http://doi.org/10.35784/iapgos.898

RANKING OF WEBSITES CREATED WITH THE USE OF ISOWQ RANK ALGORITHM

Mariusz Duka
Silesian University of Technology, Faculty of Automatic Control, Electronic and Computer Science, Gliwice, Poland

Abstract. *The purpose of this article is to present the new ISOWQ Rank ranking algorithm for the technical assessment of website quality. For evaluation purposes, the algorithm takes into account the IT technologies used on a website, compliance of the source code with the applicable standards and the structure of the text content. The paper also includes the results of comparative ranking algorithms.*

Keywords: page ranking, ISOWQ Rank, MOZ, PageRank, SEO

RANKING WITRYN INTERNETOWYCH STWORZONY Z WYKORZYSTANIEM ALGORYTMU ISOWQ RANK

Streszczenie. *Celem artykułu jest prezentacja nowego algorytmu rankingowego ISOWQ Rank do technicznej oceny jakości strony internetowej. Algorytm do oceny bierze pod uwagę wykorzystane w serwisie WWW technologie, zgodność kodu źródłowego z obowiązującymi standardami oraz strukturę tekstu. W artykule zamieszczone są wyniki badań porównawczych algorytmów rankingowych.*

Słowa kluczowe: ranking stron, ISOWQ Rank, MOZ, PageRank, SEO

Introduction

Since 1991, when the first structures of the HTML programming language appeared, together with the first website (info.cern.ch), the Internet has been filled with over 1.4 billion websites on 233 million unique domains. Since 1995, the British company, Netcraft, has been searching Internet resources and examining IT technologies and software used to build websites on the servers of hosting service providers. Based on the analysis of the IP address and source code, Netcraft calculates the number of active websites, i.e. presenting unique content. It excludes the "under construction" web services, redirections or any domains indicating same content, e.g. ones with a registrar enabling 'the parking service'. Based on the cyclical analysis, it is estimated that almost 200 million active websites are accessible on the Internet.

A dynamic development of Internet resources is shown in Figure.1

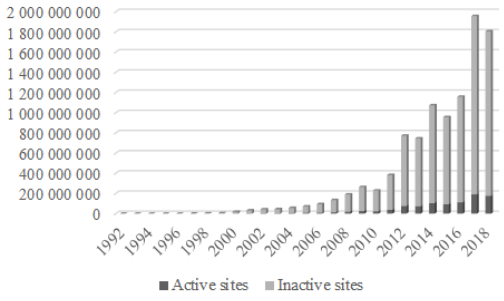

Fig. 1. The number of websites with unique content on the Internet (based on the data from Netcraft)

The growing number of websites has prompted the creation of search engines, i.e. the systems that help Internet users find specific information on the Web. The data collected in these systems is based on an analysis of Web pages in terms of their content and of the hyperlink network topology. Search engines like Google, Yahoo! or Bing, have had to develop their own algorithms to allocate value to the analysed websites to affect their ranking in the search engines results. The leading algorithm determining the quality of the website was the PageRank algorithm, developed by the creators of Google in 1998. In the following years, some other IT companies have also created their own algorithms to give websites a certain ranking value and thus determine their quality.

The popular MOZ ranking system has been created using the DA (Domain Authority) and PA (Page Authority) parameters. It is based largely on the quality of inbound links.

The quality of such links is determined by their source (catalogues, blogs, forums, social media, search engines) and their internal linking value.

This paper presents the ISOWQ Rank algorithm used in the ISOWQ (International Studies of Website Quality) ranking system, to analyse websites available under all 243 national domains (ccTLD) and the European .eu domain. Unlike the competing ranking algorithms, focusing largely on linking, such as PageRank or MOZ, the ISOWQ Rank algorithm takes into account the technologies used on the website, compliance of the source code with applicable standards (including W3C), content and text structure, as well as external rankings (MOZ, Alexa Rank). Other leading ranking systems i.e. Google PageRank, MOZ and Alexa Rank are also described here. A separate chapter presents the ISOWQ Rank algorithm and the IT system for which it was designed. The paper concludes with the presentation of the findings for the correlation between the ISOWQ Rank and MOZ.

1. Ranking methods

1.1. Google PageRank

The most well-known algorithm that determines the quality of a website is PageRank, used by Google. It was developed by Sergey Brin and Larry Page and described in the publication "The Anatomy of a Large-Scale Hypertextual Web Search Engine" [1].

