Automatic Assessment of Website Compliance to the European Cookie Law with CooLCheck

Claudio Carpineto
Fondazione Ugo Bordoni
Rome, Italy
carpinet@fub.it

Davide Lo Re
University of Rome 1
Rome, Italy
lore@di.uniroma1.it

Giovanni Romano
Fondazione Ugo Bordoni
Rome, Italy
romano@fub.it

ABSTRACT

We study the problem of automatically assessing whether a website meets the requirements of the Cookie Law, in particular to check that when some tracking cookie is installed the user is asked to give consent to its use. We present a methodology based on cookie disclosure and classification together with identification of natural language consent requests by web information retrieval techniques. This approach performs real time analysis and is very accurate. Using the 500 most popular websites in Italy as a test set, we found that the automatic diagnosis was always correct, except for the case when the consent request was expressed in a language not supported by the system. We also report the results of a systematic evaluation of the 23000 Italian Public Administration websites showing large-scale infringement. Our approach has been implemented as a web application named CooLCheck, which is currently being used by the Italian Data Protection Authority as a support tool for evaluating and monitoring the compliance of websites to the Cookie Law.

Categories and Subject Descriptors

[**Human and societal aspects of security and privacy**]: Privacy protections; [**World Wide Web**]: Web mining; [**Information retrieval**]: Content analysis and feature selection; [**Computing / technology policy**]: Privacy policies

Keywords

Privacy policies, privacy protection, web mining, information retrieval, cookies

1. INTRODUCTION

The European Cookie Law is intended to protect the users privacy by requiring that any website should inform its visitors of what type of information is being gathered. The law applies across the EU, although it is implemented differently in each country. Regardless of the specific legislations, one key common requirement is that the user must be asked for a consent to use the tracking cookies installed by the website. In this paper we study how to automatically detect whether a website installs tracking cookies and whether it asks for the user consent. This problem has not been addressed before, to our knowledge.

Consent requests are typically placed inside a banner displayed at the top or the bottom of the homepage, as exemplified in Figure 1. From a technical point of view, two main questions must be addressed. The first is disclosure and classification of tracking cookies, the second is identification of consent requests. These problems are challenging due to the dynamism of web pages and natural language understanding issues. In the remaining of the paper we describe our methodology, implemented in a system termed CooLCheck (standing for Cookie Law Check), and the results of an experimental evaluation showing the technical feasibility and practical utility of the overall approach.

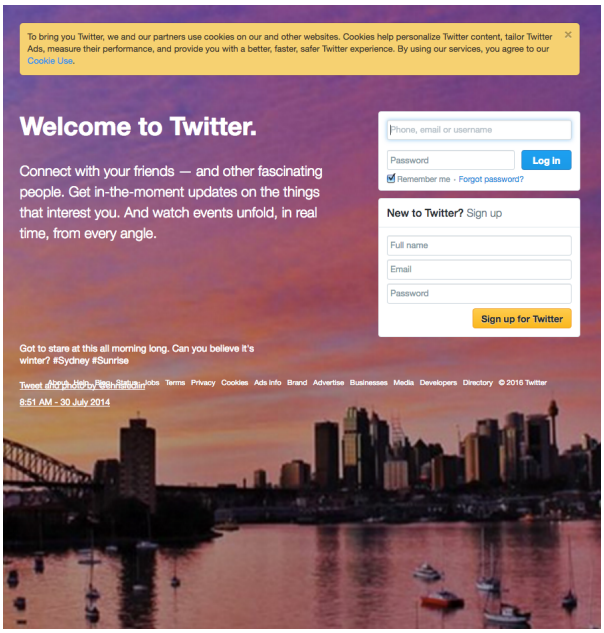

Figure 1: Homepage of Twitter.com with a notice and consent request displayed in the yellow box on top of the page. The text content of the banner can be read in the first item of Table 1.

WPES'16, October 24 2016, Vienna, Austria

DOI: http://dx.doi.org/10.1145/2994620.2994622 ...$15.00.

Table 1: Some consent requests in English and Italian on popular websites.

English
http://www.twitter.com
To bring you Twitter, we and our partners use cookies on our and other websites. Cookies help personalize Twitter content, tailor Twitter Ads, measure their performance, and provide you with a better, faster, safer Twitter experience. By using our services, you agree to our Cookie Use.
http://www.youtube.com
Cookies help us deliver our services. By using our services, you agree to our use of cookies.
http://www.theguardian.com
Welcome to the Guardian. This site uses cookies, read our policy here.
Italian
www.netflix.com
Netflix utilizza i cookie per diversi scopi, tra cui la personalizzazione dell'esperienza dell'utente e della pubblicità online. Scopri di più o modifica le tue preferenze per i cookie. Netflix supporta i principi della Digital Advertising Alliance. Continuando a utilizzare il nostro servizio acconsenti all'utilizzo dei cookie.
http://it.yahoo.com
Utilizzando Yahoo accetti che noi e i nostri partner possiamo impostare dei cookie per personalizzare contenuti e inserzioni pubblicitarie.
http://www.office.com
L'accesso al sito Web implica l'accettazione dell'uso di cookie per analisi, contenuti personalizzati e annunci.

