Classification of Web History Tools Through Web Analysis

João Rafael Gonçalves Evangelista(✉),
Dacyr Dante de Oliveira Gatto, and Renato José Sassi

Universidade Nove de Julho – UNINOVE, São Paulo, SP 01504-000, Brazil
jrafal607@gmail.com

Abstract. Web pages may contain various types of sensitive information exposed, such as user login information. Even after these pages have been corrected, the sensitive information, once exposed, can be found through the web history tools. These tools make snapshots of web pages, that is, capture the state of the pages in the most varied periods. Although these tools are widely used, it is not known which web history tool is the most accessed. A method to find out which web history tool is the most accessed is by means of classification using the web analytics technique. Therefore, in view of this scenario, the objective of this work was to classify web history tools through web analysis. The methodology used was the descriptive with quantitative approach. As for the technical procedures, this work is characterized as experimental to verify if the technique of web analysis is able to classify web history tools. The results show that the technique of web analysis produces indicators capable of classifying the web history tools by the total number of accesses received.

Keywords: Web analytics · Web history tools · Open Source Intelligence · Web technologies

1 Introduction

People are increasingly sharing information on the internet. Practices such as publishing employee lists on organizational web pages allow people with bad intentions to identify easily company employees among millions of social media users [1].

In addition to the information shared on the internet by users, other sensitive information may also be exposed. A badly configured web page can leave unprotected information such as user logins, database settings, information about active servers in the domain, services in operation, and other types of sensitive information.

Even after correcting the web pages, we can find the sensitive information exposed using web history tools. Web history tools work like a repository, collecting and archiving web pages periodically [2].

The concept used to describe the collection of information from open sources, as well as the techniques and tools used to acquire this information is Open Source Intelligence (OSINT) [3].

A. Moallem (Ed.): HCII 2019, LNCS 11594, pp. 266–276, 2019.
https://doi.org/10.1007/978-3-030-22351-9_18

Web history tools are available on the internet, but not known which is the most accessed. One method to find out which web history tool is most accessed is through web analytics. Web analytics is a technique that extracts indicators about user interaction with a web page.

Web analytics encompasses a variety of activities, such as measuring web traffic, collecting large volumes of data, analyzing web performance, mining corporate data, and visualizing data strategies [4].

Web analytics provide indicators that can analyze and classify pages on the Internet, for example: The total number of hits received in a given time period, the type of device that accessed the page, the average duration of each access, or even the average number of pages accessed.

In view of this scenario, the objective of this work was to classify web history tools through web analysis technique using the Access Rank indicator, in order to find out which are the most accessed web history tools.

2 Theoretical Background

2.1 Web History Tools

With the evolution of the internet, it is simpler to search information. You can search for any word or phrase, and in a few moments, search engines that are capable of generating results. In addition to searching the internet, another important factor for acquiring information is automated capture [5].

For this, systems and tools are developed to facilitate the proper archiving of content. Among the tools available on the internet, we have the web history tools or archiving tools of web pages [5, 6].

Web History tools have the capability to recover and access previously archived Web pages. For your use, it is enough that the user provides the URL of the desired page and navigate among those archived by the web history tool [7].

Internet Archive [8], for example, is the first web history tool to archive web pages. The tool holds more than 360 billion web pages with files since 1996, making it possible to go back in time to view previous versions of archived web pages [6].

To analyze and evaluate web pages in a determined period, web history tools are commonly used. [9] for example, present the use of the web history tool Wayback Machine [10] to highlight the growth in store sales following the introduction of new policies in Italy.

The authors [11] address another application; they present the use of the historic web tool Wayback Machine [10] to confirm the historical accuracy of a classification of informal financial systems, known as shadow banks, in fintech or non-fintech.

2.2 Open Source Intelligence (OSINT)

Open Source Intelligence (OSINT) involves the collection, analysis, and use of data from open sources for intelligent purposes. So, it can be understood that OSINT involves locating, selecting and extracting information from open sources, such as Twitter and Facebook, and, finally, analyzing extracted information [12, 13].

According to the methodology of tests of information security PTES Technical Guideline [14], OSINT, in the simplest of terms, is to find and analyze open sources. In the area of information security, this information collection process aims to produce current and relevant information that is valuable to an attacker or a competitor.

OSINT can act in several types of open sources, such as global media, blogs on the internet, web pages with government reports, satellite images, academic works, Wikipedia, YouTube and Facebook, as well as a series of other information made available through internet and other media resources [15].

