International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-05, Issue-06, Sep 2019

Benchmarking of Web scraping data on various types of Industries supporting Transport Sector

[1]Shobha Rani B R, [2]Deepa Bharadwaj

[1,2]Dr. Ambedkar Institutre of Technology, Bengaluru, India.

[1]shobhakrishna8@gmail.com, [2]deepabharadwaj.blr@gmail.com

Abstract : This paper aims to provide a detailed description of the web scraping process, comparison between different web scraping tools and the analysis of industries data. Using different web scraping tools such as Octoparse, ParseHub, Fminer and Mozenda the industries data is collected from two different websites- Justdial and sulekha.com. The data collected includes some of the parameters such as Name of Industries, Phone Numbers, Locality, Category of industries i.e. Manufacturers, Dealers etc., Sub-Category of industries such as Cashew manufacturers, Chemical Manufacturers, Steel Manufacturers etc. and the Websites of the industry. After the collection of these data from each of the above-mentioned web scraping tools, the process of data pre-processing and cleaning is carried out. The next step is analysing the industries data collected from each tool and the pictorial reports are designed based on comparison between these web scraping tools. The Analysis of industries data is done based on the number industries in particular districts who are using fleet operators, the category of industries who are using fleet operators etc., these processes of analysis is done in order to forecast the business in Karnataka and can be extended across India.

Keywords* —Data extraction, *Web Scrapping, Web crawler, Parse Hub, Octoparse, Data Preprocessing, Data visualization

I. INTRODUCTION

The World-Wide Web consists of an interlaced network of information, which is presented through webpages to the end-users. World-Wide Web has substantially changed the way we share, gather, and broadcast data. The quantity of presented info grows constantly. As the data grows in terms of quantity and quality the managers must focus on the information which is concerned the most. For businesses, it's not the complete data which is equally important.

The web is now a days also used for improving a new marketing and sales channel thus the quantity of content is multiplied. Online merchants offer large data packs to describe their products. Knowledge base providers offer access to their databases.

According to IDC Forecasts, by 2025 the global data-sphere will be increased to 163 zettabytes (i.e. Trillion Gigabytes). That's 10 times of 16.1 ZB Data is generated in 2016. All this information will unlock unique end-user experiences and a new world of business opportunities. ((IDC), 2017)

With this unorganized growth, it is no longer possible to manually track and record all available sources. At that moment, the Web Scraping was evolved. Automated methods allow to gather a enormous quantity of information from the Website compared to manual extraction of data.

Web scraping (alias web harvesting, screen scraping or web data-extraction) is a web technique to fetch data from any web page on the Internet and turning the unstructured format of data from any websites into a structured format and store it in a regional computer.

Web Scraping programme access the World-Wide-Web directly using the Hypertext Transfer Protocol (HTTP), or with a website. There are several ways to scrape a website to extract information for re-use.

We are going to address this question by looking into the different industries. To do this, we've compiled and analysed the data extracted from Industries websites, including Justdial, Sulekha.com (Including Manual Data collection).

Figure 1.1: Web Scraping

 DOI : 10.35291/2454-9150.2019.0435

The Industries data is gathered by some of the web scraping tools as follows:

1.Octoparse: This is an advanced visual website data extraction tool. As it can be used by both experienced and inexperienced individuals, the users will find it easy to work with Octoparse as the bulk information can be extracted from different websites, and most of the tasks doesn't require coding. Octoparse operates well for both static, Dynamic webpages and also for the websites which are using Ajax. There are different formats of your choice like excel, HTML, csv, txt and database (oracle, MySQL) to export the data.

Its features include filling the forms, entering an item into the textbox for searching, etc., will make it much easier to extract data from the web. The project to be extracted can run either on your own machines (Local Extraction) or in the cloud (Cloud Extraction).

2. Fminer: FMiner is a visual Web-data extraction software for scraping the websites and also Web Screen scraping. It consists of a graphical user interface which allows you to quickly exploit the Software's powerful engine of data mining for the extraction of data from web pages. The powerful engine is using webkit browser so that it will extract data from most of Websites which includes dynamic websites with Javascript/Ajax.

3.Parsehub: Parsehub is a Web-Data extraction software/Data mining tool for website scraping. It is having an easy to use graphic interface. The data can be scraped from dynamic websites, its scalable and the data can be stored in cloud. This tool helps the analysts and consultants to backup the business cases of scraped data, sales leads by scraping sales funnel, data scientist to extract data for research, clean those datasets and visualize and helps ecommerce by extracting different products and their prices.

