DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN
UNIVERSITAT POLITÈCNICA DE VALÈNCIA

P.O. Box: 22012 E-46071 Valencia (SPAIN)

Informe Técnico / Technical Report

Ref. No.:	2019/01 **Pages:** 10
Title:	A Benchmark Suite for Template Detection and Content Extraction
Author(s):	J. Alarte, D. Insa, J. Silva and S. Tamarit
Date:	January, 2019
Keywords:	Template Detection, Content Extraction, Benchmark Suite, Information Retrieval

Vº Bº
Leader of research Group Author(s)

A Benchmark Suite for Template Detection and Content Extraction*

Julián Alarte[1], David Insa[1], Josep Silva[1], and Salvador Tamarit[2]

[1] Universitat Politècnica de València, Camino de Vera S/N, E-46022 Valencia, Spain.
{jalarte,dinsa,jsilva}@dsic.upv.es
[2] Babel Research Group, Universidad Politécnica de Madrid, Madrid, Spain
stamarit@babel.ls.fi.upm.es

Abstract. Template detection and content extraction are two of the main areas of information retrieval applied to the Web. They perform different analyses over the structure and content of webpages to extract some part of the document. However, their objective is different. While template detection identifies the template of a webpage (usually comparing with other webpages of the same website), content extraction identifies the main content of the webpage discarding the other part. Therefore, they are somehow complementary, because the main content is not part of the template. It has been measured that templates represent between 40% and 50% of data on the Web. Therefore, identifying templates is essential for indexing tasks because templates usually contain irrelevant information such as advertisements, menus and banners. Processing and storing this information is likely to lead to a waste of resources (storage space, bandwidth, etc.). Similarly, identifying the main content is essential for many information retrieval tasks. In this paper, we present a benchmark suite to test different approaches for template detection and content extraction. The suite is public, and it contains real heterogeneous webpages that have been labelled so that different techniques can be suitable (and automatically) compared.

1 Introduction

Template extraction is an important tool for website developers, and also for website analyzers such as crawlers. Content extraction is essential for many information processing tasks applied to webpages. In the last decade, there have been important advances that produced several techniques for both disciplines.

* This work has been partially supported by the EU (FEDER) and the Spanish *Ministerio de Economía y Competitividad (Secretaría de Estado de Investigación, Desarrollo e Innovación)* under grant TIN2013-44742-C4-1-R and by the *Generalitat Valenciana* under grant PROMETEO/2011/052. David Insa was partially supported by the Spanish Ministerio de Educación under FPU grant AP2010-4415. Salvador Tamarit was partially supported by research project POLCA, Programming Large Scale Heterogeneous Infrastructures (610686), funded by the European Union, STREP FP7.

Hybrid methods that exploit the strong points of several techniques have been defined too. In order to test, compare and tune these techniques, researchers need:

- collections of benchmarks that are heterogeneous (to ensure generality of the techniques) and
- a gold standard (to ensure the same evaluation criteria).

A benchmark suite is essential to measure the performance of these techniques, and to compare them with previous approaches. Benchmark suites are used in the testing phase and in the evaluation phase. The testing phase allows developers to optimize the techniques by adjusting parameters. Once the technique has been tuned, the evaluation phase allows us to know its performance with objective measures. It is obvious that the set of benchmarks used in the testing phase cannot be used in the evaluation phase, thus, they need disjoint sets of webpages.

In this paper we present a benchmark suite together with a gold standard that can be used for template detection, content extraction and menu detection. All benchmarks have been labelled so that every HTML element of the webpages indicates whether it should be classified as main content or not, and whether it should be classified as template or not. In addition, the main menu of the webpages has also been labelled. The suite also incorporates scripts to automatize the benchmarking process.

This suite has been developed as the result of a research project. We developed a new technique for content extraction [4] that was later adapted for template detection [2]. In the evaluation phase, our initial intention was to use a public benchmark suite. We first tried to use the CleanEval [3] suite of content extraction benchmarks, because it has been widely used in the literature. Unfortunately, it is not prepared for template detection. Then, we contacted the authors of other techniques that had already evaluated their techniques. However, we could not use these benchmarks due to privacy (they belong to a company or project whose results were not shared), copyright (they were not publicly available) or unavailability (they had been lost). Finally, we decided to build or own benchmark suite and make it free and publicly available. Later, we developed a new menu detection technique [1], so we decided to update the suite by labelling the nodes that represent the main menu of the website. Hence, we included 25 new benchmarks, for a total of 65. The rest of this paper presents that benchmark suite.