The information discovered by OSINT is defined by [16] as information that is publicly available for anyone to acquire this information legally by request, purchase or observation. Usually the practice of Open Sources Intelligence is seen in positive terms, particularly as a conventional data collection method that does not violate human rights [17].

[15, 17–19] present other concepts that address the collection of information, where OSINT acts directly with each one. The authors as disciplines of intelligence approach the concepts. Table 1 describes the intelligence disciplines along with their ID and description.

Table 1. Intelligence disciplines.

ID	Intelligence disciplines	Description
01	COMINT	Communication Intelligence
02	CULTINT	Cultural Intelligence
03	DFINT	Digital Forensics Intelligence
04	ELINT	Electronic Intelligence
05	GEOINT	Geospatial Intelligence
06	HUMINT	Human Intelligence
07	IMINT	Image Intelligence
08	MARKINT	Market Intelligence
09	MASINT	Measurement and Signature Intelligence
10	SIGINT	Signal Intelligence
11	SOCMINT	Social Media Intelligence
12	TECHINT	Technical Intelligence
13	TELINT	Telemetry Intelligence

The practice of data collection has been discussed since 1941 when an effort to monitor German and Japanese radio broadcasts was launched with the creation of the Foreign Broadcast Monitoring Service, an organization that later became the Open Source Center [16].

From the creation of the Open Source Center to the present, numerous tools and techniques for collecting information from open sources have emerged that self-tune the search and analysis [20]. For example, address the practice of OSINT tools such as Google, Shodan, Sensys, theHarvester, Z-map and Carrot2 to find vulnerabilities in a system.

2.3 Web Analytics

Web analytics is a technique that involves the use of softwares that collect data about the behavior of users while they browse the internet. You get the data by tracking the mouse clicks or even by requesting information for the users [21].

The web analytics technique is responsible for helping to understand how users interact with web pages and mobile applications, automatically registering aspects of user behavior, and then combining, analyzing, and transforming behavior into data [22].

To run a web analytics on a web page, you must have a question or questions to answer. In Fig. 1, its show how the answers are not always as simple as we expect, and when we look at an area, we can discover new discoveries along the way. Semi-structured analysis involves data collection, transformation and analysis [22].

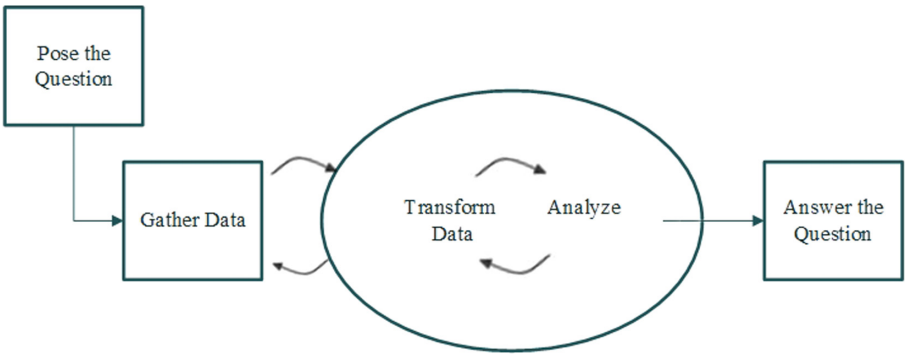

Fig. 1. The flow for performing a web analysis.

An example of web analytics application, is to use it to know information on where web traffic is coming up, what types of products users are interested in, what types of keywords users are typing in search engines for access a website [23].

Web analytics can also analyze user-generated content on social media, such as product reviews. The organization responsible for the product can use these opinions as feedback on their product to improve it, while the customer can use the same opinions to decide whether to buy the product or not [24].

Another example of application addressed by [25] is on the use of web analytics to perform performance measurement in digital marketing. Already [26] presents the web analysis to obtain and evaluate the performance indicators generated by university students inside a library [27]. Present the use of web analysis through the tool Similarweb [28] to develop a categorization of web pages. While [29] also used the Similarweb tool to explore the interest and use of the PhET website in a university.

3 Materials and Methods

3.1 Characteristics of the Studied Process

This research is of the descriptive type with a quantitative approach, since it involves the application of the web analysis in tools of historical web.

The descriptive research has as main objective the description of the characteristics of a certain population or phenomenon or the establishment of relations between variables. Its most significant characteristics are the use of standardized data collection techniques [30].

