JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 63 - 68
https://publishoa.com
ISSN: 1309-3452

Intelligent Web Data Extraction System for E-commerce

Dr P. Tamije Selvy, Ms M. Anitha, L. R. Vishnu Varthan, P. Sethupathi and S. P. Adharsh

ABSTRACT
Data is crucial in today's world for whichever domain is taken to offer a higher and more intuitive experience for a user, Sometimes data is readily available in raw format, and Sometimes we need to extract the data from other websites/sources using advanced web scraping technologies. However, a web scraper that is based on a set of rules fails in the real world because the website's content dynamically and rapidly changes over time which in turn also changes the HTML contents of the website content. This study investigates a mechanism to allow automated web data extraction, In this study, an intelligent web data extraction using convolutional and Residual Neural Networks (ResNet) is developed. Usage of ResNet in the training layer accelrates the over all learning speed of the model.

Keywords: Adaptive web scraping; Deep learning; Long Short-Term Memory (LSTM); Web data extraction; You only look once (Yolo), ResNet

I. INTRODUCTION

A. Deep Learning

Deep learning[1] is an advanced version of machine learning which works on the basis of artificial neural networks, as the neural network is going to mimic the neurons in the human brain so deep learning is also a kind of mimic of the neurons in the human brain. In deep learning, there is no need to explicitly program everything. The recent advancements in deep learning have helped us in applying these advanced techniques in Health care, eCommerce, EdTech sector etc.
Deep learning's application in extraction of web data is still in a very early stage, in addition to the collection of data from documents or web pages, this application requires visiting multiple websites and saving the data for analytics and visualization purposes. Despite the fact that web application development is undergoing a faster paradigm change with respect to dynamically displaying the content and websites within the browser, the maintenance of extraction engine for core data remains a significant challenge that brings in frequent human intervention. The correction of extraction engine scripts for hundreds of websites is wasteful and time-consuming, which brings in a need for the creation of adaptive and intelligent web data extraction systems that can work well for dynamic website changes.

B. Web data and its extraction

The increase in data in the web increases makes the development of an extraction system for web data difficult. Deep web data extraction technologies were explained in-depth, and different data extraction methods for structured and unstructured data were examined. The data of deep web extraction algorithms are super difficult to design due to the properties of deep web (e.g., larger and higher coverage of data, higher quality, and strong structure).
Deep web data extraction technologies were summarised in-depth, and different data extraction techniques for structured and unstructured data were examined. Deep web data extraction methods are tough to create due to the features of the deep web. Furthermore, the massive amount of data that has been kept concealed.
Furthermore, the vast amount of data hidden in web pages with noisy and uncleaned content (in the form of advertisements, blogs, page settings, and navigation buttons) creates further difficulties. In this sense, AI and Deep Learning approaches can aid in the construction of an intelligent web extraction system. Using layers of an artificial neural network model, the fast Region-based Convolutional Neural Network(R-CNN)model[2] was used to filter these photos., extract hidden and dynamic data from web pages, and then persist them in a relational database is one technique to extract such data.

With the advancement of intelligent systems techniques, knowledge extraction techniques with website wrappers has improved. The present web wrapper adapts to each specific web template before beginning the extraction technique. Such web wrappers train from prior patterns using rapid R-CNN, resulting in a site-independent web wrapper that, nevertheless, needs empirical validation of its outcomes. Due to the usag of the HTML Document Model (DOM) which has tree of records, many online wrappers may accurately do an assessment of tree structure similarities in the interwebs.
A new wrapper is trained using the extraction model which detetcs the features in the images Traditional object detection problems in image processing and computer vision have also advanced through the use of deep learning, particularly those involving Convolutional Neural Networks (CNNs), although this has not been applied to web data extraction. CNN is commonly used for reliable object identification, semantic segmentation (proposing plausible regions of interest using a selective search method), and object classification.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 63 - 68
https://publishoa.com
ISSN: 1309-3452

C. Yolo

Yolo [3] (You Only Look Once) is a deep learning model that was developed in the year of 2015. It is based on the classification of images and the location of objects. An image is classified into a number of categories, such as "cosmetics" and "apparels," but one category is allocated to it in image classification. Using the zone of interest pooling methodology, this technology swiftly evolved into fast R-CNN. As a result, the first fully variational models were faster R- CNN and Mask R- CNN[4] models. Yolo was also added to allow for real-time picture or object detection.

