Practice of Robots Exclusion Protocol in Bhutan

Kezang Dema[1*] Thinley Jamtsho[2]
1.Information Technology Department, College of Science and Technology, Royal University of Bhutan, Chhukha, Bhutan
2.Department of Science and Mathematics, Phuentsholing Middle Secondary School, Phuentsholing, Chhukha, Bhutan
* E-mail of the corresponding author: kelden.dema@gmail.com

Abstract
Most of the search engines rely on the web robots to collect information from the web. The web is open access and unregulated which makes it easier for the robots to crawl and index all the contents of websites so easily. But not all wish to get their websites and web pages indexed by web crawlers. The diverse crawling activities can be regulated and managed by deploying the Robots Exclusion Protocol (REP) in a file called robots.txt in the server. The method used is a de-facto standard and most of the ethical robots will follow the rules specified in the file. In Bhutan, there are many websites and in order to regulate those bots, the usage of the robots.txt file in the websites are not known since no study has been carried out till date. The main aim of the paper is to investigate the use of robots.txt files in various organizations' websites in Bhutan. And further, to analyze its content present in the file if it exist. A total of 50 websites from various sectors like colleges, government ministries, autonomous agencies, corporations and newspaper agencies were selected for the investigation to check the usage of the file. Moreover, the files were further studied and analyzed for its file size, types of robots specified, and correct use of the file. The result showed that that almost 70% of the websites investigated are using the default robots.txt file generally created by the Joomla and Word press Content Management Systems (CMS) which ultimately specifies that there is a usage of the file. But on the other hand, the file is not really taken into seriously and almost 70% of it lacks major and best protocols defined in it that will help define the access and denial to various resources to various types of robots available on the web. Approximately 30% of the URLs adopted for the study show that the REP file is not added in their web server, thus providing unregulated access of resources to all types of web robots.

Keywords: Crawler, robots.txt, search engines, robots exclusion protocol, indexing
DOI: 10.7176/JEP/11-35-01
Publication date: December 31st 2020

1. Introduction

There are various search engines namely Google, Bing, Yahoo! search, Baidu, Ask and many more, among which Google, Yahoo! And Bing being one of the top most search engines used by many users. All those search engines have their own web robots, also known as web crawlers and one of the most leading search engines is Google that depends on Googlebot. The search engines help the users with content delivery from any websites using their crawling agent called spiders or crawlers (Kolay et al., 2008). These crawlers are highly automated and because of their nature it captures every new page and sites added in the web without any obligations imposed on them. It would index about 100 Gigabytes of data consisting thousands of keywords. This is the concept of site being spidered and crawled. The captured pages are stored in a database and they are indexed, relatively making the information available to users' when queried using any search engines (Jha et al., 2014). Web search engines, digital libraries, offline browsers, and intelligent searching agents heavily depend on those robots to retrieve the content (Sun, Zhuang & Giles, 2007). There won't be any functional search engines without the presence of those web robots. By default, every public and private site is indexed by web crawlers since they are not restricted to do so. When many web crawlers are made to crawl the sites without any restriction, this may have an undesired impact on the server workload and provision of access to non-public information. There needs to be rules to regulate the crawling behavior and indexing activities of robots at individual web servers (Yang & Liao, 2010)

As per Koster (1996), the Robots Exclusion Protocol has been proposed as one of the methods that would provide advisory regulations for robots to follow. In order to do so, a file called robots.txt that will define all the advisory regulations for the robots to follow and it will be deployed at the root directory of a website (Jha et al., 2014). The file will be accessible to all possible web robots and most of the ethical robots will read the robots.txt file and obey the rules when visiting the website for indexing. Although the use of REP is not the official standard, they are used widely by nearly all commercial search engines and popular web crawlers especially to guide and regulate the robots activities. Thus, this helps in evading indexing by disallowing or allowing some of all web robots an access to certain parts or all of the content of the entire site. Despite its importance and criticality of the REP file, no work has been done to investigate its usage especially in Bhutan. Some works have been carried out in other parts of the world especially by Sun, Zhuang & Giles (2007) and Kolay (2008). In Bhutan, there are 10000 and more websites in various domains like colleges, schools, government, private companies, news and many more, but till now, none has undertaken a study and investigated the use of the robots.txt file in large scale. This paper will present the first ever investigation and analysis on the usage of robots.txt file in Bhutan that will cover the domains of 13 colleges, 10 ministries, 6 corporations, 19 autonomous agencies and 2 news agencies.