As for the technical procedures, this research is experimental, as it verifies if the web analytics application is able to classify the web history tools. The experimental research consists in determining an object of study, selecting the variables that would be able to influence it, defining the forms of control and observation of the effects that the variable produces on the object [30].

As for the theoretical background, a bibliographic survey was carried out using the key-words: "Osint", "Open Source Intelligence", "Web History Tool", "Archive Internet" and "Web Analytics" in the bases: IEEE Digital Library, Scopus, Science-Direct, EmeraldInsight, Portal Capes and ProQuest.

3.2 Computational Experiments

The computational experiments has three steps, shown in Fig. 2. In step A, we searched for OSINT Toolkits. In step B, is performed an extraction and evaluation of web history tools and web analysis tools. Finally, in the last step, step C, is performed a classification of web history tools.

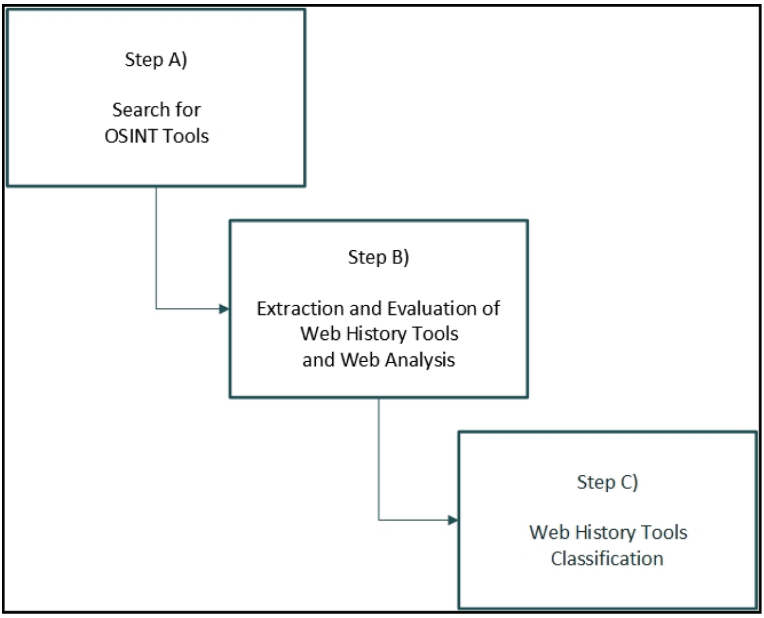

Fig. 2. The figure shows the steps of the computational experiments of this work.

- **Step A: Search for OSINT Tools:** We searched for OSINT toolkits to extract web history tools and web analytics tools. For this, we used the search engines: Biznar [31], Carrot2 [32], Google [33] and Metabear [34]. The OSINT toolkit selected was the "OSINT, Tools and Resources Handbook" of the I-Intelligence company [34].
- **Step B: Extraction and Evaluation of Web History Tools and Web Analysis:** We looked at the OSINT toolkit defined in phase A by web history and web analytics tools.

For the Web History tools, you extracted all the tools that appeared in the OSINT toolkit in the "Web History and Site Capture" category. For the validation of extracted web history tools, selected the link of each tool and performed a search with the URL of the social network domain LinkedIn [35].

For the web analytics tools, we extracted tools that could perform an online web analysis without the need for installation. For evaluation criteria, we verified which web analytics tools could be executed free of charge for a minimum period of 30 days. Selected the tools, a web analysis was done with each tool in the social network LinkedIn [35].

- **Step C: Web History Tools Classification:** A Web analysis was performed on the web history tools extracted in step B with the tool Similarweb, also extracted in step B. After the web analysis, we selected the desired indicators, and finally, we created an attribute "Rank Access", to classify the web history tools by the total number of accesses received.

4 Results and Discussions

In this section, the results of the computational experiments are presented and discussed. The experiments has three steps:

- **Step A: Search for OSINT Tools:** We searched for OSINT toolkits to extract web history tools and web analytics tools. For this, we searched the key words: "OSINT framework", "OSINT Toolkit" and "OSINT platform" in search engines: Biznar [31], Carrot2 [32], Google [33] and Metabear [34].

For the selection criteria of the OSINT toolkits, it was verified which of the toolkits found would bring the greatest amount of OSINT tools grouped by categories and which of them appear in periodicals or books. In Table 2, the OSINT toolkits found, along with their ID, Type and URL.