Despite advancements in the development of automated and intelligent information extraction systems, web data extraction is performed using modern and machine learning approaches, making the system adaptive to the layout changes in the website.

II. PROBLEM DEFINITION

Web scraping tools such as open-source, proprietary crawlers, and application programme interfaces are used to extract data. Conventional, machine learning, and deep learning-based automated end-to-end web data extraction systems all have an "Extraction Engine" as one of the fundamental subcomponents. It is implemented using a variety of techniques, including image identification, detection, and data extraction from web pages. Due to a change in website layout, the suggested system is dynamic, pulling data from a unseen location or layout of the web app.

Due to changes in layout of the website, classical, machine learning, and deep learning-based web data extraction techniques have limitations. Website not found error(404 error) , change in position of the web page, and the webpage does’nt exist arise as a result of dynamic changes in websites and damage the "data extractor and script generator." Extraction engines employing modern, machine learning, and deep learning approaches, as well as open-source tools like crawlers, are used to extract data from web pages. The logic for extraction is embedded into the primary engine using modern machine learning or deep learning approaches in all circumstances.

A. Web Crawlers

Despite advancements in the development of automated and intelligent information extraction systems, the extraction of data from the web is performed using modern and machine learning approaches, making the designed system subject to layout changes of the website.

B. Automation Tools

These are based upon the Selenium[4] automation tool for recording and playback tools that generate the Python code that we can run to capture the website layout programmatically.

C. Data Extraction

The website's data, pictures, and files could be extracted using extraction engines. It includes crawler rules and a strategy for quick and accurate extraction of data from the web. The engine which is used for extraction examines the given URLs and gathers all of the data that matches the given template.

The portions of the page are identified using a core extraction engine, which yields the XPATH[5] from which the data of significance is retrieved.

By incorporating machine learning and deep learning algorithms into the engine which does data extraction, the accuracy of extraction of the data is improved, and automated and intelligent web data extraction is possible.

III. SYSTEM ARCHITECTURE

The Yolo deep learning model, on the other hand, is a core component which is brand new that facilitates auto-correction. Yolo and the self-correcting architecture are discussed in the proposed architecture because it is introduced in a novel way. The Yolo architecture begins with a single image, from which a list of relevant images with bounding boxes is generated, each region is assigned with a label , a bounding box, and a probability score is calculated for each label for the bounding box.

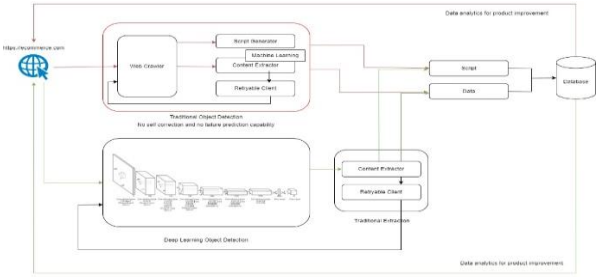

Fig. 1 Architecture of the proposed system

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 63 - 68
https://publishoa.com
ISSN: 1309-3452

As given in the Fig. 1, the propose system will have ResNet embedded in it and the crawlers will execute concurrently to fetch the data from the resources. Yolo has also subcomponents as explained the Fig. 1

The subcomponents of the Yolo architecture are as follows:

A. Convolution Layer
The image filters are trained in this layer to extract appropriate features from the image. The desired object features are then learned by these filters. The outcome of convolution is a t matrix of two dimensions called a feature map, which is produced by sliding the filter all along the input pictures. The input object, for example, may be a set of image files of a product (such as "mobile"), from which the CNN extracts features of several "mobile" objects before applying a filtering procedure to find the target "mobile" object.

B. MaxPool Layer
The CNN features are used to determine a preset number of bounding boxes that may include objects in the MaxPool[7] layer. The MaxPool layer, for example, the bounding box is applied to the regions (i.e., product information within the "mobile" object [e.g., "brand," "price," and "ratings"]) once the "mobile" object is detected. The data which we need is extracted from these areas. Because the object may have multiple zones, categorization and bounding boxes are used.