The PageRank algorithm is based on the idea that the substance and quality of the site is evidenced by the number of references from other sites i.e. the value of the page measured by the number of hyperlinks pointing to that page from other pages [2].

The PageRank algorithm is calculated according to the following formula:

$$PR(A) = (1-d) + d\left(\frac{PR(T1)}{C(T1)} + \cdots + \frac{PR(Tn)}{C(Tn)}\right) \qquad (1)$$

where: *PR(A)* is the PageRank value of the side *A*, *PR(T1)* is the PageRank value of the side *T1*, *C(T1)* is the number of links pointing to the side *T1*, *d* is the dumping factor, in the range of $0<d<1$, usually accepted as 0.85. PageRank is calculated as the total value of all pages pointing to the selected page divided by the number of links on each of these pages.

The PageRank value is measured on the scale from 0 to 10, which means that the higher the value, the better the website quality. The value above 6 are usually reserved for professional sites, such as Facebook, Yahoo! or Twitter. Note that keywords, text or source code do not affect the ranking value, relevant are links only. The highest score in PageRank is usually obtained by the home page of the website. One of the reasons being, it is most often linked to and from other websites. However, there is also a

possibility that a higher score in PageRank may get another page of this website, which due to its rich content and substance, is linked more. An example of this may be a subpage containing a blog [2].

Google still uses the PageRank algorithm, but, from 2013, the data are not publicly available, which makes it go down in history for the SEO (Search Engine Optimization) industry.

1.2. MOZ DA and PA

Founded in 2004, MOZ, offers the marketing analytics tools at moz.com, It has also developed its own algorithm to give websites a ranking value. The intention of the MOZ ranking is to predict the position of a given website in search engines. Each page is awarded a so-called "authority" ranging from 0 to 100 points. The higher the authority of a domain or a page, the higher the chances of getting a good position in search engines results. The MOZ ranking is calculated for the entire website – DA (Domain Authority) and for a specific page – PA (Page Authority) [3].

In addition to analysing the quality of inbound links, the algorithm takes into account many additional factors in order to calculate the "authority". "Authority" is estimated on a 100-point logarithmic scale, which means that an increase from 20 to 30 points is much easier than that from 70 to 80 points. The algorithm structure is strictly confidential and protected by the MOZ business secrets.

1.3. Alexa Rank

Founded in 1996, Alexa Internet, currently controlled by Amazon, provides information related to Web traffic. The Alexa ranking is based on the amount of traffic generated by Internet users on a particular website. It is measured by using specific software (plugins for the Chrome or Firefox browser) and calculated on the basis of several factors, such as the number of people who visited the website and the frequency of these visits [4].

The Alexa ranking value is interpreted as an indicator of the popularity of a websites on the Internet. The lower the indicator is, the more popular a particular website is on the Internet. It is assumed that if the Alexa Rank value is below 100,000, the website is very popular among Internet users.

2. ISOWQ Rank algorithm

The ISOWQ Rank algorithm allocates a specific value to its analysed websites. It does so by measuring the three qualities. It is calculated according to the following formula:

$$IR = \frac{PM+PK+PT}{3} \quad (2)$$

where *PM* – number of points for technologies used and their ranking positions, *PK* – number of points for the source code optimisation, *PT* – the number of points for content and text structure.

The ISOWQ Rank value is calculated on a scale of 0 to 20 points. To calculate the score for the technologies used and ranking positions, ISOWQ Rank takes in to account such factors as MOZ and Alexa Rank rankings, the number of incoming links, the use of social media plugins, SSL encryption, geolocation, public e-mail addresses on the website and the registration of the hosting servers' IP addresses in DNSbl databases.

Points for the search engine optimisation are calculated using several other factors. These include: the use of HTML tags for SEO, optimization of the size of the code, the ratio of text size to the source code, the use of cascading styles and JavaScript in external files, the ratio of the number of external to internal links, the use of microformats, the number of URLs returning the error code.

The value for the content and structure of the text is affected by the size and formatting of the visible text on the page (the use of headers, highlights etc.). Text readability tests are also considered, however, as they do not apply to all languages, they are complementary. Points for PM, PK and PT is calculated only for the first 30 URLs (subpages) found in the source code of the website. The scoring can be affected by the UE factor, which is calculated on the basis of the number of internal links pointing to non-existing pages, i.e. when the server returns HTTP 4xx or 5xx error code.