The toolkits Osintframework and Inteltechniques presented few or even none categorized as web analytics. Thus, the "OSINT, Tools and Resources Handbook" toolkit of the company I-Intelligence [20] was selected for the variety and quantity of categorized tools.

Table 2. OSINT toolkits.

ID	OSINT Toolkit	Type	URL
01	Osintframework	Web Page	https://osintframework.com/
02	Inteltechniques	Web Page	https://inteltechniques.com/menu.html
03	Osint Tools and Resources Handbook	Ebook (PDF)	https://www.i-intelligence.eu/wp-content/uploads/2016/11/2016_November_Open-Source-Intelligence-Tools-and-Resources-Handbook.pdf

– **Step B: Extraction and Evaluation of Web History Tools and Web Analysis:** We looked at the OSINT toolkit defined in phase A by web history and web analytics tools. For Web History tools, all tools from the "Web History and Website Capture" category found in the OSINT toolkit was extracted. Then, it was verified which tools could be executed online, without the need of installation. Table 3 shows the web history tools extracted along with your ID and URL.

Table 3. Web history tools.

ID	Web history tool	URL
01	Archive.is	http://archive.is
02	Archive.fo	http://archive.fo
03	CashedPages	http://www.cachedpages.com
04	CachedView	http://cachedview.com
05	Common Crawl	http://commoncrawl.org
06	Screenshots.com	http://www.screenshots.com
07	Wayback Machine	http://archive.org/web/web.php

To evaluate previously extracted web history tools, you have accessed each tool and searched the LinkedIn social network domain. All selected tools have managed to bring historical pages of the social network.

For the web analytics tools, we extracted tools that could perform an online web analysis without the need for installation. For selection criteria, it was verified which web analytics tools could be executed free of charge for a minimum period of 30 days. The Table 4 presents the extracted web analytics tools along with their ID and URL.

Table 4. Web analytics tools.

ID	Web analytics tools	URL
01	Similarweb	https://www.similarweb.com/
02	Crunchbase	https://www.crunchbase.com

For the evaluation of the web analysis tools extracted, a web analysis was performed on the LinkedIn social network with each of them. The Crunchbase tool [37] provided much more qualitative rather than quantitative information, being unable to find indicators about users' use of the social network.

The web analytics tool Crunchbase [37] provided values such as: Name of the founders, e-mail addresses of some employees, investors, links to social media, current price of the organization, name of the organization's team, among others.

In addition to the Crunchbase tool [37], the tool Similarweb [28] provided quantitative information on user interaction with the social network, such as: Global rank, total number of accesses received, average monthly accesses, average access time and rate mean of rejection. Thus, the tool Similarweb [28] was selected to perform the web analysis in this work.

The indicators selected to perform the classification of web-based tools by the total number of accesses received were Global rank and total number of accesses received between June 2018 and August 2018, the most recent date available in the tool at the time of execution of this work.

– **Step C: Web History Tools Classification:** A web analysis was performed on the web history tools extracted in Phase B with the tool Similarweb.

The following indicators selected were Total accesses received by the tool between June 2018 and August 2018, in addition to the global rank reported by the tool Similarweb. To perform classification, the attribute "Access Rank" was created based on the indicators selected previously.

Table 5 presents the web history tools sorted by rank access.

Table 5. Web history tools classify by access rank.

ID	Web history tool	Rank global	Total of accessed received	Access rank
07	Wayback Machine	237	301,3 M	1
01	Archive.is	3193	58,74 M	2
02	Archive.fo	9848	18,33 M	3
03	CashedPages	337600	437,821	4
04	CachedView	245786	388,303	5
05	Common Crawl	1199003	78,918	6
06	Screenshots.com	3281034	31,769	7

The web history tool that presented the highest number of accesses received was the Wayback Machine with 301.3 million accesses between June 2018 and August 2018. The Fig. 3 shows the graph of the web history tools and the total accesses of each tool received between June 2018 and August 2018.