2 The TECO Benchmark Suite

TECO (TEmplate detection and COntent extraction benchmarks suite) was created as a benchmark suite specifically designed for template detection and content extraction. It can be used for testing and evaluation of these techniques, and it is formed from 65 real websites downloaded from Internet. We selected heterogenous websites such as blogs, companies, forums, personal websites, sports

websites, newspapers, etc. Some of the websites are well known, like the BBC website or the FIFA website, and others are less known like personal blogs or small companies websites. The downloading of the webpages was done in some cases using the OS X software SiteSucker, and in other cases using the Linux command `wget`.

It is important to know how the websites were downloaded and stored, so that other researchers can increase the suite if it is needed. The following command downloads a website from the Linux terminal using the `wget` command:
`$ wget --convert-links --no-clobber --random-wait -r 3 -p -E -e robots=off -U mozilla` *`http://www.example.org`*
The meaning of the flags used is:

- `--convert-links`: Converts links so they can work locally.
- `--no-clobber`: Do not overwrite any existing file.
- `--random-wait`: Random waits between downloads.
- `-r 3`: Recursive downloading up to 3 levels of links.
- `-p`: Downloads everything.
- `-e robots=off`: Act as not being a robot.
- `-E`: Get the right file extension.
- `-U mozilla`: Identify as a Mozilla browser.

Each benchmark is composed of:

– A principal webpage, called *key page*. It is the target webpage from which the techniques should extract the main content or the template—note that it is not necessarily the main webpage of the website (e.g., index.html)—.
– A set of webpages that belong to the same website as the key page. This set contains all those webpages that are linked by the key page, and also the webpages linked by them.

2.1 Producing the gold standard

The suite comes with a gold standard that can be used as a reference to compare different techniques. The gold standard specifies for each key page what parts form the template. This is indicated in the own webpage by using HTML classes that indicate what elements are classified as *notTemplate*. It has been produced manually by careful inspection of the websites and mixing the opinion of several people.

In particular, once all the websites were downloaded (the key page and two levels of linked webpages in the same domain), four different engineers did the following independently:

– They manually explored the key page and the webpages accessible from it to decide what part of the webpage is the template and what part is the main content.
– They printed the template and the main content of the webpage.

Then, the four engineers met and performed again these two actions but now all together sharing their individual opinions. Using the results of this agreement, each website was prepared for both, template extraction and content detection. On the one hand, all elements from the key page not belonging to the template were included in a HTML class called *TECO_notTemplate*. This way, a template extraction tool can automatically compare its output with the nodes not belonging to the *TECO_notTemplate* class. On the other hand, all elements belonging to the main content were included in an HTML class called *TECO_mainContent*. Therefore, a content extraction tool can easily compare its output with the nodes belonging to that class. In addition, the node that represents the main menu of the webpage was included in an HTML class called *TECO_mainMenu*. Consequently, a menu extraction tool can compare its output with the nodes belonging to that class.

2.2 Benchmark details

A classification of the benchmarks is important and useful depending on the application and technique that is being fed with them. We provide different classifications according to the purpose and properties of the benchmarks. First, all benchmarks have been classified into five groups:

Companies / Shops, Forums / Social, Personal websites / Blogs, Media / Communication, Institutions / Associations.

Table 1 shows this classification together with the URLs from which we extracted the benchmarks.

Table 2 shows some properties of the benchmarks. Here, column **Nodes** indicates the total number of DOM nodes in the key page, column **T. Nodes** shows the number of DOM nodes that belong to the template and column **M.C. Nodes** refers to the number of DOM nodes that belong to the main content.