2. Literature Review

The software robots are one of the most widely used tools for data searching services. Yang and Liao (2010) made amendments to the existing robots.txt and robots Meta tags that would help in allowing and refusing the software robots, hence helping in evading Google indexing. It helped robots in preventing conflicts by making robots follow given policies and some in breaching the law if ignoring the policy, thus allowing the webmaster in expressing their will explicitly and avoiding online copyright authorization policies. The easiest method to optimize the search engines for any sites is the use of Robot Exclusion Protocol that is easily deployed in the server because many search engines rely on the available web robots to crawl and index the web pages and make the information easily available from the web. Sun, Zhuang & Giles (2007) carried out a large scale study on the use of the REP in the file called robots.txt. The study collected and investigated a total of 7593 websites covering various domains like government, business, and news and education sites. The analysis of the data showed the use of the robots.txt file in their site to regulate the crawling activities by the robots. Moreover, the survey showed the increased usage of robots.txt by various domains over the time. The study carried out by Sun, Zhuang & Giles (2007) shows that many sectors wish to have the REP rules be used to regulate the robots.

The robots Exclusion Protocol was being deployed on the server side to regulate the crawling of the sites by the robots but found some bias to search engines from robtos.txt. Sun et al. (2007) surveyed the usage of robots.txt file that regulate crawling activities and found some bias among the robots and its crawling activities. A total of 7925 websites were investigated and collected 2925 distinct robots.txt files for the analysis. The result showed that there is a serious bias towards some of the popular robots especially the robots of popular search engines and information portals like Google, Yahoo, and MSN. They are mostly preferred by the websites that they have sampled and analyzed. In contrast to what Sun et al. (2007) has shown in their analysis and result, Kolay et al. (2008) conducted a study to investigate the bias towards some search engines as found and concluded by Sun et al. (2007) in their work. In his work, he surveyed about six million sites. The team conducted a comprehensive and large scale study of the robots.txt file usage in those six million sites. The result of his work showed a very insignificant bias towards search engines that was contrary to the result proved by Sun et al. (2007).

Many search engines largely depends on web crawlers to crawl and collect the information and this hence, led to the web traffic generated by various crawlers. In order to regulate the behaviors of web crawlers at a web server, many adopts the usage of robots.txt file. But the question in here is the extent to which the robots respect and violets the regulations put in the robots.txt file. In this regards, Giles, Sun & Councill (2010) conducted a study on the web crawlers and their ethics based on their behaviors by proposing a vector space model. The result obtained was that most of the commercial web crawler's behavior are ethical and are guided as per the rules specified in the file. However, there are few crawlers that violate certain rules. The use of this technique to regulate the crawlers can be efficient since they are found to be obeying the rules. The search engines rely on software robots to crawl web pages and to create indices for users to search. Thom (1999) used the robots.txt file and the robots Meta tag to provide guidance to robots on whether and how to catalogue a site they have contacted. The approach helps in creating robots.txt and robots Meta tag files that would enable webmasters to reduce the load placed on web server by legitimate robots

As per Ge and Ding (2016), the general-purpose web crawlers are not able to crawl and fetch the deep web pages and Ajax pages available in the websites with its current Robots Protocol designed and developed by many search engine companies. In order to help robots to crawl any kind of pagers, the authors have proposed a Robots Exclusion and Guidance Protocol (REGP) by integrating their proposal along with the current available robots protocols developed. Various series of experiments were conducted to verify the efficiency of their proposed REGP. The study and the experiments carried out by the authors seems to have provided an additional input so that the capabilities of the robots are enhanced for them to crawl those restricted pages.

