of product made, and analyzes what can be enhanced in future. Besides that, there is special part on labor safety in public organization. Also there is methodic part, which consists on lection and lab work about parsing websites.

Pages – 83, Tables – 3, References – 35, Applications – 2.

Keywords: web-site, analysis, clusterization, parser, robot, automation, algorithm, program product, program library.