2. DESCRIPTION OF COOLCHECK

CooLCheck takes as input an URL homepage and returns a compliance report. The overall architecture and workflow of CooLCheck is illustrated in Figure 2, with its main components being described in detail in the following sections.

2.1 Cookie disclosure and classification

Modern websites are heavily dynamic, using JavaScript, loading content via XHR requests or using third party content, so that parsing the HTML code is not enough to collect all cookies set by a website. The strategy used by CooLcheck is to create realistic simulation of virtual users visiting Web sites. Through the Selenium framework, CooLcheck runs Firefox and extracts the Document Object Model of the intended webpage once it has been loaded. The hardest task is to determine when the page is fully loaded, because third party cookies are all about loading external resources. To tackle this problem, we iteratively wait for all network requests to be completed, including those triggered by earlier loaded resources. A further complication is represented by an optimization technique used by many (especially news) websites, in which the home-page only contains little text and resources, and the rest of the webpage is loaded when the user scrolls down or interacts with some page elements. To workaround this optimization, we mimick the user behavior; e.g., scrolling down and moving the cursor.

After gathering the contents of the whole cookie jar, cookies are classified according to their purpose. The classification task is necessary because the Cookie Law makes a distinction between analytics and tracking cookies, and only for tracking cookies is the consent request strictly compulsory. As there is no purely technical way of knowing the purpose of a cookie, we used a whitelist approach to filtering (third-party) tracking cookies. This approach is justified by a number of studies showing that the leaders hold a big share of the tracking market [4, 1]. We merged two lists, available at https://services.disconnect.me/disconnect-plaintext.json and http://privacy.aol.co.uk/cookies-list/, resulting in 1403 domains labeled as 'advertising'. We empirically checked that that the final list covered the majority of the websites used in the experimental part. We also found that cookies with the same domain were installed by a number of different sites, following a distribution similar to a power law; e.g. 'doubleclick.net' occurred in 60% of websites.

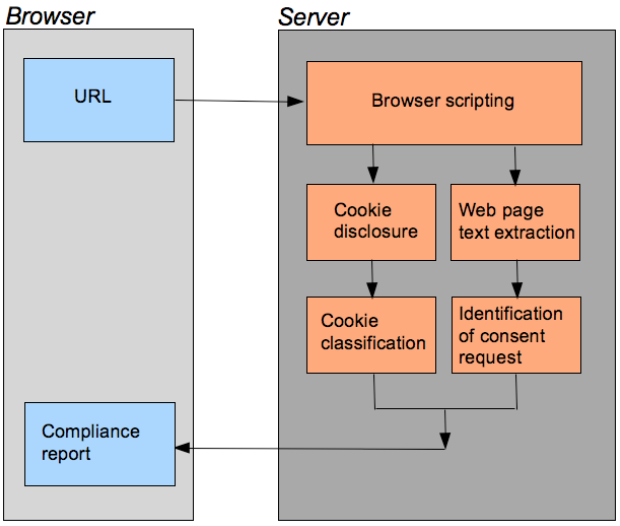

Figure 2: Architecture and flowchart of CooLCheck.

2.2 Text extraction and consent request identification

CooLCheck needs to extract text out from a web page as a preliminary step to find consent requests. This shares many of the technical challenges with the cookie disclosure task, because the text shown at runtime may have been generated using JavaScript or iframes. To address this issue, we extract text not only from the the web page itself, but we also recursively visit the nodes of type text in the DOM tree, including the iframes DOM (if any).

A web page's text may be very large and the identification of a consent request within it is not trivial, in general. Note that a simple heuristic based on the occurrence of the string 'cookie', for instance, would be useless because the word 'cookie' may appear many times in quite different contexts.[1] The problem of recognizing the sought context is difficult because the content of consent requests has not been standardized. As shown in Table 1, it may differ substantially across websites, not only in length and vocabulary but also in the type of information provided.

Table 2 shows statistics about the number of words, number of informative words (i.e., those that were not determin-

[1]There are home pages where the string 'cookie' appears tens of times.

Table 2: Statistics about consent requests in English and Italian.