As many studies show that the use of REP in the web servers not necessarily causes all the web crawlers to follows the rules specified in the robots.txt file. The robots are found to be disobeying the rules defined in the file thereby causing a bigger threats to the webserver being overloaded and harming the network security with distributed denial of service (DDoS) attacks (Li, Liao & Zeng, 2019). The research displayed the efficiency of the rules specified in the file and the factors leading to the disobey situation using Nash equilibrium of the game theory model.

3. Methodology

The online desk research technique was adopted for the research whereby a total of 70% of the website URLs were accessed from the National Portal of Bhutan (NPB) and remaining were accessed by visiting the selected sectors' own websites. The methodology followed for the study is depicted in Figure 1.

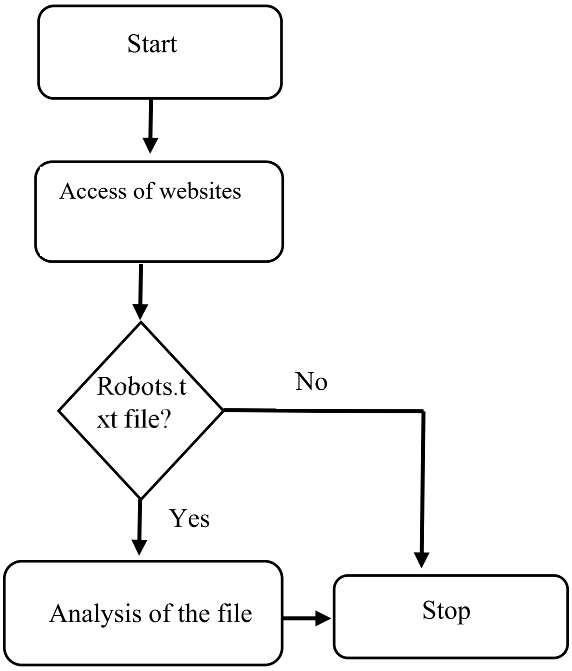

Figure 1. Methodology

3.1 Collection of websites of various domains

The main source from which the URLs were collected was from the National Portal of Bhutan (NPB). The NPB portal provided access especially to three domains namely ministries, dzongkhags and business domains among which the ministries domain was selected for the research study. The other domains were collected through Internet search. A total of 50 website URLs consisting of different domains like 10 ministries, 13 colleges, 19 autonomous agencies, 2 media agencies and 6 corporations were identified for the evaluation as depicted in Table 1 and Figure 1.

Table 1. Different Sectors

Sl. No	Sector	Total URLs/websites
1.	Colleges	13
2.	Ministries	10
3.	Corporations	6
4.	Autonomous Agencies	19
5.	News Agencies	2

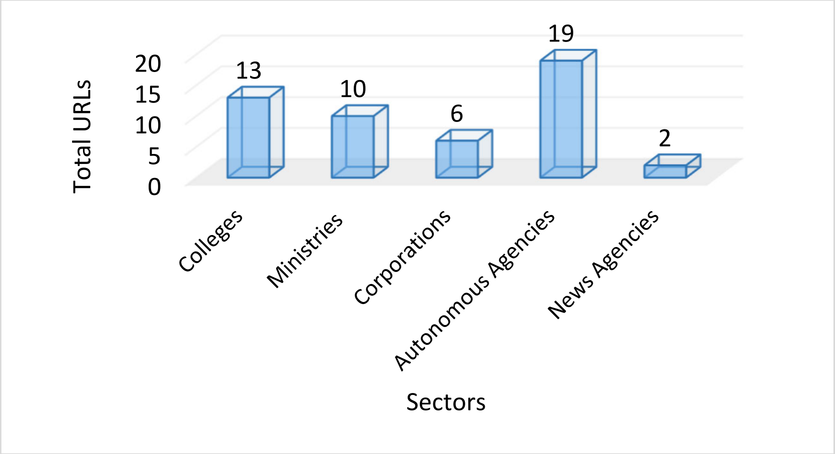

Figure 2. Different Sectors

3.2 Check for the presence of robots.txt file

All the 50 selected sectors' websites were examined to check the existence of the robots.txt file (REP). The details of the findings are shown in Figure 3.

