

# Contents

<b>Table of Contents</b>	<b>19</b>
<b>List of Figures</b>	<b>25</b>
<b>List of Tables</b>	<b>28</b>
<b>I Introduction</b>	<b>33</b>
<b>1 Preamble</b>	<b>35</b>
1.1 Motivation . . . . .	35
1.2 Contributions of the thesis . . . . .	37
1.3 Structure of the thesis . . . . .	39
<b>2 Block Detection Techniques</b>	<b>43</b>
2.1 Data Mining . . . . .	43
2.2 Web Mining . . . . .	44
2.3 Web Content Classification . . . . .	45
2.4 Wrappers and Unsupervised Learning . . . . .	47
2.5 Block Detection . . . . .	48
2.6 Conclusions . . . . .	52
<b>II Foundations</b>	<b>53</b>
<b>3 The DOM tree</b>	<b>55</b>
3.1 Brief history of DOM . . . . .	56
3.2 Main characteristics of DOM . . . . .	57
3.3 From the document to the browser's screen . . . . .	58
3.4 Conclusions . . . . .	63

---

<b>4</b>	<b>Preliminary Definitions and Notation</b>	<b>65</b>
4.1	Basic definitions . . . . .	65
4.2	Site-level techniques . . . . .	68
4.2.1	Candidates selection . . . . .	69
4.2.2	Mapping . . . . .	72
4.3	Web page blocks . . . . .	72
4.3.1	Web page menu . . . . .	72
4.3.2	Template . . . . .	73
4.3.3	Main content . . . . .	74
4.3.4	Relationship between web page menu and template .	74
4.3.5	Relationship between the template and the main content . . . . .	74
4.4	Evaluation metrics . . . . .	76
<b>III</b>	<b>Page-level Block Detection Algorithms</b>	<b>79</b>
<b>5</b>	<b>Page-level Menu Detection</b>	<b>81</b>
5.1	Related Work . . . . .	82
5.2	Menu detection algorithm . . . . .	83
5.2.1	Rating DOM nodes . . . . .	83
5.2.2	Selection of candidates . . . . .	88
5.2.3	Selection of root nodes . . . . .	88
5.2.4	Selection of the menu node . . . . .	91
5.3	Implementation . . . . .	92
5.3.1	Empirical evaluation . . . . .	92
5.3.2	Evaluation phase: Precision/recall measurement . .	96
5.4	Conclusions . . . . .	100
5.5	Contributions . . . . .	102
<b>6</b>	<b>Page-level Content Extraction</b>	<b>103</b>
6.1	Related Work . . . . .	104
6.2	Main content extraction . . . . .	105
6.2.1	The web page's main content . . . . .	106
6.2.2	Weighting DOM nodes . . . . .	107
6.2.3	Properties standardization . . . . .	111
6.2.4	$c$ - <i>SET</i> computation . . . . .	113
6.2.5	Selecting the main content nodes . . . . .	114
6.2.6	Final post-process . . . . .	115
6.3	Implementation . . . . .	117

---

6.3.1	Empirical evaluation . . . . .	118
6.4	Conclusions . . . . .	125
6.5	Contributions . . . . .	126
<b>IV</b>	<b>Site-level Block Detection Algorithms</b>	<b>127</b>
<b>7</b>	<b>Candidates selection algorithms</b>	<b>129</b>
7.1	Related Work . . . . .	130
7.2	Identifying web pages that implement the same template . .	131
7.2.1	Complete subdigraphs . . . . .	132
7.2.2	Hyperlink analysis . . . . .	134
7.2.3	Finding web page candidates in a website . . . . .	139
7.3	Implementation . . . . .	143
7.3.1	Empirical evaluation . . . . .	144
7.4	Conclusions . . . . .	151
7.5	Contributions . . . . .	152
<b>8</b>	<b>Equal Top-Down Mapping</b>	<b>153</b>
8.1	Related Work . . . . .	153
8.2	Comparing DOM nodes . . . . .	154
8.2.1	Template extraction from a complete subdigraph . .	154
8.3	Implementation . . . . .	159
8.3.1	Empirical evaluation . . . . .	160
8.4	Conclusions . . . . .	165
8.5	Contributions . . . . .	166
<b>9</b>	<b>Site-level Template Detection</b>	<b>167</b>
9.1	Related work . . . . .	167
9.2	Template detection . . . . .	170
9.2.1	The web page's template . . . . .	170
9.2.2	Building a complete subdigraph . . . . .	171
9.2.3	Web pages implementing several templates . . . . .	173
9.2.4	Template detection from a complete subdigraph . .	174
9.3	Implementation . . . . .	175
9.3.1	Empirical evaluation . . . . .	175
9.4	Conclusions . . . . .	184
9.5	Contributions . . . . .	184

