UNIVERSITY OF SCIENCE AND TECHNOLOGY

COLLEGE OF GRADUATE STUDIES AND ACADEMIC ADVANCEMENT

Evaluate Web Content Mining Tools for Effective Extraction of Informative Data from the Web Pages

A Thesis

Submitted to the College of Graduate Studies and Academic Advancements in Partial Fulfillment of the Requirement for the Degree of Master of Science in Information Systems

Prepared By:

Huzaifa Abdalmonem Yassen

Supervisor By:

Dr. Hassan Fadlallah

Jan 2017

Abstract

Nowadays, the Web has become one of the most widespread platforms for information change and retrieval. As it becomes easier to publish documents, the number of users, publishers and documents increased. This make searching for information and extracting informative and related data from the web pages too difficult. Web Content Mining is one of the elements of Web Mining. One of the objective of web content mining is to extract useful information from the content of the web document. There are many tools that can be used to extract the required data out of the web pages. However these tools are different from each other with respect to many features and characteristics. As example there are some tools designed to work with specific platform while others work to extract specific data structure out of web page. Studying and analyzing these tools still needed in order to enhance their work for extracting the required data from web document. The thesis evaluate two tools; Web Content Extractor and Automation Anywhere to detect which tools is better in extract the data out of web page. The evaluation base on four parameters they are (data type argument, recording of the data argument, browser type argument, and usability measurement argument).The result of evaluation proved that Automation Anywhere is better than Web Content Extractor in the field of extraction useful data and information from the web pages.

المستخلص

في الوقت الحاضر،أصبح الويب واحدا من المنصات الاكثر انتشارا لتغير واسترجاع المعلومات،كما اصبح من السهل نشر الوثائق فيه،حيث أن عدد المستخدمين و الناشرين في ازدياد وبالتالي فأن عدد الوثائق في نمو متزايد،وهذا يجعل البحث عن المعلومات وأستخلاص البيانات المفيدة وذات الصلة من صفحات الويب أمرا صعبا جدا. التنقب في محتويات الويب أحد عناصر التنقيب في الويب.وأحد أهداف التنقيب في محتويات الويب هو أستخلاص المعلومات المفيدة من محتوي ملفات الويب.وهنالك العديد من الأدوات المستخدمة في أستخلاص البيانات المطلوبة من صفحات الويب.ومع ذالك فأن هذه الأدوات تختلف في بعضها من ناحية العناصر والخصائص.وكمثال هنالك بعض الأدوات التي صممت لتعمل في منصات محددة في حين أن بعضها الأخر تغمل لأستخلاص بيانات معينة من صفحات الويب.ومازالت الحاجة لدراسة Web و Automation Anywhereوتحليل هذه الأدوات مطلوبة في سبيل تحسين العمل في هذه الأدوات لأستخلاص البيانات المطلوبة من ملفات الويب.هذا البحث يقيم وهما أثنين من الأدوات التي تعمل في Content Extractor مجال أستخلاص المعلومات المفيدة من صفحات الويب لأكتشاف ايهما الأفضل في هذا المجال. التقيم كان مبنيا علي Automationأفضل منWebأربعة معاملات أساسية هي(معامل نوع البيانات،معامل تسجيل البيانات،معامل نوع المتصفح ومعامل قياس سهولة الاستخدام) ونتيجة التقيم أثبتت أنAnywhere Content في مجال أستخلاص البيانات والمعلومات المفيدة من صفحات الويب.

1.1 Introduction

Web content mining tools and its applications are very important today. Because World Wide Web has a popular place for dissipating and accumulates the information. Extracting the useful information from web pages becomes essential task. Web is a medium for accessing the information store in different sources.

Web Mining is a developing research area motivated on resolving these problems. The various techniques include Web Content Mining, Web Usage Mining, and Web Structure Mining. Web Pages contain large amount of undesired information, which is called noisy or irrelevant content. The navigational panel, header, Footer, copyright, advertisements known as noisy content [1].

Extracting the information from various resources has many problems like finding the useful information from the web page and ignoring the irrelevant or not useful data and stay away from noisy information.

However to ignore noisy or irrelevant information and extract useful data and information we using web content mining tools for better extraction informative data and information from the web sides [2].

1.2 Problem Statement

They’re a many tools used in extraction the data from the web pages. But these tools have different accuracy and effectiveness in extraction the data.

The process of chosen the appropriate tool to extract the data from the web pages considered as problem. In this research we will make evaluation between two tools (Automation Anywhere and Web Content Extractor) from the tools that use in this field to determine which of two tools is best in extraction the data from the web pages.

1.3 Research Objective

This research aims to accomplish the following objectives:

1. To investigate and explore how these tools work.
2. To validate the applicability of the selected tools.
3. To propose a new method to evaluate the tools.

1.4 Research Methodology

1. Selecting and describing the tool.
2. Selecting evaluation parameter.
3. Executing the measurement process.
4. Discuss the results.

1.5 Thesis Structure

The thesis contains five chapters the second chapter contain literature review, which discusses the web mining background and related work. And the Third chapter describes web content mining tools and the research methodology. The Fourth includes the implementation and results discussion. While the fifth chapter includes conclusions and recommendation