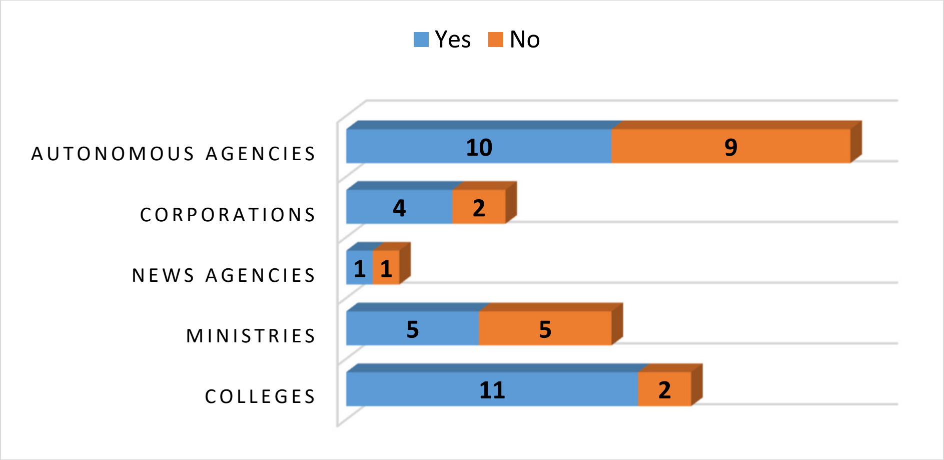

Figure 3. Presence of REP robots.txt file in websites

The analysis on the presence of the REP file in those 50 websites of various sectors show that, at least there is the existence of the file in majority of the organizations' webserver even though the usage is not taken into account strictly. As depicted in Figure 3, out of 19 autonomous agencies, 10 of them has made use of the file while the remaining 9 don't have the file. Out of the 6 corporations, 4 have made use of the file and the remaining 2 do not have the file when checked for it. Among the 2 selected new agencies, one have not included the file whereas the other has specified the file. When it comes to the 10 selected ministries, 5 of them have added the file while the other 5 have not included the file at all in their webservers. In the case of educational institutions, 11 colleges have added the REP file to monitor the web crawling of their web pages whereas 2 colleges do not have the file.

3.3 Analysis of the files

All the REP robots.txt files available or being used by those sectors were then extracted for further analysis particularly on the given areas: (i). Length of the file, (ii). Types of robots used and (iii) Appropriate use of the rules in the file.

4. Results and Discussions

The overall findings after studying and analyzing all those available robots.txt file of different organizations are depicted in Table 2.

Table 2. Statistics of files from various selected organizations

Sl.No	Organization	File presence	Robots Type	File Length	Sector/Domain
1.	College of Science and Technology (CST)	Yes	General	15 lines	Colleges
2	Gedu College of Business Studies (GCBS)	Yes	General	14 lines	
3	Paro College of Education (PCE)	Yes	General	47 lines	
4	Samtse College of Education (SCE)	Yes	General	3 lines	
5	Sherubtse College (SC)	Yes	General	3 lines	
6	CLCS	No	NA	NA	
7	Gyelposhing College of Information Technology (GCIT)	Yes	General	3 Lines	
8	Jigme Namgyel Engineering College (JNEC)	Yes	General	3 Lines	
9	College of Natural Resources (CNR)	Yes	General	33 Lines	
10	Royal Thimphu College (RTC)	Yes	General	14 lines	
11	Norbu Rigter College (NRC)	Yes	General	16 Lines	
12	Younphula Centenary College (YCC)	Yes	General	3 lines	
13	Kgumsb	No	NA	NA	
14	Ministry of Agriculture and Forests (MoA)	Yes	General	4 lines	Ministries
15	Ministry of Economic Affairs (MoEA)	No	NA	NA	
16	Ministry of Education (MoE)	No	NA	NA	
17	Ministry of Finance (MoF)	Yes	General	3 Lines	
18	Ministry of Foreign Affairs (MoFA)	No	NA	NA	
19	Ministry of Health (MoH)	Yes	General	3 lines	