---

<b>10 Site-level Content Extraction</b>	<b>185</b>
10.1 Related work . . . . .	185
10.2 Main content extraction . . . . .	188
10.2.1 The web page's main content . . . . .	189
10.2.2 Set of web pages selection . . . . .	192
10.2.3 Web pages mapping . . . . .	193
10.2.4 Candidate set reduction . . . . .	194
10.2.5 Main content branch detection . . . . .	196
10.2.6 Discarding candidates . . . . .	198
10.2.7 Main content selection . . . . .	199
10.3 Implementation . . . . .	201
10.3.1 Empirical evaluation . . . . .	202
10.4 Conclusions . . . . .	216
10.5 Contributions . . . . .	217
<b>11 Hybrid Technique for Template Detection</b>	<b>219</b>
11.1 Related work . . . . .	220
11.2 Hybrid template detection . . . . .	221
11.2.1 HTML to DOM corresponding to page-level ConEx	222
11.2.2 Content extraction . . . . .	223
11.2.3 Hyperlink analysis . . . . .	224
11.2.4 Complete subdigraph extraction . . . . .	225
11.2.5 HTML to DOM corresponding to TemEx . . . . .	225
11.2.6 Template detection . . . . .	225
11.3 Implementation . . . . .	226
11.3.1 Empirical evaluation . . . . .	227
11.4 Conclusions . . . . .	230
11.5 Contributions . . . . .	231
<b>V Comparison with the State of the Art</b>	<b>233</b>
<b>12 Comparison with the State of the Art</b>	<b>235</b>
12.1 Selection and description of Web template detectors . . . . .	237
12.1.1 Methodology for the selection of template detectors	237
12.1.2 Search results . . . . .	239
12.2 A workbench for template detection . . . . .	244
12.3 Comparison of template detectors . . . . .	248
12.3.1 Computation time . . . . .	250
12.3.2 Scalability . . . . .	250

---

12.3.3 Asymptotic costs . . . . .	251
12.4 Comparison of content extractors . . . . .	252
12.5 Conclusions . . . . .	256
12.6 Contributions . . . . .	257
<b>VI Implementations</b>	<b>259</b>
<b>13 TeCo Benchmark Suite</b>	<b>261</b>
13.1 Benchmark Suite Structure . . . . .	262
13.2 Producing the Gold Standard . . . . .	264
13.3 Benchmark details . . . . .	265
13.4 Guidelines for using the suite . . . . .	269
13.4.1 Downloading and configuring the suite . . . . .	269
13.4.2 Rules for using the suite and report . . . . .	273
13.5 Conclusions . . . . .	278
13.6 Contributions . . . . .	279
<b>14 Implementation</b>	<b>281</b>
14.1 Implementation of the WebExtensions . . . . .	281
14.1.1 Architecture . . . . .	281
14.1.2 Structure . . . . .	285
14.1.3 Evaluation environment . . . . .	287
14.2 Usage scenario . . . . .	288
14.3 Tools information . . . . .	290
14.3.1 Differences between different browsers . . . . .	291
14.4 Conclusions . . . . .	292
14.5 Contributions . . . . .	293
<b>VII Conclusions and Future Work</b>	<b>295</b>
<b>15 Conclusions</b>	<b>297</b>
<b>16 Open Lines of Research</b>	<b>303</b>
<b>Bibliography</b>	<b>306</b>
<b>A Glossary of Acronyms</b>	<b>327</b>

<b>B Scientific Contributions</b>	<b>331</b>
B.1 Conference papers . . . . .	331
B.2 Journal Publications . . . . .	332
B.3 List of derived artifacts . . . . .	334