2.3 Guidelines for using the suite

2.3.1 Downloading and configuring the suite

TECO is freely distributed and can be downloaded from the URL:

http://www.dsic.upv.es/∼jsilva/retrieval/teco

After downloading the suite, a directory that contains 65 folders, one for each website, is created. Table 3 shows the path to the key page of each benchmark.

2.3.2 Rules for using the suite and report

All researchers and developers that use TECO must follow two basic principles:

1. They must publish their results so that they are publicly available.
2. They must provide enough information so that anyone can easily duplicate their experiments.

Table 1: Sources of the benchmarks

Website type	**Original URL of the webpage**
Forums / Social	es.sharelatex.com/learn/Uploading_a_project
	github.com/DawidStankiewicz/forum
	forum.skyscraperpage.com
	en.citizendium.org
	www.filmaffinity.com/es/
	stackoverflow.com
	www.meneame.net/faq-es/
	www.strangehorizons.com/2004/20040906/greenglass-f.shtml
	www.accountkiller.com/en/delete-activision-account
	study.com/learn/science-questions-and-answers.html
	c.mi.com/it/
	frances.forosactivos.net
	alumni.harvard.edu/help/message-board/
Personal / Blogs	www.cocinaconmarta.com/2015/04/empanadillas-chinas-de-gambas-y-verduras.html
	www.javiercelaya.es
	markahall.blogspot.com.es
	www.trendencias.com
	users.dsic.upv.es/∼dinsa/en/
	googleblog.blogspot.com.es
	www.robyncarr.com/qa/
	www.annmalaspina.com
	users.dsic.upv.es/∼jsilva/wwv2013/index2.html
	foodsense.is/a-list.html
	diarium.usal.es/lguich/pagina-personal-de-luis-arturo-guichard/
	www.folj.com/puzzles/difficult-logic-problems.htm
	oneminutelist.com/16-browser-alternatives-to-desktop-programs/
Companies / Shops	www.eclipse.org
	www.swimmingpool.com
	www.emmaclothes.com
	www.arduino.cc/en/Main/Software/
	today.java.net/pub/a/today/2004/07/06/3ddesktop.html
	clotheshor.se
	ruzafagallery.com/calendario/
	www.raspberrypi.org/resources/teach/
	doodle.com/online-calendar/
	www.newprosoft.com/web-content-extractor.htm
	worryfreelabs.com/about/
	www.intelligencetest.com
	www.ikea.com/gb/
Media / Communication	edition.cnn.com
	www.neoteo.com/star-wars-the-force-awakens-el-regreso-de-viejos-personajes/
	riotimesonline.com
	www.bbc.co.uk/news/
	techcrunch.com/gadgets/
	www.turfparadise.com
	www.cleanclothes.org
	www.afp.com/es/contact/
	news.discovery.com/tech/robotics/artificial-intelligences-hawkings-fears-stir-debate-141206.htm
	www.history.com
	detroit.cbslocal.com/2018/12/04/high-school-newspaper-suspended-after-publishing-disruptive-investigation/
	www.rocklists.com/91x-1983.html
	www.lashorasperdidas.com
Institutions / Associations	web.mit.edu/institute-events/visitor/
	www.isoc-es.org
	www.museodelprado.es
	www.jdi.org.za
	www.u-tokyo.ac.jp/en/about/history.html
	www.savethechildren.net/what-we-do/our-humanitarian-work/
	college.harvard.edu/financial-aid/
	www.unicef.org/where-we-work.html
	www.linuxfoundation.org/about/
	clinicaltrials.gov/ct2/search/index/
	cordis.europa.eu/fp7/ict/fire.html
	www.informatik.uni-trier.de/∼ley/pers/hd/s/Silva_Josep.html
	parents.berkeley.edu/advice/babies/laundry.html