C. Object detection
Yolo algorithm is used to detect images or objects. To extract characteristics and feature sets from images and forecast bounding box output coordinates, twenty-four CNN layers and two fully connected network layers are used. Yolo CNN is deployed, which can recognise numerous text objects on a webpage picture. Yolo DarkNet-19 architecture is used to recognise text items. The DarkNet[8] architecture was chosen because it requires less processing power than other systems like ImageNet and GoogleNet.

D. Data extraction
Tesseract is utilised to extract the text once the output image or product detail image has been discovered. Tesseract is an open source Optical Character Recognition(OCR) system that uses artificial intelligence for searching for text and recognition of images, as well as a two-step method for adaptive and intelligent recognition. Tesseract, on the other hand, employs LSTM for text recognition. The Yolo model is used to identify the text objects on the page using a stepwise technique to accurately predict text from webpage images. The text is recognized once the bounding boxes with text have been discovered.

E. ResNet
ResNet has won several competitions, and its architecture allows deeper network learning. The code is modified to include the YOLO classifier at the end utilizing ResNet50 weights.

F. Bounding Boxes Prediction
The MaxPool layer proposes regions as an input to a fully connected network. Then Support Vector Machine (SVM)[9] is used classification algorithm to forecast object class and construct bounding boxes around the categorised objects. The object “mobile” has regions marked and categorised and forecasted using SVM algorithm in the case of the mobile website, and data is retrieved from these regions.
The goal of this research is to better understand and implement the Yolo model, which is based on deep learning, to enable the online system of extraction for data to self-correct in the case of various constraints in website. The proposed auto-correction capacity in the end-to-end online data extraction system is realised once the Yolo has been properly trained with photos and the appropriate loss function has been tuned.

IV. INPUT DATA MODELLING AND TRAINING

The ability to accurately specify the input data model
is one of the most important aspects of developing a successful deep learning based system. Specific product pages are referred to as "records" in the rest of the article, while product attributes are referred to as "regions." Each record belongs to a category of product, whereas each area belongs to a product detail. There are several regions in each record. The amount of data to the given input data model and the loss sustained during the object detection process are two parameters that can influence standard web raw data using deep learning models. The precision of the output data can be greatly improved by increasing or reducing the number of data or fine-tuning the learning parameters inside the loss function.
Creating a high-quality set of data: Pictures are critical for deep learning systems, and it takes a long time to create 1000+ image files from the retail and non-retail websites, including Bing single and multiple product details. In the retail domain, the Yolo system is trained to detect a single class of products (e.g., chocolate) and different class of products (e.g.,

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 63 - 68
https://publishoa.com
ISSN: 1309-3452

chocolate and mobile), whereas in the nonretail domain, it is taught to detect a single class of products (e.g., chocolate1) and different class of products (e.g., book1 and mobile1).
The weights of the ResNet and DarkNet are pre-trained before the model is fine-tuned by adjusting the weights of the final two layers. This method helps to keep the same extraction layers for the features while only training again the decision component, which cuts down on training time. The dataset is relatively limited when considering the scope of the study. The training, validation, and test sets each include 1000 photos in the ratio of 80:10:10 . The dataset is created using photos of product pages from various shopping websites, which are then bounding boxes and labels for each image present in the dataset.
Yolo identifies the top K bounding boxes for each image. Three separate groups of text items (series title, author, and market price) are chosen for identification in this study. The learning rate, weight decay, and momentum parameters used to train Yolo DarkNet-19 on the Common Objects in Context[10] (COCO) and Visual Object Challenge[11](VOC) datasets are similar to those used for training Yolo DarkNet-19 on the COCO and VOC datasets. Pascal VOC is an XML file that contains one file for each image in the dataset, whereas COCO is a JSON file that has only one file for the complete dataset.

The suggested framework is used in this experiment to extract data from a single product page (for example, "mobiles") on the Flipkart retail website without logging into the Myntra retail website, which mimics the regular situation. The next step is to create an account on Myntra's retail website and search for the same book using a different URL and layout on the product page. There is an error condition due to the change in product page layout and linked URL, which indicates a change in website layout. The input data consists of an image for a single category in which the output can be extracted and the bounding boxes can be drawn without the need of the user to be authenticated into the system.