4.Mozenda: The extraction of information/images with point and click interface. It is composed of "Agent-Builder" and also a "Web-Console". The Agent created with the help of Agent-Builder can then be run in console which enables to manage, organize, view, export and publish the information. All agents can be run in Mozenda Data Centre.

After the data extraction, the process of Data Cleansing or Data Cleaning is done. This process of data cleaning involves detecting the inaccurate information from the record set, table and database and replacing it, modifying it or deleting the coarse information.

The process of data analysis and visualisation is done in order to provide:

- The comparison charts and pictorial graphs of different web scraping tools based on their standard plans, ratings, usability, functionality, easy to learn, storage capacity, crawler configuration, URL queries etc.

Some different graphs are plotted based on different category of industries in each district.

II. LITERATURE REVIEW

A. *General Facts about Web Scrapping*

Numerous definitions of web-scraping arose during the literature research. All the three presented definitions below mention the data extraction from multiple sources. They vary in the form of the initial sources for the extracted information.

Often it is essential to collect data from web which aims for Human Readers, not the Software Agents. The process here is known as "web scraping". (Apress, 2009).

The First definition mentions the data sources, which are originally designed for human readers. Such definition has proved itself as obsolete. With the evolution of automated techniques the possibility also emerged by extraction from software readable sources. However, it must be considered that the date of publication was 2009. At the times there were a very limited sources of Application Programming Interface (API). There were approx. 750 available sources available, compared to 17175 listed in 2017 held on Public API directory available on Programmable Web website (Berlind, 2015).

Web scraping alias web extraction or harvesting, is a method for extracting information from WWW and saved the info in a file or database in personal computer for future use or analysis. Frequently, web info is scrapped using HTTP or through a web browser. It is achieved either by a manual approach by the users or automatically by a bot or web crawler. Since enormous amount of heterogeneous data is frequently produced on the WWW, web scraping is broadly recognized as an efficient and robust technique for gathering big data. (Mooney, 2001)

The current state of affairs is more precisely portrayed by the second definition, where Web Scraping is mentioned as one of sources for big data collection. The Web Scraper receives, processes and parses the data from a specified source as shown in Figure 2.1.

Web scraping tools can be used for multiple purposes in various scenarios. Few uses are:

1. Collect Data for Market Research

Web scraping tools can assist keep you informed on where your company or industry is positioned in the next six months, serving as a powerful tool for market research. The tools can retrieve data from multiple data analytics providers and some market research firms, and consolidate them into one spot for easy reference and analysis.

2. Extract Contact Info

These tools can also be utilised to extract data such as emails and phone numbers from several websites, thereby facilitating to have a list of suppliers, manufacturers and

 DOI : 10.35291/2454-9150.2019.0435

other individuals of interests to one's own business or company, beside their respective contact addresses.

Figure 1.2: process of web scraping

The definition below does not mention many details. However, it succinctly captures the activities of Web scraping most precisely.

Web Scraping is the process of querying the origin of information from websites, retrieving the outcomes and parsing the page to retrieve the results. (John J. Salerno, 2003).

B. Related Work

Anand V. Saurkar, Kedar.G. Pathare and Sweta.A.Gode [1] discussed that from the evolution of World Wide Web, the internet user's scenario and the data exchange is rapidly changing. The new technologies are promoted day by day to boost up network as common people join the internet and start to use it. Daily use of internet causes that a tremendous amount of data is available/generated on internet which may be related to business, research or academics. Earlier most of the user were using the manual technique of copy-paste for gathering data and analysing them, but it was a tedious technique as it was time consuming. To solve this issues, new technique called Web Scraping was introduced which is used to generate the data from any unstructured format from web and store those data in structured format either in central database or spreadsheets. This paper is mainly focused on to get information from website with an automatic process. Web scraping is more reliable, efficient and effective technique where this process of automatic data retrieval system is fast compared to Dom Parsing, HTML parsing, Manual Data Collection etc. By using web scraping terminology user can easily extract unstructured data on single or multiple websites into a structured data automatically. The main aim of this technique is to get information from web and aggregate into a new dataset with minimal time and effort.