Table 2: Benchmark properties

Id	Benchmark	Nodes	T. Nodes	M.C. Nodes	Menu Links
1	es.sharelatex.com	1091	877	214	118
2	github.com/DawidStankiewicz/	1236	453	783	5
3	forum.skyscraperpage.com	3371	146	3225	6
4	en.citizendium.org	1083	414	633	35
5	www.filmaffinity.com	1333	351	976	32
6	stackoverflow.com	6450	447	5891	5
7	www.meneame.net	760	207	423	11
8	www.strangehorizons.com	634	149	403	23
9	www.accountkiller.com	501	222	279	8
10	study.com	287	103	184	74
11	c.mi.com/it/	3490	2949	541	37
12	frances.forosactivos.net	813	318	495	9
13	alumni.harvard.edu	2004	1785	219	40
14	www.cocinaconmarta.com	5482	3411	1999	10
15	www.javiercelaya.es	754	682	258	12
16	markahall.blogspot.com.es	3137	697	2440	22
17	www.trendencias.com	2426	1139	1042	7
18	users.dsic.upv.es/∼dinsa	241	74	160	5
19	googleblog.blogspot.com.es	5084	3574	1507	2
20	www.robyncarr.com	292	92	200	4
21	www.annmalaspina.com	400	190	206	8
22	users.dsic.upv.es/∼jsilva/	203	169	34	14
23	foodsense.is	334	104	192	5
24	diarium.usal.es/lguich	603	79	524	3
25	www.folj.com	559	175	384	4
26	oneminutelist.com	490	273	217	5
27	www.eclipse.org	256	156	92	8
28	www.swimmingpool.com	607	499	176	25
29	www.emmaclothes.com	1080	374	706	8
30	www.arduino.cc	830	490	340	20
31	today.java.net	696	342	354	6
32	clotheshor.se	459	231	228	8
33	ruzafagallery.com	439	258	176	5
34	www.raspberrypi.org	392	140	252	14
35	doodle.com	572	490	82	5
36	www.newprosoft.com	832	151	681	6
37	worryfreelabs.com	424	321	103	7
38	www.intelligencetest.com	366	284	82	18
39	www.ikea.com	1545	407	1138	10
40	edition.cnn.com	3934	192	3742	15
41	www.neoteo.com	1034	635	399	18
42	riotimesonline.com	2063	1094	747	23
43	www.bbc.co.uk	2991	572	1360	22
44	techcrunch.com	2576	1890	586	35
45	www.turfparadise.com	1057	838	213	98
46	www.cleanclothes.org	1335	266	928	7
47	www.afp.com	1199	404	789	16
48	news.discovery.com	2896	1209	800	68
49	www.history.com	1246	669	260	12
50	detroit.cbslocal.com	1259	1161	98	7
51	www.rocklists.com	765	533	232	6
52	www.lashorasperdidas.com	1822	553	722	12
53	web.mit.edu	393	252	141	9
54	www.isoc-es.org	271	171	100	17
55	www.museodelprado.es	534	148	386	7
56	www.jdi.org.za	626	401	199	10
57	www.u-tokyo.ac.jp	602	499	97	30
58	www.savethechildren.net	751	690	54	21
59	college.harvard.edu	1090	669	397	5
60	www.unicef.org	1052	671	381	4
61	www.linuxfoundation.org	588	534	54	4
62	clinicaltrials.gov	543	423	120	37
63	cordis.europa.eu	959	335	179	19
64	www.informatik.uni-trier.de	3085	64	3021	9
65	parents.berkeley.edu	283	99	184	8