Sl.No	Organization	File presence	Robots Type	File Length	Sector/Domain
20	Ministry of Home and Cultural Affairs (MoHCA)	No	NA	NA	
21	Ministry of Work and Human Settlement (MoWHS)	Yes	General	1 Line	
22	Ministry of Labour Human Resources (MoLHR)	No	NA	NA	
23	Ministry of Information and Communications (MoIC)	Yes	General	3 Lines	
24	Bhutan Broadcasting Service (BBS)	No	NA	NA	News Agencies
25	Kuensel	Yes	General	3 Lines	
26	Bhutan Telecom (BT)	Yes	General	3 lines	Corporations
27	Bhutan Power Corporation (BPC)	Yes	General	3	
28	Druk Green Power Corporations (DGPC)	Yes	General	3 Lines	
29	Druk Air	No		NA	
30	Royal Insurance Corporation of Bhutan Limited (RICBL)	No		NA	
31	Bhutan Post (BP)	Yes	General	2 Lines	
32	Bhutan Narcotics and Control Agency (BNCA)	Yes	General	3 Lines	Autonomous Agencies
33	Bhutan Infocomm and Media Authority (BIMA)	No	NA	NA	
34	Bhutan Council for School Examination and Assessment (BCSEA)	No	NA	NA	
35	Bhutan Olympic Committee (BOC)	Yes	General	3 Lines	
36	Bhutan Standards Bureau (BSB)	No	NA	NA	
37	Dzongkha Development Commission (DDC)	Yes	General	12 Lines	
38	Civil Society Organizations Authority (CSOA)	No	NA	NA	
39	Drug Regulatory Authority (DRA)	No	NA	NA	
40	Bhutan Studies (BS)	Yes	General	35 lines	
41	Construction Development Board (CDB)	Yes	General	2 Lines	
42	GNH Commission (GNHC)	No	NA	NA	
43	NCWC	Yes	General	2 Lines	
44	National Environment Commission (NEC)	Yes	General	3 Lines	
45	National Land Commission Secretariat	No	NA	NA	
46	National Statistics Bureau	No	NA	NA	
47	Office of Attorney General	Yes	General	4 Lines	
48	Royal Education Council	Yes	General	3 Lines	
49	Royal Institute of Management	No	NA	NA	
50	Tourism Council of Bhutan	Yes	General	2 Lines	

4.1 Length of the files

The addition of the strict REP with the use of robots.txt file in the websites for the robots crawling and for effective Search Engine Optimization (SEO) is very minimal in our country Bhutan. It seems that many sectors opt to use other various ways and methods for SEO. Accordingly, even if the file is being used, the number of lines that indicates various rules applied to various web robots are very minimal and it's not a very strict rules. The length of the files range from 1 line being the least number of rules to a maximum of 35 lines in some cases. The distribution of the number of lines per file in those selected websites of different sectors taken for this particular study is given in Figure 4.

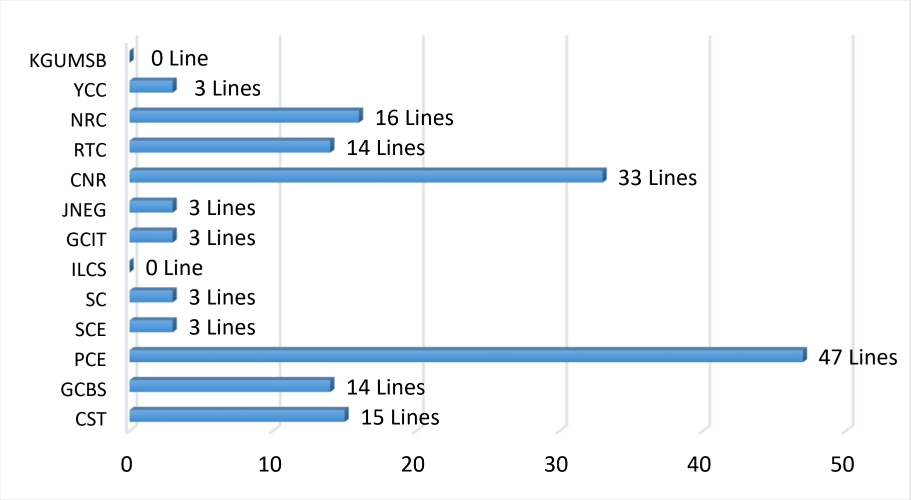

Figure 4. Robots.txt file length in colleges

The college sector consist of 13 selected colleges available in our country. The analysis shows that at least 2 colleges have not added the robots.txt file. The other 11 colleges have added the file as shown in Figure 4, and the length of the files range from 3 lines of rules to a maximum of 47 lines of rules specified in PCE.

