Indian Journal of Science and Technology, *Vol 12(26), DOI: 10.17485/ijst/2019/v12i26/145479, July 2019*

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Cricket Scrapper: A Tool Developed to Extract Cricket Players Data

Raheela Asif[1], Saba Izhar Haque[2], Saman Hina[2*], Syed Dayan Qaim[1], M. Hammad Haider[1] and Haris Ali Khan[1]

[1]Department of Software Engineering, NED University of Engineering and Technology, Pakistan; rahmed@neduet.edu.pk, dayanqaim95@gmail.com, hammadhaider96@gmail.com, haris.css@gmail.com

[2]Department of Computer Science and IT, NED University of Engineering and Technology, Pakistan; saba@neduet.edu.pk, samhaque@neduet.edu.pk

Abstract

Objective: To develop a generic web scraping tool that can extract cricket player's data for research purposes. **Methods:** Cricket as a sport has greatly plagued the world with its importance and entertaining nature. For a team the performance of each individual player has great importance for their victory. In this paper, an automated data scraping mechanism is devised to extract and store data of cricket players from a cricket website. This website contain web-enabled database of cricket players where the data is contained on multiple web pages and proposed mechanism is designed to extract this variation as per requirement of researcher. **Findings:** This tool will be helpful for data scientists/researchers to extract dataset as per their requirements. **Applications:** This generic tool will be applicable in analyzing the formation of cricket team as per rating of individual player. Moreover, this tool will contribute as an essential ingredient for research related in the domain of cricket analytics.

Keywords: Cricket Analytics, Information Extraction, Scrapper, Web Content Extraction

1. Introduction

The sports industry is well known for a large number of people including sports enthusiasts, professional in sports industry and has great contributions for researchers. Researchers and other analysts are interested in exploring different patterns of games and individual players as well. In parallel to this, the volume of data is increasing incredibly that need to be analyzed for different purposes (predictive analysis and current analysis).

The cricket industry has gone through immense height in prestige and became an important aspect of our economy. Massive amounts are being spent for the development of a team and heavy revenue is generated after the success of that team in any tournament or tour matches. With the economic factor involved, the need to statistically analyze the performance of a player and important decision making process based on that analysis has never been greater.

Cricket, a game of immense thrill, entertainment and pleasure has taken the sports industry by storm. Due to the monumental importance lay towards cricket, the economical factor involved has also sky rocketed. Nowadays, analysis of each player to evaluate their performance in past games has gathered great significance towards the success of any franchise or international teams. To enable a successful analysis, the need of gathering consistent, concrete and complete data is imperative.

In the world of internet and technology, web scraping is generally considered as web indexing, which performs the operation of indexing on web pages with the help of a

*Author for correspondence

web crawler or a bot. A web scraper outlining standards and strategies are differentiated. Development and design of a web scraper is based on the basic parameters on, provide the link of the desired data source, a concrete mechanism to extract the data from specified source and finally storing the extracted data into a desirable format to be used hereafter.

In this study, it was inevitable to deny that the availability of cricket players' data is scares. On the web there are numerous websites providing information of players around the globe as well as a detailed informative data on their performance (match by match).

In order to analyze a player's performance by implementing various statistical approaches, and performing cricket analytics methodologies, a mechanism to rapidly extract multidimensional data of that player must be present. Our study presents a mechanism, where innings by innings data of a player is gathered through a tool, which works on the sole principles of web scraping[1]. The rest of the study in this paper is organized as follows. Section 2 presents the detail why the need of a scraping tool is vital for rapidly extracting the data available on the web. Section3 presents the methodology and multiple channels used to develop such a tool and Section 4 and 5 finally discuss and conclude this study.

2. Aim of the Present Study

To extract content from the web is critically important if one wants to work on updated content. This process of accessing desired content is known as "web content extraction"[2]. This process increases storage operations and indexing of data. The presented tool is generic for all web-enabled databases having cricket player's information.

The main objective is to develop an automated generic scraping framework that can work dynamically to extract information about cricket for data analytics. This web scraping tool can enable a user to just type the name of the player, after which the developed scraper will automatically indexes the page where that player's data is located and delivers the question to the user to select the format of which the data is to be extracted of that player.

Storing, tabulated data from a web page into a .csv format file is time consuming and challenging to manage. Researchers have explored automated data scraping for different case studies such as soccer analysis[3] weather analysis[4].

Some researchers have worked on specific features of soccer, for instance performance rating index of a player with its usage[5]. For cricket, researchers have used publication databases to extract data for assessment of workload and its results on fast bowlers in terms of injury and performance[6].

In contrast with the related work, it was observed that there is no benchmark data or data extraction source available for analysts and researchers for exploration of data for dynamic purposes. In addition, there is no generic tool available that can extract data for cricket analytics with customized settings. Due to this purpose, it was decided that an automated scraping framework should be devised to fetch the specified data from a website containing consistent and detail amount of that data.

For this purpose, a cricket analytics website, proven to be a perfect foundation to implement our scraping mechanism and extract data of every player around the world. The realization of this tool did not only scrimp the utilization of resources such as time and effort but also provided the user to extract data of every match, innings by innings, batting and bowling both into a .csv format with separate files for batting and bowling data of a desired player.