Vojtech Draxl [2] explained about available techniques of web scraping such as manual data collection, DOM Parsing, HTML Parsing etc, the complete procedure of web scraping and their progress in the previous years is explained with an example. Currently available software tools such as import io, parsehub, fminer, dexiio, outwithub, scraper etc are listed with a brief summary of their functionalities.

Chandni Saini ,Vinay Arora_[3] explained that, At present, the World Wide Web (WWW) is flooded with massive volume of data. According to this increase in Popularity of the Internet, the tough task is to search for the meaningful data among billions of sources in WWW. The end user is provided with the documents from the Information Retrieval which satisfy his needs. The precious information can be collected from internet using different search engines. Web Crawler is a major part of search engines; which is an Automatic script or program that can browse the data from World Wide Web in an Automatic Manner which even uses indexes for browsing. This Process is called web crawling. This paper covers, study on different strategies of Information retrieval in Web Crawling has been introduced which covers classification of 4 categories Which is: focused, distributed, incremental and hidden web crawlers. Thus, on this basis of user customized parameters the comparative analysis of various IR strategies has been performed.

C. Disadvantages

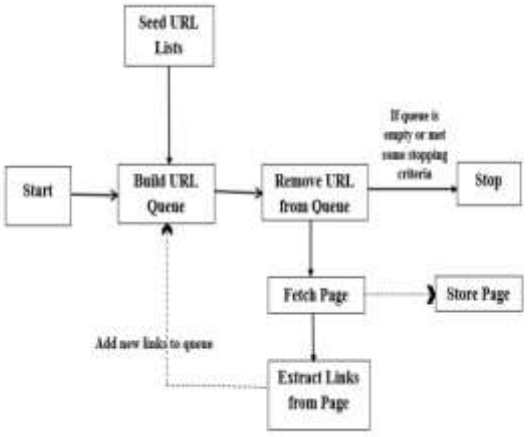
Clients used to find the data and retrieve the Industrial details manually and save the obtained information in any specific format as preferred in the Organisation themselves. As the Data collection is done manually its prone to many errors [inconsistency in entering data and mis keying] and the process is time-consuming, the individual would then choose the categorisations constantly each time.
Another existing system is web crawler. A web crawler (alias Web Spider or Web Robot) is an Automated Script which browse the WWW in an organized and Automated manner. This approach is called Web Crawling or Spidering. Several legal websites, in a particular search engine, spidering is used as a way of ensuring current or present-day data.

III. METHODOLOGY

The detailed description of working of 4 different tools: Octoparse, Fminer, Parsehub and mozenda of web scraping. The comparison between 4 different tools of web scraping. The analysis of industries data collected throughout the process and visualising it. To create a highly capable web scraper for fetching the information of the industries who are using fleet operators for transportation of their goods

 DOI : 10.35291/2454-9150.2019.0435

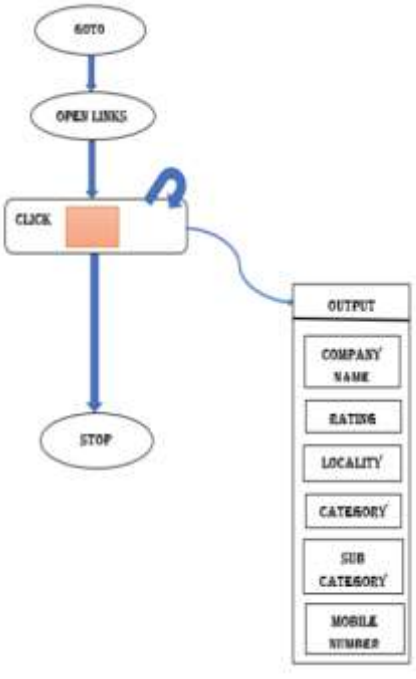
realising different API's, collecting certain information from the API sources and formatting it in order to retrieve the equivalent info of the data extracted.



DATA ACQUISITION:

It is also known as data-collection which is the first stage of this project. Usually, we can't find the entire dataset required in one place as it is spread across the **line-of-business** (**LOB**) applications and all the different systems. Thus as a part of this project the data related to industries in every particular districts of Karnataka have been collected through manual process and even using the web scraping tools. The data collected contains some of the contact information such as address or locality, phone numbers, the type of industries and its categories, Ratings related to that industry etc.

The process of data acquisition done with the help of different web scraping tools accordingly, as per the steps of each tool is as follows:



PARSEHUB PROCESSING STEPS

Step 1: Select a website Url. Find websites which have the details of industries data that you want to use.