The ISOWQ Rank algorithm is shown below:

ISOWQ Rank Algorithm
Input:
array T(U) with technical analysis of subpages
array P(S) with technical analysis of the web server
Output:
value in the range of <0,20>

- Create variables PM1, PM2, PK, PT, UE
- Create an auxiliary array V
- For each S_i € P calculate:
  - analyse the data
  - put for the result in PM1
- For each U_i € T, calculate penalty points for each tuple:
  - analyse the data
  - put the result in UE
- Reset the V array
- For each U_i € T calculate the score for each tuple:
  - analyse the data
  - put the result in V_i
- Insert the average value of V array into PM2
- If UE > 0, lower the PM2 value, including UE
- PM = PM1 + PM2
- If PM < 0: PM = 0 elif PM > 20: PM = 20
- Reset the V array
- For each U_i € T calculate the score for each tuple:
  - analyse the data
  - put the result in V_i
- Insert the average value of V array into PK
- If UE > 0 lower the PK value, including UE
- If PK < 0: PK = 0 elif PK > 20: PK = 20
- Reset the V array
- For each U_i € T calculate the score for each tuple:
  - if number of words in the text > 100
    - analyse the data:
    - put the result into V_i
  - elif:
    - to V_i insert 0
- Insert the average value of V array in PT
- If UE > 0 lower the PT value, including UE
- If PT < 0: PT = 0 elif PT > 20: PT = 20
- Return (PM + PK + PT) / 3

In practice, to get more than 5 points would mean that the website has met the basis technical requirements and is present in popular rankings. For example, melex.com.pl [5] obtained 8.73 ISOWQ Rank points, which should be considered a good result. Based on the analysis, it can be seen that this website also uses the YouTube service to publish video materials. It is worth noting that its code is optimised at 87%, and the <title> attribute occurs in 79% of <a> tags (hyperlinks), which again is a good value. Unfortunately, this website does not use other IT technologies such as social media plugins or the SSL encryption. Besides, the <alt> attribute only occurs in 25% of <img> tags, and public email addresses are not encoded.

Obtaining more than 10 points means that the website uses marketing techniques to communicate with other services, has many inbound links, contains a lot of text content and has a high ranking position. An example of such web service is alivia.org.pl [6] which scored 14.21 ISOWQ Rank points, indicating a very good result. It uses social media plugins, the SSL encryption, and its system is based on the modern CMS (Content Management System) – WordPress which allows the publication of the search engine – optimised text and video materials. Marketing strategies were implemented, as evidenced in the use of YouTube service. It also offers a large number of shares on Facebook and a high position in MOZ and Alexa Rank rankings. The website contains

large unique text content on every subpage. Code optimisation is at 91% and the <alt> attribute is found in 73% of <img> tags, which again is a good result.

Additional optimisation in terms of using the <title> attribute in the <a> tags and encoding email addresses would increase the chance of getting a score of 15 ISOWQ Rank points.

The score of more than 15 points is usually reserved for websites that use dominant IT technologies in Web marketing, are optimised according to SEO guidelines [7] and have high ranking positions. In practice, the score of 15 points is exceeded very rarely and is usually obtained by large web services.

An example of such website is lazienkaplus.pl [8] service, which scored 15.39 ISOWQ Rank points.

It uses social media plugins, microformats, has high positions in MOZ and Alexa Rank, has a well optimised code (93%), the <title> attribute in the <a> tag reaches 91% and the <alt> attribute in the <img> tag reaches 93%.

3. The ISOWQ ranking system

In 2010, with the increasing popularity of the SEO (Search Engine Optimisation) industry, I designed the ISOWQ (International Studies of Website Quality) ranking system as a suggestion for a new technical analysis of websites. The ISOWQ system comprises of two independent segments. The first segment consists of the subsystem responsible for automatic typing with the analysis of subsequent website while the second is responsible for the data presentation and the system user support. The user can add additional web addresses to the system and place an order for the web analysis. Both segments have direct access to the database server. In order to create the system, I have used IBM eServer x345 and Dell PowerEdge 1950 servers. I have also written the software in PHP using the PNCTL library for concurrency service [9]. The data archiving system uses MySQL and SQLite3 database engines.