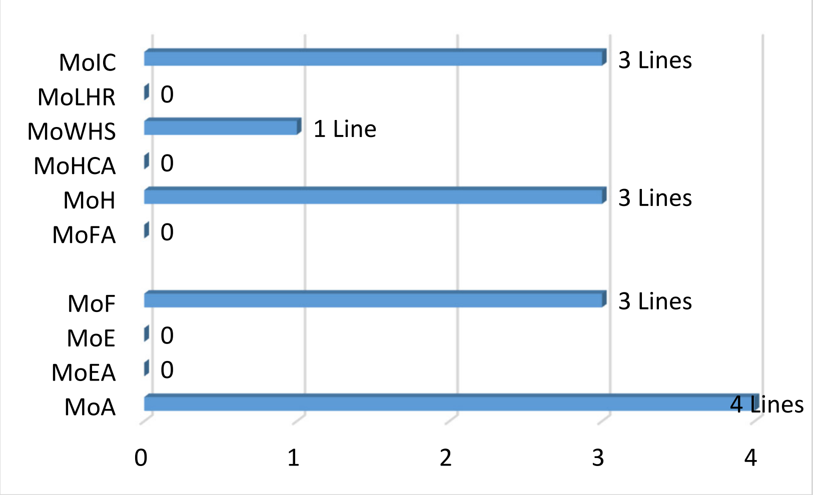

Figure 5. Robots.txt File Length in Ministries

The length of the file of their robots.txt for those 10 selected ministries are too minimal wherein the maximum rules written is 4 lines in MoA as shown in Figure 5. The other three ministries namely MoF, MoH and MoIC has only three rules applied. The MoWHS has one line of rule added in their file. The other remaining 5 ministries are not using the file. The consideration of using the file to monitor the crawling of web pages by the robots are not well monitored in this sector.