Step 2: Select each and every field you want and rename.

Step 3: Train & Test the data. Extract html, text, images, links from sites.

Step 4: Download the Data. Retrieve the scraped data in any of the forms such as CSV/Excel, JSON or the API.

FMINER PROCESSING STEPS

Step 1: Create a new project by giving a name and start the recording by clicking on a record button.

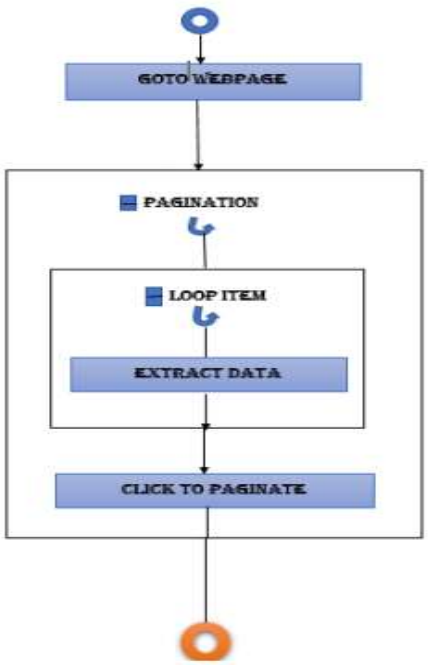
Step 2: Select a website Url. Find websites which have the details of industries data that you want to use.

Step 3: Click on scrape button and then create a table with the columns as needed by you to collect the data.

Step 4: Train & Test the data. Extract html, text, images, links from sites.

Step 5: Download the Data. Retrieve the scraped data in any of the forms such as CSV/Excel, JSON or the API

OCTOPARSE PROCESSING STEPS



Step 1: Select an advanced mode and switch to workflow mode for processing the task. Enter the URL of the page you need to process.

Step 2: Create a pagination loop to process with pages and the loop items for all the items you need to select.

Step 3: Extract Data is selected from action tips and data preview is selected to review the data.

Step 4: Save and start extraction.

 DOI : 10.35291/2454-9150.2019.0435

MOZENDA PROCESSING STEPS

Step 1: Open the Agent Builder**,** Type URL of website where your data to be extracted is present, Click on start new Agent from specified page.

Step 2: Click the first item in the list you want to create and select capture list and create an action for it.

Step 3: You can loop for the page by creating a page list and then can check out the preview of them.

Step 4: Export the data and save the content

3.2.2 DATA CLEANING:

Data cleaning is the process of Identifying the Incomplete, Incorrect, Inaccurate or Irrelevant parts of data (such as when the data is being collected from different web scraping tools some of the fields will be empty as there is no such data in webpage to fill that field/Data in a specified is mismatching etc. such fields are identified) and correcting/ removing/ modifying/ replacing those corrupted or Inaccurate records from Excel sheet, Record-set, Table, or Database.

IV. ANALYSIS OF TOOLS

After the Industries data collection, it is the time for deeper Data analysis. The Data analysis is a procedure for evaluating the industries data utilising some analytical and logical reasoning for the review of each piece of data provided by different web-scraping tools. This type of analysing is one of the steps which should be finalized when conducting a research project. Information is gathered from different resources, review all data, later on analysed to form some sort of findings.

	Octoparse	Parschub	Mozenda	Fminer
Customer support	Email, phone, community	Email, live chat, Forum	Phone, video chat, Email	Email
Price	$0-$249	$149-$499	$100-$500 Page credits	$168
Trial/Free version	Free version	Free version (15 days trial)	30 days trial	Free version (15 days trial)
OS Specification	Win	Win, Mac, Linux	Win	Win, Mac
Data Export formats	Txt, CSV, XLS Database	CSV, JSON	CSV, XML, XLS, JSON	CSV, JSON, XLS
Multi Thread Processing	Yes	Yes	Yes	Yes
API	Yes	Yes	Yes	Yes
Scheduling	Yes	Yes	Yes	Yes

Fig 4.1: General feature comparison

	Octoparse	Parschub	Mozenda	Fminer
All Paid plans	-	-	-	Yes
Download images and files to drop box	No	Yes	Yes	Yes
Download images and files to Amazon 83	No	Yes	Yes	No
Outer proxy	Yes	Yes	No	Yes
IP Rotation	Yes	Yes	Yes	-