3. Methodology

To conduct web scraping and develop a fully-fledged scraping tool, the use of any programming language mostly python is considered for different case studies[4,7]. Some researchers have also worked using R studio for data scraping[8]. However, in our scenario, it was decided to use C# as the development language of the tool and implementation was carried out in asp.net 4.5 framework because it was aimed to develop cricket analytics tool that can be used for research purposes.

As it was decided to develop the tool, by gathering data from Howstats.com, the Document Object Model (DOM) of that website was taken into account. This study reflected us with the insight that Howstats.com was indeed a dynamic website and displayed data through AJAX requests.

However, the aim was to extract data of a player (innings by innings) which was present in tabular format on that website. There was a limited use of JavaScript, in displaying the match data, so it was not difficult to

decipher a link to the data available. Assuring that the use of typical and traditional methods for scraping should not be adopted. The main reason for this was to avoid any unwanted circumstances and legal issues from the web server.

Therefore, the process of scraping had to be automated. HTML Agility pack was discovered, HTML Agility pack is an agile HTML parser that builds read/write DOM and supports plain XPATH or XSLT[9]. It is a library in C# that enables us to parse HTML, PHP or even .aspx files. For better understanding, HTML Agility pack is used to implement scraping of multiple web pages present on the internet. Our objective to use this library was to breakdown the HTML page on which the tabular data of player has been located and extracts data from that HTML page.

After breaking down the desired page, the major focus was revolving around how to extract and save the tabular data from that web page to a ".csv file" format. This issue was addressed, by discovering Language Integrated Query (LINQ) in asp.net framework with AngleSharp. The use of AngleSharp enables us to define which capabilities are present for the browsing engine and making the engine appear in the form of a "browsing context", which makes the browsing engine considered a headless tab, resolving the legal issues of scraping the web page as well[10]. Through the utilization of AngleSharp library, a new document can be opened and in that document the elements and that were required can be extracted and saved in a .csv format by providing LINQ queries on that document. This mechanism provides automated extraction of player's data that can be used for further analysis in this domain[11].

4. Discussion

By the succession of this study, it was discovered that scraping techniques constitutes of many tools to be implemented and considered. Despite of the fact there are web-scraping techniques available but for dynamic websites, it is still challenging researchers to extract data of their use. Researchers are still facing problems of unavailability of quality research data to conduct advance analysis on cricket data. In addition, there is an abundance of data available on cricket players but the methodologies implemented to scrap the data are widespread and heavily dependent on the nature of the website.

5. Conclusion

The nature of this study was to reflect and exhibit different tools that can be implemented to leverage a scraping tool, for the possession of any desired data available on the web. An importance of scraping is also demonstrated, when there are no API's available that can be used to extract data from a website directly. After analysis of available resources and their justification of use, an automated scraper was developed to extract information of cricket players with customized settings. This research work aims to contribute towards statistical comparisons between players of different teams irrespective of their team ranking in international formats of cricket, generating a rating system based on their performance in recent and past matches and ranking them according to that rating.

6. References

1. Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. Date accessed: 20/01/2015. https://www.amazon.com/Automated-Data-Collection-Practical-Scraping/dp/111883481X.
2. Rahman AFR, Alam H, Hartono R. Content extraction from HTML documents. International workshop on Web document Analysis; 2001. p. 7–10.
3. Asif R, Haque SI, Zaheer MT, Hasan MA. Football (Soccer) analytics: A case study on the availability and limitations of data for football analytics research, International Journal of Computer Science and Information Security. 2016; 14(11):516–18.
4. Bonifacio C, Barchyn TE, Hugenholtz CH, Kienzle SW. CCDST: A free Canadian climate data scraping tool, Computers and Geosciences. 2015; 75:13–16. https://doi.org/10.1016/j.cageo.2014.10.010.
5. McHale IG, Scarf PA, Folker DE. On the development of a soccer player performance rating system for the English Premier League, INFORMS Journal on Applied Analytics. 2012; 42(4):339–51. https://doi.org/10.1287/inte.1110.0589.
6. McNamara DJ, Gabbett TJ, Naughton GJSM. Assessment of workload and its effects on performance and injury in elite cricket fast bowlers, Sports Medicine. 2017; 47(3):503–15. https://doi.org/10.1007/s40279-016-0588-8. PMid: 27435575.
7. Web Scraping with Python: Collecting Data from the Modern Web. Date accessed: 2015. https://yanfei.site/docs/dpsa/references/PyWebScrapingBook.pdf.

8. Khalil S, Fakir M. RCrawler: An R package for parallel web crawling and scraping, Software X. 2017; 6:98–106. https://doi.org/10.1016/j.softx.2017.04.004.
9. Getting Started with HTML agility pack. Date accessed: 26/07/2017. https://www.c-sharpcorner.com/UploadFile/9b86d4/getting-started-with-html-agility-pack/.
10. Uzun Erdinç, Nusret Buluş H, Doruk Alpay, Özhan Erkan. Evaluation of Hap, Anglesharp and Htmldocument in Web Content Extraction; 2017. Date accessed: 11/2017. https://www.researchgate.net/publication/321228381_EVALUATION_OF_HAP_ANGLESHARP_AND_HTMLDOCUMENT_IN_WEB_CONTENT_EXTRACTION.
11. Khan JR, Biswas RK, Kabir E. A quantitative approach to influential factors in One Day International cricket: Analysis based on Bangladesh, Journal of Sports Analytics. 2019; 5(1):57–63. https://doi.org/10.3233/JSA-170260.