In the first few years of the ISOWQ operation, all the analysed websites were hosted on the national domains only, that is, belonging to the European Union countries, the EU candidate countries, Russia with other members of the Commonwealth of Independent States, the United States of America, as well as those under the .eu domain.

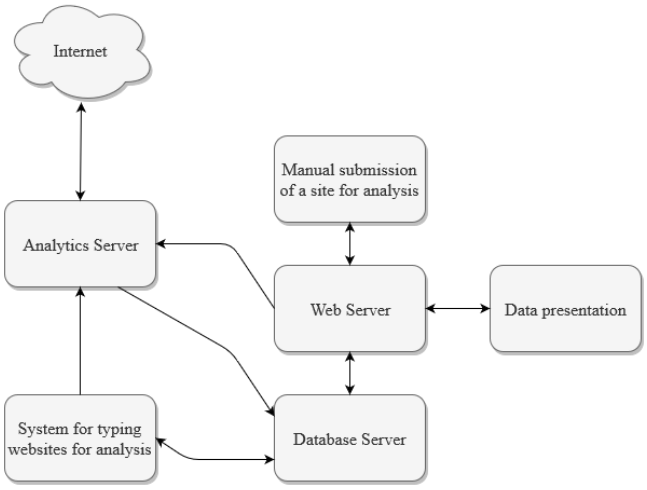

Fig. 2. The structure of ISOWQ system

Today, the main purpose of the ISOWQ system is to analyse websites operating under all (243) national domains (ccTLD). ISOWQ also produces reports to make it available publicly on the Internet. The data provided by the ISOWQ can also be used by other IT systems or be subjected to further analysis. The database contains the detailed analyses of over 1.3 million websites which creates over 25 million URLs [10].

The structure if the ISOWQ system is shown in Figure 2.

In addition to presenting the technical information online, the ISOWQ system enables the analysed website to be presented in a PDF format, with advice to specific questions regarding technical, marketing and text related aspects. Based on the collected data, one can generate many interesting combinations, e.g. regarding the use of multimedia on websites, the use of social media plugins or versions of the HTML markup language in a specific group of ccTLD domains.

Table 1 shows the summary for European national domains and the dependent territories (2016).

Table 1. The application of HTML5, YouTube and Social Media on European national domains based on the analyses conducted in 2016

Place	ccTLD	HTML5	YouTube	SM	No of audits
1	gg (Guernsey)	83,60%	19,22%	50,70%	98
2	no (Norway)	78,19%	19,29%	38,19%	1375
3	ba (Bosnia and Herzegovina)	77,49%	33,57%	8,20%	170
4	mc (Monaco)	77,49%	33,57%	8,20%	69
5	al (Albania)	65,29%	25,37%	26,71%	153
6	ax (Åland)	77,20%	15,78%	18,69%	135
7	je (Jersey)	72,56%	11,35%	25,15%	46
8	gi (Gibraltar)	62,99%	20,27%	24,68%	135
9	ad (Andorra)	56,09%	24,40%	22,92%	98
10	ch (Switzerland)	57,70%	19,10%	19,82%	1373
20	fr (France)	51,32%	9,16%	14,80%	2404
25	va (Vatican City)	62,62%	5,00%	2,50%	35
33	pl (Poland)	39,70%	8,58%	16,73%	2391
34	lv (Latvia)	39,62%	11,06%	14,04%	1456
35	eu (European Union)	40,26%	11,42%	12,05%	2235
40	gr (Greece)	35,48%	10,10%	16,24%	2302
45	ru (Russia)	39,38%	10,38%	9,28%	1546
48	de (Germany)	36,69%	10,17%	9,82%	2812
51	tr (Turkey)	39,47%	6,20%	6,58%	1890
52	mt (Malta)	36,45%	5,98%	6,90%	246

The Table above shows that in the websites with the .gg domain (analysed in 2016), HTML5 markup language was used in almost 84% of them, while those with the .pl domain used it only in 39,70%. The next interesting list in Table 2 is the share of unencrypted email addresses in the page code and registration of the IP addresses of the hosting servers as spam carriers in DNSbl databases.

