CHAPTER 7

RELATIONAL WEB WRAPPER

(RWW)

The information available over the Internet is growing at rapid rate, so Web data
extraction systems are required to extract the required information. One such technique
to extract the data is Web wrapper, which is a supervised learning approach in it a
template (program) is developed by the programmer to extract some specific data. This
research work provides a Web wrapper known as Relational Web Wrapper (RWW)
which extracts related information of the Webpage. RWW acts as an example to
describe the working of Web Wrapper induction system i.e. how to develop the
program for Web Wrapper and how to test the performance of a Web Wrapper. The
performance evaluation of RWW is shown in the results.

This chapter is organized into four sections, Section 7.1 consists of the introduction,
Section 7.2 provides the System model (RWW), Section 7.3 shows the experimental
results and Section 7.4 provides summary of this research work.

7.1 INTRODUCTION

WWW has becomes the most extensively used information retrieval system among all
the other information sources available and hence growing at a very fast rate. Because of
this rapid growth and large information available Web data extraction systems are
necessary to use. Web wrapper is one of such technique which is a supervised learning
approach where a programmer develops a program (template) and applies it to a number
of WebPages. Web wrapper may solve a particular problem i.e. may extract a particular
pattern of data. Web wrapper translates the data included inside the WebPages into a
relational structure and store into some specific format like xml, excel sheets etc.
Wrappers may be developed manually, through wrapper induction or automatic
approach. First method of wrapper generation is very difficult, error-prone and specific.
Hence, semi-automatic or automatic forms are more preferable (89, 107-109). In semi-
automatic or wrapper induction, system a program known as web wrapper is developed based on the template structure of the Website. In case of automatic approach a program is developed which scans the Websites and develop template structure by itself according to the website. Hence it is free from a particular template structure of a website.

This research work provides Relational Web Wrapper which uses wrapper induction approach i.e. a program is developed by the developer. This wrapper extracts the related data of a Webpage which may further be used for extending the information of the Webpage or user.

### 7.2 SYSTEM MODEL (RWW)

Relational Web Wrapper takes a Webpage as input and extracts the related data of Webpage like Title, Summary and Keywords. The extracted data like keyword (related data of the webpage) are used for further searching so that more results may be provided. The extracted data is stored in excel sheets. It provides the user an easy approach to search the related data of the Webpage.

The working of RWW is shown in Figure 7.1

![Figure 7.1: Relational Web Wrapper Working](image)
Algorithm 7.1: Relational Web Wrapper (RWW Algorithm)

It describes how to develop a Web wrapper and hence the algorithm is developed in coding format to show the development of Web Wrapper.

Phase 1: Extracting the data from the Webpage and calculation of time to extract the data

Step 1: var start = new Date().getTime()

Declare variables t as title of document, l as location of document and E for keywords from meta tag as follows:

t = (document.title)
l = (document.location.href)
E = document.getElementsByTagName('meta')

Step 2: var desc

for (i = 0, j = E.length; i < j; i++) {
if (E[i].name.toLowerCase() == "desc") {
desc = E[i]
endfor

Step 3: print ('Title of the webpage is ' + '<b>' + t + '</b>')

Step 4: print ('Location of the document is ' + ' &lt;a href="#" + l + ""> +
desc.content )

Step 5: for (i = 0, j = E.length; i < j; i++) {
if (E[i].name.toLowerCase() == "keywords") {
desc = E[i]
endfor

Step 6: print ('keywords extracted from the webpage are' + ' &lt;a href="#" + l + ""> + t +
'&lt;/a&gt;' + desc.content)

Step 7: var end = new Date().getTime()

Step 8: var extraction_time = end - start

Step 9: print ('Time taken to extract the data is ' + extraction_time)

Phase 2: Storing Data to Excel Sheets

Step 10: Declare uri = 'data: application/vnd.ms-excel; base64,'
Declare n as namespace for excel workbook as
xmlns: n="urn: schemas-microsoft-com: office:excel"

**Step 11:** Use n to set options of excel workbook

```xml
<n:ExcelWorkbook>
  <n:ExcelWorksheets>
    <n:ExcelWorksheet>
      <n: Name> {worksheet} </n:Name>
      <n:WorksheetOptions><n:DisplayGridlines/> </n:WorksheetOptions>
    </n:ExcelWorksheet>
  </n:ExcelWorksheets>
</n:ExcelWorkbook>
```

**Step 12:** base64 = function(p) {return window.btoa (unescape
(encodeURIComponent(p)))

**Step 13:** format = function (p, c) {return p.replace (/ {(\w+)}/g, function (c, p) {return
  c[p];}) }

**Step 14:** return function (t, name) {if (!t.nodeType) t = document.getElementById(t)
  var w = {worksheet: name || 'Worksheet', t: t.innerHTML}
  window.location.href = uri + base64 (format (template, w))}

### 7.3 EXPERIMENTAL RESULTS AND DISCUSSION

RWW is developed using JavaScript platform. Three Websites named Website1 (MMIM), Website2 (MMEC) and Website3 (MMCE) were developed in HTML and JavaScript for experimentation.