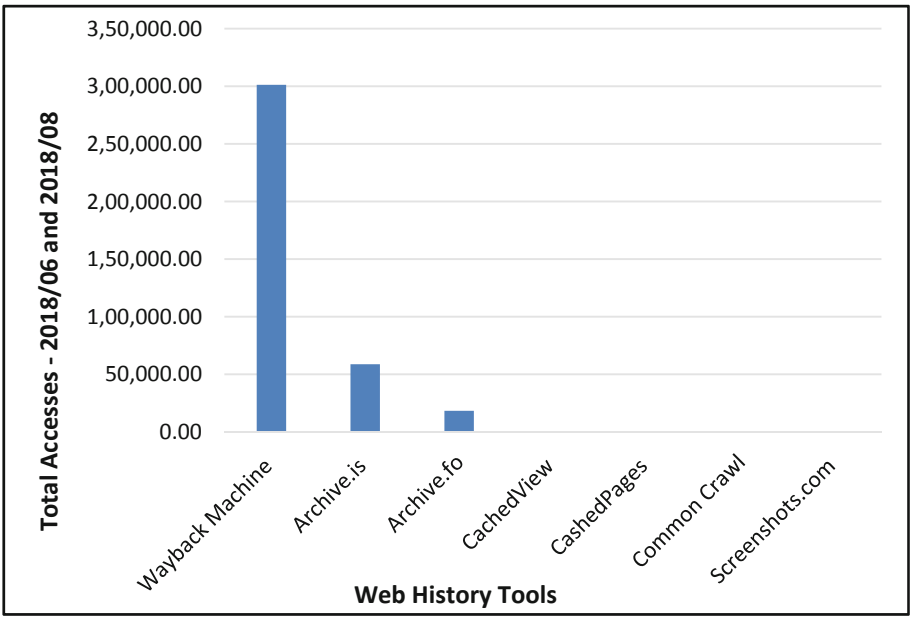

Fig. 3. The figure shows the total number of hits received by web history tools between June 2018 and August 2018.

5 Conclusion

In this work, we approached the classification of tools of historical web by means of the technique of web analysis with the objective of evidencing the ones that are the most accessed.

The application of the web analytics through the tool Similarweb generated important indicators, of which, were used the "Total of incomes received" and "Global Rank". These Indicators, which were able to classify the web history tools by the total access received. Thus, one can see which Web-based tools are the most accessed by the number of accesses received.

As a contribution of this work, the technique to classify the web history tools can be applied not only to classify OSINT online tools, but other types of web pages, in different areas, such as marketing and education. In addition, this work also presents the OSINT toolkits, where one can explore the other categories of tools, such as search engines, geo-localization tools, among others.

As a suggestion for future work, it will be interesting to continue the evaluation of the OSINT tools, since incorporating other categories and not just the web history tools, it becomes possible to develop a framed OSINT framework, tool or platform or information security using the most accessed tools.

References

1. Edwards, M., Larson, R., Green, B., Rashid, A., Baron, A.: Panning for gold: automatically analysing online social engineering attack surfaces. Comput. Secur. **69**, 18–34 (2017)
2. Li, Y., Arora, S., Youtie, J., Shapira, P.: Using web mining to explore triple helix influences on growth in small and mid-size firms. Technovation **76**, 3–14 (2018)
3. Glassman, M., Kang, M.J.: Intelligence in the internet age: the emergence and evolution of open source intelligence (OSINT). Comput. Hum. Behav. **28**(2), 673–682 (2012)
4. Čegan, L., Filip, P.: Webalyt: open web analytics platform. In: 27th International Conference, RADIOELEKTRONIKA 2017, pp. 1–5. IEEE (2017)
5. Singhal, A., Srivastava, P., Dawn, S.: Computational transformation from web to Ebook archiving. In: 5th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering, UPCON 2018, pp. 1–6. IEEE (2018)
6. Alnoamany, Y., Alsum, A., Weigle, M.C., Nelson, M.L.: Who and what links to the internet archive. Int. J. Digit. Libr. **14**(3–4), 101–115 (2014)
7. Kanhabua, N., Kemkes, P., Nejdl, W., Nguyen, T.N., Reis, F., Tran, N.K.: How to search the internet archive without indexing it. In: Fuhr, N., Kovács, L., Risse, T., Nejdl, W. (eds.) TPDL 2016. LNCS, vol. 9819, pp. 147–160. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43997-6_12
8. Internet Archive homepage. https://archive.org/. Accessed 31 Aug 2018
9. Carrieri, V., Madio, L., Principe, F.: Light cannabis and organized crime: evidence from (Unintended) liberalization in Italy. Eur. Econ. Rev. **113**, 63–76 (2019)
10. Wayback Machine homepage. https://archive.org/web/. Accessed 31 Aug 2018
11. Buchak, G., Matvos, G., Piskorski, T., Seru, A.: Fintech, regulatory arbitrage, and the rise of shadow banks. J. Financ. Econ. **130**(3), 453–483 (2018)
12. Howells, K., Ertugan, A.: Applying fuzzy logic for sentiment analysis of social media network data in marketing. Procedia Comput. Sci. **120**, 664–670 (2017)
13. Koops, B., Hoepman, J., Leenes, R.: Open-source intelligence and privacy by design. Comput. Law Secur. Rev. **29**(6), 676–688 (2013)
14. Penetration Testing Execution Standard – Technical Guidelines homepage. http://www.Pentest-Standard.Org/Index.Php/PTES_Technical_Guidelines. Accessed 25 Aug 2018
15. Quick, D., Choo, K.R.: Digital forensic intelligence: data subsets and open source intelligence (DFINT + OSINT): a timely and cohesive mix. Future Gener. Comput. Syst. **78**, 558–567 (2018)
16. Clarke, C.S.: Open source intelligence. An Oxymoron or real intelligence? Marine Corps Gazette **99**(8), 22 (2015). Professional Journal of U.S. Marines
17. Hribar, G., Podbregar, I., Ivanuša, T.: OSINT: a "grey zone"? Int. J. Intell. Counterintelligence **27**(3), 529–549 (2014)
18. Akhgar, B., Bayerl, P.S., Sampson, F.: Open Source Intelligence Investigation: From Strategy to Implementation. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-47671-1
19. Chauhan, S., Panda, N.K.: Hacking Web Intelligence: Open Source Intelligence and Web Reconnaissance Concepts and Techniques, 1st edn. Syngress (2015)
20. Lee, S., Shon, T.: Open source intelligence base cyber threat inspection framework for critical infrastructures. In: Future Technologies Conference (FTC) 2016, pp. 1030–1033. IEEE (2016)
21. Salini, A., Malavolta, I., Fabrizio.: Leveraging web analytics for automatically generating mobile navigation models. In: International Conference on Mobile Services (MS), pp. 103–110. IEEE (2016)