Table 2. Use of IP addresses of hosting servers in DNSbl databases and the share of unencrypted email addresses in the code of websites analysed in 2016

Place	ccTLD	DNSbl	E-mail	No of audits
1	tn (Tunisia)	81,28%	57,56%	144
2	et (Ethiopia)	75,86%	52,01%	91
3	ne (Niger)	70,16%	39,05%	30
4	bt (Bhutan)	63,16%	66,39%	137
5	sd (Sudan)	59,95%	45,18%	136
6	bf (Burkina Faso)	59,71%	42,47%	75
7	sn (Senegal)	53,69%	53,11%	140
8	mm (Myanmar)	53,26%	48,83%	101
10	mv (Maldives)	49,92%	61,67%	109
28	mk (Macedonia)	31,13%	43,05%	467
29	kz (Kazakhstan)	30,88%	41,49%	557
122	pl (Poland)	9,89%	55,79%	2391
131	fr (France)	8,96%	39,70%	2404
169	lv (Latvia)	6,87%	60,63%	1456
173	de (Germany)	6,75%	60,53%	2812

The above statement shows that servers classified as a spam carrier, operate mainly in African countries. It also shows that the problem of unencrypted e-mail addresses on websites has a global dimension.

4. Comparative study of ranking algorithms

The t-Kendall ratio was used to estimate the correlation between the results obtained by MOZ and ISOWQ Rank algorithms. 10 websites were selected for the analysis and included the following technologies:

- HTML5 (+31 additional types, including HTML 2, 3, 4 and XHTML 1 and 2),
  - code analysis – meta, http-equiv, HTML tags, diagrams, microformats,
- Social media (Facebook, Twitter, Google, LinkedIn),
- Google services i.e. Ads (Adwords), Adsens, Analytics, Maps,
- JS framework (jQuery, Bootstrap +24 additional),

- Multimedia (YouTube, Flash, Silverlight),
- SSL encryption, spam protection (e-mail and DNSbl).

The study was conducted in April 2019. The scale of results has been unified to a range of 0 to 100. The results are presented in Table 3.

Table 3. Websites analysed for the purpose of research in April 2016

Website	ISOWQ Rank (0–100)	MOZ DA (0–100)
4wsk.pl [11]	55	31
alivia.org.pl [7]	71	45
argos.org.pl [12]	48	31
brightmedia.pl [13]	15	34
link2europe.pl [14]	40	30
machineryzone.pl [15]	52	30
melex.com.pl [6]	44	31
mlyny-rozdrabniacze.pl [16]	28	7
palacporaj.pl [17]	29	22
wrobywatel.pl [18]	19	14

Microsoft Excel 2016 software was used together with the "The Real Statistics Resource Pack" [19] to calculate this correlation. At the declared significance level of 0.05, the t-Kendall ratio is 0.44, which means that the algorithms have a moderate relationship with each other.

Figure 3 shows the correlation between ISOWQ Rank and MOZ DA.

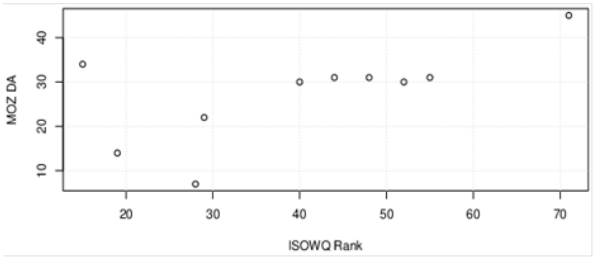

Fig. 3. Correlation between the ISOWQ Rank and MOZ DA ranking

By comparing the analyses of the ISOWQ Rank scores, it can be seen that the main differences relate to the use of social media technologies, multimedia, SSL encryption and content management system. The comparative summary of these analyses is presented in Table 4.

Table 4. A comparison of the results of different analyses for extreme scores

	brightmedia.pl	alivia.org.pl
ISOWQ Rank	3,05	14,21
MOZ DA	34	45
Alexa Rank (the higher the score the better)	1 095 670	496 768
Social media plugins usage	No	Yes
YouTube video clips usage	No	Yes
SSL (HTTPS)	No	Yes
CMS – WordPress usage	No	Yes
Source code optimisation	81%	91%
Usage of <title> attribute in the <a> tag	91%	0%

The study showed that there is a 'positive' correlation between ISOWQ Rank and MOZ, i.e. an increase in one should result in an increase in the other. This means that the technical aspect of the website (technologies used, code optimisation, text structure) is an important factor in the MOZ algorithm, whose detailed specification is not publicly available.

The 'unconventional' method of placing the text content in the code of brightmedia.pl unable the correct detection and analysis, which is reflected in the number of points for the text (PT).