	Min	Max	Avg	SD
English				
No. of words	12	105	26.15	13.46
No. of informative words	9	55	15.07	7.48
No. of sentences	1	7	2.23	0.69
Italian				
No. of words	8	114	46.32	19.14
No. of informative words	6	62	25.43	10.01
No. of sentences	1	5	2.68	0.75

ers, prepositions, conjunctions, interjections, and pronouns), and number of sentences, drawn from the 500 most popular U.K. and Italian websites (according to Alexa). We also studied the distribution of distinct words across requests. The majority of words occurred only once in only one request, and even the most common words were shared by a limited number of requests (except for the words 'cookie' and 'cookies'). These features may have a negative impact on the adoption of text processing techniques based on bag-of-words representations, as discussed below.

It is more useful to provide a conceptual characterization of consent requests. In a well formed request there may be, at an abstract level, three statements of information: (S1) the site makes use of cookies, (S2) what the cookies are used for, (S3) how a user can accept the use of cookies. On close inspection, one can realise that statement S2 may be specified or not, while sometimes it is provided elsewhere through a link. Even when it is specified in the banner, its formulation is very site specific and difficult to recognize automatically. Also, S1 and S3 are usually both present in the banner, but not always. Furthermore, two statements of information are sometimes merged in a single sentence. These observations motivate the following procedure to assess whether or not a sequence of sentences is a consent request.

We split the problem in three subtasks: (a) text segmentation in homogeneous blocks, (b) selection of candidate blocks, (c) fine-grain analysis of candidate blocks. Step (a) is performed assuming that the content of a request banner is rendered by using some formatting tag which makes it distinguishable from other items in the page. A natural choice are HTML block-level elements (e.g., , <h1> - <h6>, <p>), which always start on a new line and take up the full width available. We thus extract the text between block tags. Step (b) keeps only the blocks where the word 'cookie' or 'cookies' co-occur with a word used to express 'use' (U) or 'agreement' (A), according to the two following lists of words extended with their inflected forms; i.e., noun plural and verb conjugations:

$U_{en} = \{use, employment, utilisation, employ, utilise\}$
$A_{en} = \{agree, accept, consent, be\ happy\ with, be\ okay\ with\}$

The corresponding lists for the Italian language are the following:

$U_{it} = \{uso, utilizzo, impiego, usare, utilizzare, avvalersi\}$
$A_{it} = \{consenso, accettazione, accettare, acconsentire\}$

This operation greatly reduces the number of blocks to be analyzed but it may retain false positives; e.g., navigation menus and privacy policies unrelated to the cookie law. Step (c) applies lexical-syntactic patterns to the blocks selected at step (b). We first segment each block in sentences and split each sentence in a sequence of words. Then we check that there is at least one sentence that corresponds to one of the following regular expressions:

$(w^*)u(w^*)c(w^*)$,
$(w^*)a(w^*)c(w^*)$,

where $u \in U_{en}$, $a \in A_{en}$, c is the word 'cookie' or 'cookies', and w is a word other than u and c. This control is currently performed for English and Italian, supported by a procedure for automatic language identification.

We would like to emphasize that we tried an alternative method based on machine learning to analyze text blocks [3]. Using MLPACK [2], we trained two state-of-the-art classifiers on a set of texts (manually labeled either as consent requests or not) and evaluated their performance on a set of test texts. However, the predictions made by the classifiers were not accurate (e.g., yielding both false positives and false negatives), probably due to the sparseness and varying length of data as as well as to the limited size of the training set.

We also experimented with a third method, based on language modeling [3]. We built a language model of the consent request from a set of training instances, and then measured the difference between the word probability distribution of the consent model and that of the text block to be analysed using Jensen-Shannon divergence as a similarity measure. However, we found that many truly consent texts exhibited low similarity to the consent model while many non-consent texts showed high similarity, thus confirming the inadequacy of probabilistic methods for the task at hand.

2.3 User interface and system implementation

In response to a URL, the systems outputs a report stating whether tracking cookies were detected and, in the affirmative case, whether a notice and consent banner was found or not. The user has an option to see the list of tracking cookies and/or the content of the banner. The system has been implemented as a web application, available at http://spai.fub.it with a password-protected access.

3. EXPERIMENTS WITH COMMERCIAL WEBSITES

We used the Alexa list of the 500 most popular websites in Italy, as of December 2015. The URLs were extracted from Alexa website and given as an input to CooLCheck, setting a time out for idle sessions. For each URL, we saved the corresponding page and the output of CooLCheck, including the list of tracking cookies and the consent request extracted by the system. The system detected at least one known tracking cookie for 324 sites, confirmed by a manual inspection. This can be seen as a lower bound of the number of sites in the sample subject to the Cookie Law.

Once the system has detected some tracking cookie, it must decide whether or not there is a banner with a con-

Table 3: Classification performance on the 324 websites (in the Alexa top 500 Italian sites list) installing tracking cookies.