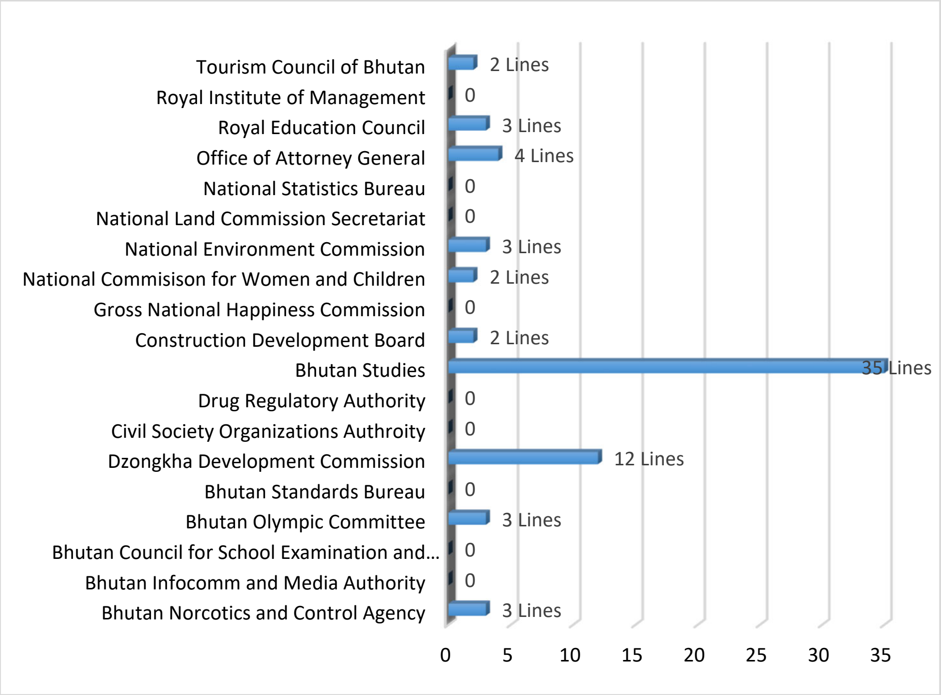

Figure 6. Robots.txt File Length in Autonomous Agencies

A total of 19 autonomous agencies considered for the study too showed a very minimal use of the REP file in their webserver to monitor the web crawling by various web robots as shown in Figure 6. The REP file could not be accessed in 10 agencies which clearly shows that they don't use the file. The remaining agencies have specified a very minimal rules in their file ranging from 2 lines, 3 lines and 4 lines. The DDC has more number of lines consisting of 12 rules and the maximum rules specified was in Bhutan studies agencies which has 35 lines in their robots.txt file.

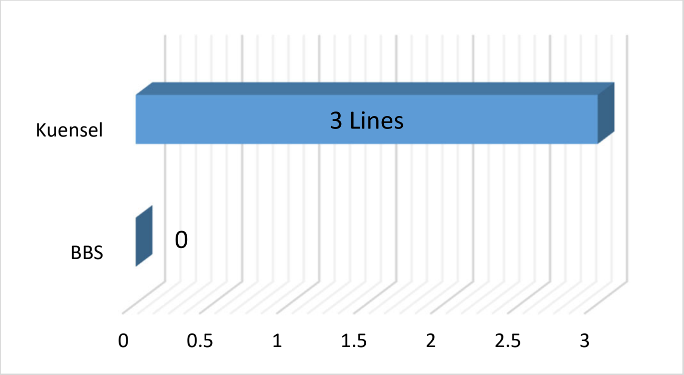

Figure 7. Robots.txt File Length in Media Agencies

In the context of news media, only two organizations were selected namely Kuensel and BBS. Between those two, BBS has not included the robots.txt file. The other organization Kuensel did have the file but the number of rules specified is only 3 lines as shown in Figure 7. The file consisting of only 3 lines won't be having much of better rules specified which can better monitor the web robots from crawling the authentic and appropriate web pages.

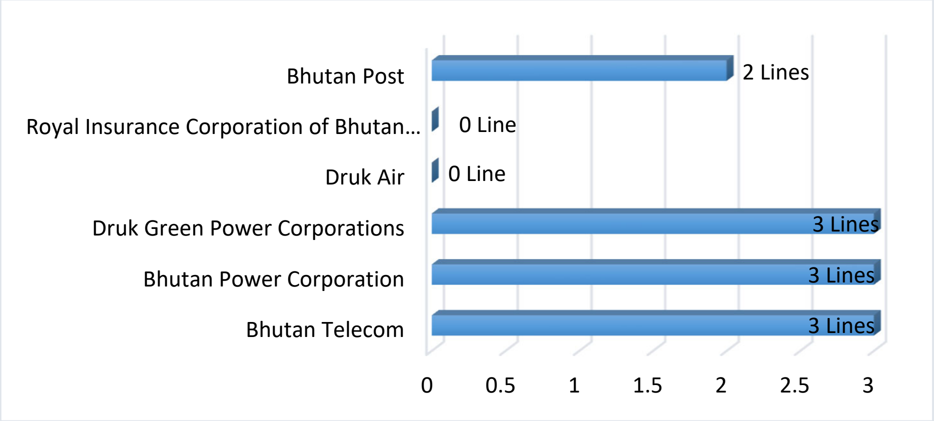

Figure 8. Robots.txt File Length in Corporations

Same case can be seen in those 6 selected corporation sectors. The length of the files consisting of the rules specified is very less that ranges from 2 lines to the maximum 3 lines as shown in Figure 8. As seen in many other sectors, this sector too has developed their websites using the CMS like Joomla and Wordpress which automatically adds the default REP file with very minimal rules or 3-4 lines.