Figure 7.2 shows the snapshots of the experiment performed on Website1 (MMIM). It shows one of the Webpage of Website1 along with the extracted data, starting time and end time on top of it.
Figure 7.2: Webpage with Extracted Data

Figure 7.3 shows the running of RWW which provides result in the form of extracted data and total evaluation time.
Figure 7.3: Extracted Data and Evaluation Time

Figure 7.4 shows the snapshot of storing the extracted data into the excel sheet.

Figure 7.4: Storing the Extracted Data to Excel Sheet
Performance Evaluation of Relational Web Wrapper

Table 7.1 shows the performance of running the RWW for the Website1 in terms of extraction time. The program developed for RWW was applied over each Webpage of Website1 for 20 times. The Results found for 2, 5, 10 and 20 times is shown in Table 7.1.

<table>
<thead>
<tr>
<th>Name of Webpage</th>
<th>Average Time to Extract the Data (Number of runs)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Index.html</td>
<td>5.0</td>
</tr>
<tr>
<td>Life.html</td>
<td>6.0</td>
</tr>
<tr>
<td>Prog.html</td>
<td>6.0</td>
</tr>
<tr>
<td>Student.html</td>
<td>5.0</td>
</tr>
<tr>
<td>Approach.html</td>
<td>5.0</td>
</tr>
<tr>
<td>Message.html</td>
<td>5.0</td>
</tr>
<tr>
<td>Facilities.html</td>
<td>6.0</td>
</tr>
<tr>
<td>Knowledge.html</td>
<td>6.0</td>
</tr>
</tbody>
</table>
Table 7.2 shows the performance of running the RWW for the Website2 in terms of extraction time. The program developed for RWW was applied over each Webpage of Website2 for 20 times. The results found for 2, 5, 10 and 20 times is shown in Table 7.2.

**Table 7.2:** Time to Extract the Related Data from Website2

<table>
<thead>
<tr>
<th>Name of Webpage</th>
<th>Average Time to Extract the Data (Number of runs)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Index.html</td>
<td>6.0</td>
</tr>
<tr>
<td>Life.html</td>
<td>8.0</td>
</tr>
<tr>
<td>Prog.html</td>
<td>7.0</td>
</tr>
<tr>
<td>Student.html</td>
<td>6.0</td>
</tr>
<tr>
<td>Approach.html</td>
<td>6.0</td>
</tr>
<tr>
<td>Message.html</td>
<td>6.0</td>
</tr>
<tr>
<td>Facilities.html</td>
<td>6.0</td>
</tr>
<tr>
<td>Knowledge.html</td>
<td>7.0</td>
</tr>
</tbody>
</table>

Table 7.3 shows the performance of running the RWW for the Website3 in terms of extraction time. The program developed for RWW was applied over each Webpage of Website3 for 20 times. The Results found for 2, 5, 10 and 20 times is shown in Table 7.3.
### Table 7.3: Time to Extract the Related Data from Website3

<table>
<thead>
<tr>
<th>Name of Webpage</th>
<th>Average Time to Extract the Data (Number of runs)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Index.html</td>
<td>7.0</td>
</tr>
<tr>
<td>Life.html</td>
<td>6.0</td>
</tr>
<tr>
<td>Prog.html</td>
<td>7.0</td>
</tr>
<tr>
<td>Student.html</td>
<td>7.0</td>
</tr>
<tr>
<td>Approach.html</td>
<td>7.0</td>
</tr>
<tr>
<td>Message.html</td>
<td>7.0</td>
</tr>
<tr>
<td>Facilities.html</td>
<td>7.0</td>
</tr>
<tr>
<td>Knowledge.html</td>
<td>7.0</td>
</tr>
</tbody>
</table>

Table 7.1, Table 7.2 and Table 7.3 revealed that RWW program takes more time when it is run for the first time than it is run for the subsequent number of time.

**Accuracy Measurement of RWW:** Accuracy of RWW is measured on the basis of following formula

\[
\text{Accuracy} = \frac{\text{Data extracted by the RWW}}{\text{Data extracted on manual basis}} \quad \ldots(7.1)
\]
Table 7.4 shows the comparison of accuracy of RWW and Manual in terms of data extraction.

**Table 7.4: Accuracy Measurement (RWW vs. Manual) for Website1**

<table>
<thead>
<tr>
<th>Name of Webpage</th>
<th>%age of data extracted by Relational Web Wrapper</th>
<th>%age of data extracted on Manual basis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Title</td>
<td>Description</td>
</tr>
<tr>
<td>Index.html</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Life.html</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Prog.html</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Student.html</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Approach.html</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Message.html</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Facilities.html</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Knowledge.html</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

The relational web wrapper provides 100% accuracy in extraction of title and description of document and more than 95% accuracy in extracting the keywords of most of the WebPages.
7.4 SUMMARY

RWW provides an example of wrapper induction system which may extract related data or meta-data of the Webpage efficiently and with minimal user effort. The work may further be extended to more dynamic content and saving the results to xml format. Accuracy of RWW may be extended by applying some improved technique on the same wrapper.