22. Beasley, M.: Practical Web Analytics for User Experience: How Analytics Can Help You Understand Your Users, 1st edn. Newnes, Oxford (2013)
23. Reshma, K., Rajendran, V.V.: An enhanced approach for querying integrated web analytics ontology using Quepy. In: International Conference on Intelligent Computing and Control (I2C2), pp. 1–6. IEEE (2017)
24. Rathan, M., Vishwanath, R.H., Murugeshwari, P., Sushmitha, H.M.: Every post matters: a survey on applications of sentiment analysis in social media. In: International Conference on Smart Technologies for Smart Nation, SMARTTECHCON 2017, pp. 709–714. IEEE (2017)
25. Järvinen, J., Karjaluoto, H.: The use of web analytics for digital marketing performance measurement. Ind. Mark. Manage. **50**, 117–127 (2015)
26. Fagan, J.C.: The suitability of web analytics key performance indicators in the academic library environment. J. Acad. Librarianship **40**(1), 25–34 (2014)
27. Choi, D., Han, J., Chun, S., Rappos, E., Robert, S., Know, T.T.: Bit.Ly/Practice: uncovering content publishing and sharing through URL shortening services. Telematics Inform. **35**(5), 1310–1323 (2018)
28. Similarweb homepage. https://www.similarweb.com/pro. Accessed 2 Sept 2018
29. Zhang, M.: Who are interested in online science simulations? Tracking a trend of digital divide in internet use. Comput. Educ. **76**, 205–214 (2014)
30. Gil, A.C.: Métodos E Técnicas De Pesquisa Social, 6th edn. Ediitora Atlas SA, São Paulo (2008)
31. Biznar homepage. https://biznar.com/biznar/desktop/en/search.html. Accessed 1 Sept 2018
32. Carrot2 homepage. http://Search.Carrot2.Org/Stable/Search. Accessed 1 Sept 2018
33. Google homepage. https://Www.Google.Com.Br/. Accessed 1 Sept 2018
34. Metabear homepage. http://www.metabear.com/. Accessed 9 July 2018
35. I-Intelligence homepage. https://www.i-intelligence.eu/wp-content/uploads/2016/11/2016_November_Open-Source-Intelligence-Tools-and-Resources-Handbook.pdf. Accessed 16 July 2018
36. LinkedIn homepage. https://br.linkedin.com/. Accessed 1 Sept 2018
37. Crunchbase homepage. https://www.crunchbase.com/. Accessed 2 Sept 2018