	With banner	*Without banner*
System's prediction	219	105
Errors	0	3

Table 4: Analysis of cookies and consent banners in the Italian Public Administration websites.

Sites with analyzable content	*Sites with cookies*	*Sites with tracking cookies*	*Sites with consent banner*	*Sites w/o consent banner*
17073	5269	1930	790	1140

sent request. There are two possible types of errors: when the system does not recognize a truly consent request (false negative), and when it detects a false consent request (false positive).

We evaluated the accuracy of the system's prediction by manually labeling the 324 sites installing tracking cookies as 'with banner' or 'without banner', and then measuring the performance on this gold standard. The results are reported in Table 3. When the system recognized a consent request it was always right. Of course, this does not mean that it is not possible to produce a text that is erroneously considered as a consent request. More simply, this event never happened in the test set. By contrast, for three times a banner with a consent request was present but it was not recognized. This is not surprising because it is easy to produce a truly consent request misleading the system. On the other hand, we checked that the three errors were due to the use of languages that were not supported by CooLCheck (i.e., German and French), whereas there were zero false negatives in Italian and English.[2]

On the whole, these findings suggest that the output of CooLCheck is very reliable and that our modelization can effectively match the conceptual structure of most existing consent requests.

4. EXPERIMENTS WITH ITALIAN PUBLIC ADMINISTRATION WEBSITES

We used the Index of Italian Public Administrations, maintained and made available as open data by the Agency for Digital Italy (AGID). The Index can be downloaded from http://www.indicepa.gov.it/documentale/index.php. At the end of 2015, the dataset contained about 23000 administrations, with very noisy URLs. We performed a number of pre-processing steps to remove ill-formed and unresolved URLs as well as to cope with anomalous behaviors and non-standard homepage implementation techniques which make automatic content extraction impossible or very difficult; e.g., 404 errors, redirect, refresh, frames, sublinks. After this treatment, we were left with 17073 sites, which were given as an input to CooLCheck. The outputs of CooLCheck were saved to analyze the results.

In Table 4, we report the statistics about the number of sites installing cookies, the type of cookies, and the presence of a consent banner. The results show that only a fraction of websites install cookies and only a small fraction install tracking cookies, as expected, although the number of sites installing tracking cookies was not negligible. Table 4 also shows that the majority of the latter sites did not show a consent banner. On an absolute scale, we found that at least 1140 websites were not compliant, which indicates that the process of adoption of the Cookie Law is still slow.

These results were generated automatically. To gain some insights about their reliability we performed a small validation test by randomly selecting 100 sites installing tracking cookies and then manually labeling them as as 'with banner' or 'without banner', similar to the procedure described above to evaluate the system's performance. We found consistent results with those achieved on commercial webistes, with 100% accuracy (thanks to the lack of non-Italian consent requests in the selected sample).

[2]This means that at least 102 sites out of 500 were not compliant to the Cookie Law as of December 2015.

5. CONCLUSIONS

We presented a methodology for automatic assessment of the compliance of websites to the Cookie Law. The methodology is fast, accurate, and can be easily adapted to different languages. It is suitable for monitoring the process of adoption of the law over time and on a large scale, and also for identifying specific non-compliant websites, which by the law now in force shall be liable to be fined. The prototype is currently being used by the Italian Data Protection Authority as a support tool. Our experimental study highlights that websites are slowly meeting the requirements posed by the Cookie Law, with the institutional sites (at least in Italy) being much later in adopting than commercial sites.

Future work includes extension to and experimentation with other languages, as well as use of specific information extracted from the privacy policy (e.g., the name of the owner of the website) to enable more fine-grained assessment of compliance to the law.

6. REFERENCES

[1] A. Cahn, S. Alfeld, P. Barford, and S. Muthukrishnan. An empirical study of web cookies. In *WWW '16: Proceedings of the 25th international conference on World Wide Web, Montreal, Canada*, pages 891–901, 2016.

[2] R. R. Curtin, J. R. Cline, N. P. Slagle, W. B. March, P. Ram, N. A. Metha, and A. G. Gray. MLPACK: A Scalable C++ Machine Learning Library. *Journal of Machine Learning Research*, 14:801–805, 2013.

[3] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[4] F. Roesner, T. Kohno, and D. Wetherall. Detecting and defending against third-party tracking on the web. In *Proceedings of the 9th USENIX Symposium on Networked. Systems Design and Implementation, San Jose, California, USA*, pages 12–12, 2012.

7. ACKNOWLEDGMENTS

We gratefully acknowledge funding by the Italian Ministry of Economic Development and support by the Italian Data Protection Authority. We would also like to thank Alessandro Mei and Eugenio Nemmi for their collaboration.