Fig 4.2: Plans and support configuration

	Octoparse	Parschub	Mozenda	Fminer
Data Export				
API	Yes	Yes	Yes	Yes
CSV	Yes	Yes	Yes	Yes
JSON	NO	Yes	Yes	Yes
Google sheet	No	Yes [API]	Yes	Yes
Tableau	No	Yes	No	No
Web	No	Yes	No	Yes
Data Storage				
Free	×	14 days	-	15 days
Standard	3 months	14 days	1 GB storage	1 month
Professional	3 months	30 days	5 GB storage	3 months
Enterprise	-	30 days	5 GB Storage	-

Fig4.3: Data download and storage comparision

	Octoparse	Parschub	Mozenda	Fminer
Solving Captcha	Yes (Local)	Yes (Text inputs captcha)	No	Yes (Local)
Error report/Debug process	Yes	No	Yes	Yes
Scheduling	Yes	Yes	Yes	Yes
Test Run	No	Yes (Max 5 pages)	Yes	Yes (Max 3 pages)
IP Rotation	Yes (Cloud)	Yes	Yes	No
Visual mode	Yes	No	Yes	No
Scrape mode	Local & cloud	Cloud	Cloud	Local & cloud

Fig 4.4: Data extraction comparison

V. VISUALIZATION

At this stage, we begin data-manipulation process in a number of different ways, such as plotting the graphs(Bar graph, pie chart etc.) if appropriate which helps to calculate an average for the different trials of the experiment and discovering the correlations or by even creating tables in Excel formats. If the data is stored in form of table, the data can be easily filtered by different variables and also let you to calculate the measure of central tendency for your data.

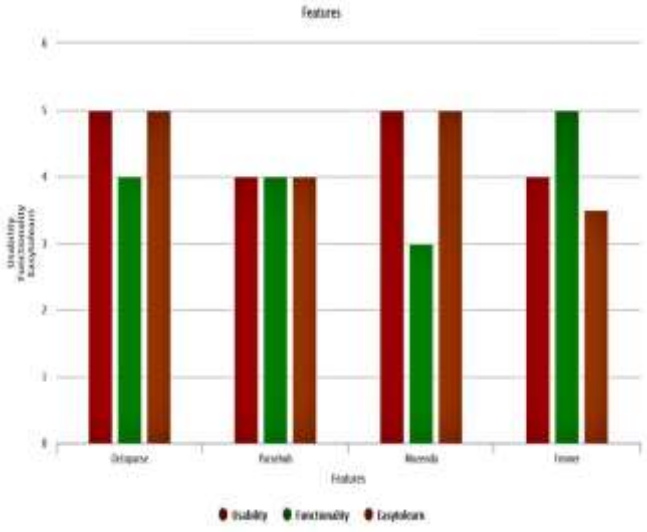


Fig 5: Comparison of different tools for their usability, Functionality and Easy to Learn in terms of rating them from 1-5

 DOI : 10.35291/2454-9150.2019.0435

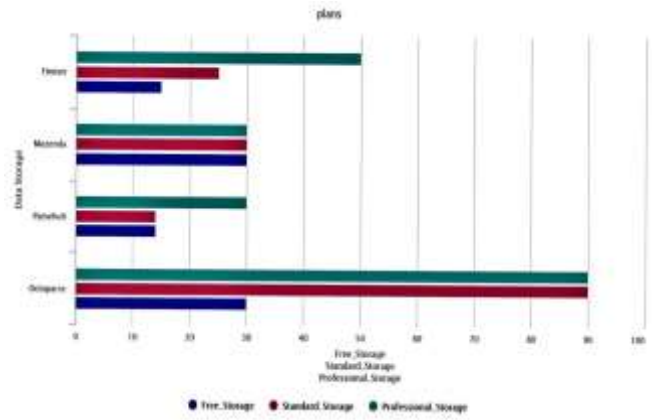


Fig 6: Comparison of different tools for their storage capacity in terms of days for different storage plans such as professional, Standard and free storage.