4.2 Types of robots used

In most of the cases, the type of robots defined in the file for access and denial of contents from a particular website will demand the need to mention the robots name explicitly. In this study, almost 95% of the websites taken for the study do not have explicit mention of the popular robots like GoogleBot, Bingbot, Slurp Bot, Baiduspider, DuckDuckBot and many more. Many default REP file added automatically by the Joomla and wordpress CMS have made use of the asterisk (*) symbol in the place where the robots name have to be mentioned and this asterisk symbol will refer to all general robots as given below:

User-agent: *

The protocol line for user-agent in almost all the files are given with the value of asterisk symbol. With the use of asterisk symbols for the user-agent, this will imply that the rules written in the robots.txt file are applied to any general robots. The kinds of robots specified in various REP files are given in Table 2 that shows the complete statistics of the findings.

4.3 Incorrect use

As per the selected domains of study, all the files were added correctly as per its standard method. None of the files were added incorrectly, thus when the files were searched for by writing their domain name followed by the robots.txt file name, the output was displayed by showing all the detailed rules. The most interesting thing is that some files were added as a default file especially for those websites built around Joomla and wordpress CMS.

5. Conclusion

There are lots of search engines and almost all of them depend on their own search engine robots to crawl the pages of a various websites and provide the information to the users. Many studies carried out in other parts of the world by various authors showed that they make better use of the Robots exclusion protocols by specifying in the file named robots.txt. With its use, they make sure that they regulate the access and denial of their web pages to be crawled by different types of web robots by specifying the protocols in the REP file. In some cases, some studies showed that they have worked into adding more rules in the REP file so that the accessibility and capability of the robots are enhanced to crawl some of the pages that can't be accessed if followed the protocols of normal REP file. In the context of Bhutan, from the 50 sample websites that were used for the study showed that the usage of robots.txt file in order to regulate the robots in providing access and denial to certain webpages of their sites are not really common and are not considered in greater depth.

And also, 38% of them do not have the file added in their web server which ultimately shows that almost all the robots are allowed to access any kind of web pages and contents without much restriction. On the other hand, for 62% of the websites that have used the REP file showed that they do have the robot.txt file added but it lacks proper and well-defined rules whereby some have only few lines of rules which do not really regulate the robots efficiently and appropriately, thus need to re-look into it and define the rules properly if one want to restrict some of the vital contents from being crawled by the robots and make accessible in the Internet. The future study may include the study of additional URLs of various other domains not included in the current study and do the analysis of the REP files. And also, the current URLs taken for the study can be re-evaluated again and see whether the rule are added appropriately or not and also see the obedience of the rules by various web crawlers.

References

Ge, D., & Ding, Z. (2016). Robots Exclusion and Guidance Protocol. *Tsinghua Science and Technology*. 21, 643-659. doi: 10.1109/TST.2016.7787007

Giles, C.L., Sun, Y., & Councill, I.G. (2010). Measuring the Web Crawler Ethics. *Proceedings of the 19th international conference on World wide web 2010 ACM*. 1101-1102

Jha, P. et al. (2014). Robots exclusion protocol*. International Journal of Emerging Science and Engineering (IJESE)*. 2

Kolay, S. et al. (2008). A Larger Scale Study of Robots.txt. *Proceedings of the 17th International Conference on World Wide Web,ACM Digital Library*. 1171-1172

Koster, M. (1996). A Method for Web Robots Control. *The Internet Draft The Internet Engineering Task Force (IETF)*

Li, W., Liao, J., & Zeng,J. (2019). Efficiency Analysis on Robots Exclusion Protocol Based on Game Theory. *IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID).*1-5. doi: 10.1109/ICASID.2019.8925189.

Sun, Y. et al. (2007). Determining Bias to Search Engines from Robots.txt. *In Proc. Of International Conference on Web Intelligence IEEE*. 149-155

Sun, Y., Zhuang, Z., & Giles, C.L. (2007). A Large-scale Study of Robots.txt. *Proceedings of the 16th International Conference on World Wide Web,ACMDigitalLibrary*.1123-1124

Sun, Y., Councill, I.G., & Giles, C.L. (2008). BotSeer: An Automated Information System for Analyzing Web Robots. *Eighth International Conference on Web Engineering,IEEE Xplore.* 108-114.

Yang, C., & Liao, H. (2010). Using the Robots.txt and Robots Meta Tags to Implement Online Copyright and a Related Amendment. *Library Hi Tech*. 28, 94-106