Table 3: Path to the key page of each benchmark

Id	Path to the key page
1	es.sharelatex.com/learn/Uploading_a_project
2	github.com/DawidStankiewicz/forum.1
3	forum.skyscraperpage.com/index.html
4	en.citizendium.org/index.html
5	www.filmaffinity.com/es/main.html
6	stackoverflow.com/index.html
7	www.meneame.net/faq-es.html
8	www.strangehorizons.com/2004/20040906/greenglass-f.shtml.html
9	www.accountkiller.com/en/delete-activision-account.html
10	study.com/learn/science-questions-and-answers.html
11	c.mi.com/it/index.html
12	frances.forosactivos.net/index.html
13	alumni.harvard.edu/help/message-board.html
14	www.cocinaconmarta.com/2015/04/empanadillas-chinas-de-gambas-y-verduras.html
15	www.javiercelaya.es/index.html
16	markahall.blogspot.com.es
17	www.trendencias.com
18	users.dsic.upv.es/∼dinsa/en/index.html
19	googleblog.blogspot.com.es
20	www.robyncarr.com/qa.html
21	www.annmalaspina.com/index.html
22	users.dsic.upv.es/∼jsilva/wwv2013/index2.html
23	foodsense.is/a-list.html
24	diarium.usal.es/lguich/pagina-personal-de-luis-arturo-guichard
25	www.folj.com/puzzles/difficult-logic-problems.htm
26	oneminutelist.com/16-browser-alternatives-to-desktop-programs/index.html
27	www.eclipse.org/index.html
28	www.swimmingpool.com/index.html
29	www.emmaclothes.com/index.html
30	www.arduino.cc/en/Main/Software.html
31	today.java.net/pub/a/today/2004/07/06/3ddesktop.html
32	clotheshor.se/index.html
33	ruzafagallery.com/calendario/index.html
34	www.raspberrypi.org/resources/teach/index.html
35	doodle.com/online-calendar.html
36	www.newprosoft.com/web-content-extractor.htm
37	worryfreelabs.com/about.1.html
38	www.intelligencetest.com/index.htm
39	www.ikea.com/gb/en.html
40	edition.cnn.com/index.html
41	www.neoteo.com/star-wars-the-force-awakens-el-regreso-de-viejos-personajes/
42	riotimesonline.com/index.html
43	www.bbc.co.uk/news/index.html
44	techcrunch.com/gadgets
45	www.turfparadise.com/index.html
46	www.cleanclothes.org/index.html
47	www.afp.com/es/contact.html
48	news.discovery.com/tech/robotics/artificial-intelligences-hawkings-fears-stir-debate-141206.htm
49	www.history.com/index.html
50	detroit.cbslocal.com/2018/12/04/high-school-newspaper-suspended-after-publishing-disruptive-investigation/index.html
51	www.rocklists.com/91x-1983.html
52	www.lashorasperdidas.com/index.html
53	web.mit.edu/institute-events/visitor
54	www.isoc-es.org
55	www.museodelprado.es/index.html
56	www.jdi.org.za/index.html
57	www.u-tokyo.ac.jp/en/about/history.html
58	www.savethechildren.net/what-we-do/our-humanitarian-work.html
59	college.harvard.edu/financial-aid.html
60	www.unicef.org/where-we-work.html
61	www.linuxfoundation.org/about.1.html
62	clinicaltrials.gov/ct2/search/index/index.html
63	cordis.europa.eu/fp7/ict/fire.html
64	www.informatik.uni-trier.de/∼ley/pers/hd/s/Silva_Josep.html
65	parents.berkeley.edu/advice/babies/laundry.html

3 Conclusions

This paper presents a benchmark suite composed of 65 heterogeneous websites. This benchmark suite can be used to test any technique that works with webpages, but it is specially useful for template detection, content extraction and menu detection because it includes a gold standard for them. Concretely, the gold standard identifies for each benchmark the template, the main content and the main menu. Thus, it can be used to evaluate and compare techniques and implementations of these disciplines. The suite is publicly available and free.

References

1. Julián Alarte, David Insa, and Josep Silva. Webpage menu detection based on DOM. In Bernhard Steffen, Christel Baier, Mark van den Brand, Johann Eder, Mike Hinchey, and Tiziana Margaria, editors, *SOFSEM 2017: Theory and Practice of Computer Science - 43rd International Conference on Current Trends in Theory and Practice of Computer Science, Limerick, Ireland, January 16-20, 2017, Proceedings*, volume 10139 of *Lecture Notes in Computer Science*, pages 411–422. Springer, 2017.
2. Julian Alarte, David Insa, Josep Silva, and Salvador Tamarit. Template extraction based on menu information. In *Proceedings of the 9th International Workshop on Automated Specification and Verification of Web Systems (WWV 13), page Article*, volume 5, 2013.
3. Marco Baroni, Francis Chantree, Adam Kilgarriff, and Serge Sharoff. Cleaneval: a competition for cleaning web pages. In *LREC*, 2008.
4. David Insa, Josep Silva, and Salvador Tamarit. Using the words/leafs ratio in the DOM tree for content extraction. *The Journal of Logic and Algebraic Programming*, 82(8):311–325, 2013.