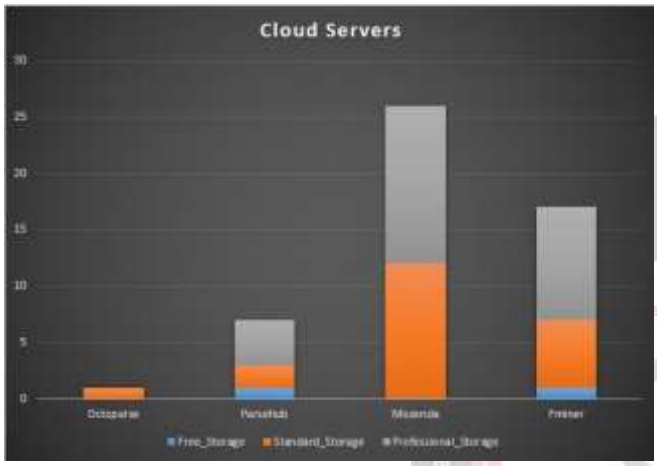


Fig 7: Comparison chart for the number of cloud servers available for each type of plan.

VI. IMPLEMENTATION

In order to collect the data needed from different sources of internet, manual data collection process was used. This Manual Data collection process were time consuming when the quantity of data is very large and also prone to human errors such as mis-keying. Thus, web scraping was introduced to collect large amount of data in a limited amount of time. There are many web scrapings tools which are used to collect/scrape data from different websites. The basic process of web scraping involves:

1. Identifying the URL of the webpage from where you want to collect your data.

2. If your information is present in more than one webpage, identify how to navigate automatically to those different web pages.

3. Record/Draw a flow chart to trace the features on webpages that label the information you want extract ie find the elements you want and identify the pattern in webpage that can be exploited (even the loops can be created for the process to extract all other similar elements).

4. Run the process so that the data will be extracted automatically (even test runs are available in some web scraping tools to check pattern and information being collected).

5. Save the data in Excel/CSV/JSON formats.

6. Analysis and visualisation is done using some R library files.

7. Some of the graphs are plotted in comparison with different web scraping tools considering parameters such as their functionality, usability, reviews from customers etc.

8. Some other graphs are plotted for analysis of industries present in each districts considering some parameters such as industry names, located in, Ratings etc.

VII. CONCLUSION

In the manual approach for analysis in any business through data analysis is more time consuming. Web-scraping overall approach is efficient and fast way to collect the data.

The comparison of different web scraping tools support retrieve and analysis of data , where each one of the tools have its own pro's and con's and they are suitable for different people in some ways or the others. Octoparse and Mozenda web scraping tool are from afar easier to employ than any other web scrapers. They are developed to make the web scraping possible for even non-programmers, thus you can foresee to master of it pretty quickly by watching some of the video tutorials. Parsehub is a powerful web scraper with robust functionalities. Though, they also require some of the programming language skills to master in it. Fminer is a web scraping tool with macro recording which scrape the dynamic pages with no differences.

Thus the Accuracy of data grows your business but this depends on the data and more efficiently the data will depend on tools of web scraping

REFERENCES

[1] G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in Plastics, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

[2] W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[3] H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch. 4.

[4] B. Smith, "An approach to graphs of linear forms (Unpublished work style)," unpublished.

[5] E. H. Miller, "A note on reflector arrays (Periodical style—Accepted for publication)," IEEE Trans. Antennas Propagat., to be published.

 DOI : 10.35291/2454-9150.2019.0435

[6] J. Wang, "Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication)," IEEE J. Quantum Electron., submitted for publication.

[7] C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.

[8] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style)," IEEE Transl. J. Magn.Jpn., vol. 2, Aug. 1987, pp. 740–741 [Dig. 9th Annu. Conf. Magnetics Japan, 1982, p. 301].

[9] M. Young, The Techincal Writers Handbook. Mill Valley, CA: University Science, 1989.

[10] (Basic Book/Monograph Online Sources) J. K. Author. (year, month, day). Title (edition) [Type of medium]. Volume(issue). Available: http://www.(URL)J. Jones. (1991, May

[11] 10). Networks (2nd ed.) [Online]. Available: http://www.atm.com

[12] (Journal Online Sources style) K. Author. (year, month). Title. Journal [Type of medium]. Volume(issue), paging if given. Available: http://www.(URL)

[13] R. J. Vidmar. (1992, August). On the use of atmospheric plasmas as electromagnetic reflectors. *IEEE Trans. Plasma Sci.* [Online]. *21(3).* pp. 876—880. Available: http://www.halcyon.com/pub/journals/21ps03-vidmar

 DOI : 10.35291/2454-9150.2019.0435