Here is a snippet of source code for this site:

```
<p class="textF2" id="text1_1">Nie jesteśmy nowicjuszami,</p>
<p class="textF3" id="text1_2">nie wzięliśmy się znikąd. Mamy wiedzę</p>
<p class="textF4" id="text1_3">i doświadczenie, bo pracowaliśmy</p>
<p class="textF5" id="text1_4">w znanych agencjach dla znanych marek.</p>
```

The ISOWQ Rank algorithm did not detect any text zones in the source code above, as it considered that there was sufficient number of characters in a single paragraph in the in the <p> tag.

If the above code snippet was replaced with:

```
<p>
Nie jesteśmy nowicjuszami,<br>
nie wzięliśmy się znikąd. Mamy wiedzę<br>
i doświadczenie, bo pracowaliśmy<br>
w znanych agencjach dla znanych marek.
</p>
```

The algorithm would then detect the text zone and the website would receive a higher score for the structure, which would have a positive impact on the final ISOWQ Rank. However, based on the source code, it can be assumed that this change would negatively affect the form of presentation of the text on the page.

5. Conclusion

This paper presents the main principles of the ISOWQ system and ISOWQ Rank algorithm. The findings obtained during these analyses confirm the thesis that the use of modern IT technologies and the appropriate structure of the text are as important as the number of hyperlinks to the website. The correct detection and analysis of the text zones understood by humans is a very complicated task, especially where the text plays a secondary role. An interesting aspect of the text analysis could be the tests checking its readability level. Although most of the tests were created with the English language in mind, there are also some tailored variants to accommodate to the Polish language intricacies. The ISOWQ Rank value is moderately convergent with the MOZ ranking. Therefore, despite the lack of access to the algorithm specification (MOZ), it can be presumed that also here, as in the case of ISOWQ Rank, the technologies used and the scope of content and code optimisation are taken into account.

References

[1] Brin S., Page L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Stanford University, Stanford 1998.
[2] Bailyn B., Bailyn E.: Outsmarting Google: SEO Secrets to Winning New Business. Que Publishing, Indianapolis 2011.
[3] https://moz.com/learn/seo/domain-authority (available: 20.06.2019).
[4] https://www.alexa.com/about (available: 20.06.2019).
[5] https://www.isowq.org/website/melex.com.pl/1439980/ (available: 02.04.2019).
[6] https://www.isowq.org/website/alivia.org.pl/1441262/ (available: 26.04.2019).
[7] https://support.google.com/webmasters/answer/35769 (available: 20.06.2019).
[8] https://www.isowq.org/website/lazienkaplus.pl/1322181/ (available: 12.04.2019).
[9] https://www.php.net/manual/en/book.pcntl.php (available: 06.12.2019).
[10] https://www.isowq.org/about (available: 06.12.2019).
[11] https://www.isowq.org/website/4wsk.pl/1443221/ (available: 08.04.2019).
[12] https://www.isowq.org/website/argos.org.pl/1441579/ (available: 05.04.2019).
[13] https://www.isowq.org/website/brightmedia.pl/1439337/ (available: 01.04.2019).
[14] https://www.isowq.org/website/link2europe.pl/1439533/ (available: 01.04.2019).
[15] https://www.isowq.org/website/machineryzone.pl/1440087/ (available: 02.04.2019).
[16] https://www.isowq.org/website/mlyny-rozdrabniacze.pl/1439111/ (available: 01.04.2019).
[17] https://www.isowq.org/website/palacporaj.pl/1439800/ (available: 02.04.2019).
[18] https://www.isowq.org/website/wrobywatel.pl/1444142/ (available: 10.04.2019).
[19] http://www.real-statistics.com/free-download/real-statistics-resource-pack (available: 10.05.2019).

M.Sc. Eng. Mariusz Duka
e-mail: mariduk781@student.polsl.pl

Ph.D. student at the Faculty of Automatic Control, Electronic and Computer Science of the Silesian University of Technology. Professionally associated with the IT industry since 1993, in particular in software design and its implementation. Author of many Web projects, including the ISOWQ (International Studies of Website Quality) ranking system. His scientific interests related to the issues of data exploring, concurrent programming and IoT.

http://orcid.org/0000-0001-5773-0459

otrzymano/received: 16.12.2019 *przyjęto do druku/accepted: 26.06.2020*