V. LIMITATIONS AND FUTURE SCOPE

There are numerous use cases available on web data extraction utilising regular, standard analytical algorithms, and machine and deep learning algorithms. Web data extraction research involves unstructured, semistructured, and structured data webpages, as well as data presentation with static HTML [12] or dynamic displaying using JavaScript. Extraction rules or wrappers suited to a certain data source are commonly used in modern and machine learning-based extraction systems because most automatic and intelligent data extraction systems can only handle a limited number of document formats and are unable to adapt to changes in document structure, HTML documents with interesting data must be identified.
Data of relevance must be discovered on the website page, and rules for data extraction must be defined. To extract data from the enormous range of page formats available on the web, a system for creating data extraction rules must either be sufficiently generic or simple to implement.
Since web data providers constantly change the configuration or data of their pages, the information extraction system must be able to adapt to changes in the webpage structure.

VI. RESULT

Rules of extractions and wrappers suited to a certain data source are commonly used in modern and machine learning-based extraction systems.
Data of the required features must be located in the web page

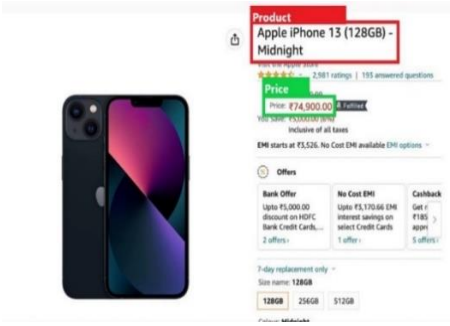

Fig. 2 Labelling of features by the proposed system

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 63 - 68
https://publishoa.com
ISSN: 1309-3452

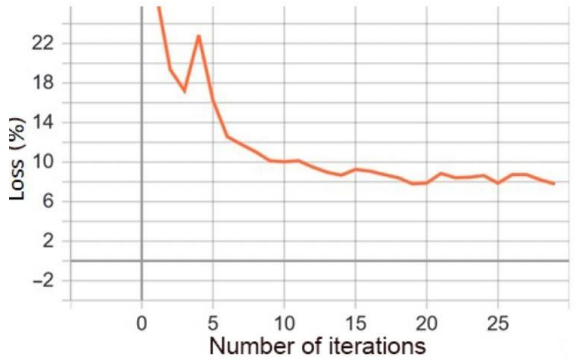

Fig. 3 Decreasing of loss (Y-Axis) with respect to the increase of iterations(X-Axis)

VII. CONCLUSION

This study proposes an intelligent and automated web data extraction system that uses YOLO and Tesseract LSTM[13] neural networks, ResNet, which not only is adaptive to the changes in website layout dynamically but also the data is extracted dynamically. The web data extraction system removes the subcomponents of the extraction engine, which are important to deep learning and machine learning techniques for data extraction to be done efficiently. As a result, the system has the potential to revolutionise the automated data extraction process from websites in the future.
Extraction of product detail from single and several web pages using real-world examples from the retail and non-retail worlds, demonstrating intelligence and adaptability. Advances in deep learning networks will inspire future research to reinvent the automated web data extraction process.

VIII. REFERENCES

[1] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436-444

[2] Albawi, Saad, Tareq Abed Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network." 2017 international conference on engineering and technology (ICET). Ieee, 2017.

[3] Huang, Rachel, Jonathan Pedoeem, and Cuixian Chen. "YOLO-LITE: a real-time object detection algorithm optimized for non-GPU computers." 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018.

[4] Bures, Miroslav, and Martin Filipsky. "SmartDriver: Extension of selenium WebDriver to create more efficient automated tests." 2016 6th International Conference on IT Convergence and Security (ICITCS). IEEE, 2016.

[5] Benedikt, Michael, and Christoph Koch. "XPath leashed." ACM Computing Surveys (CSUR) 41.1 (2009): 1-54.

[6] He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.

[7] Wang, Jie, et al. "CGFNet: Cross-Guided Fusion Network for RGB-T Salient Object Detection." IEEE Transactions on Circuits and Systems for Video Technology (2021).

[8] Setiyono, Budi, Dyah Ayu Amini, and Dwi Ratna Sulistyaningrum. "Number plate recognition on vehicle using YOLO-Darknet." Journal of Physics: Conference Series. Vol. 1821. No. 1. IOP Publishing, 2021.

[9] Chen, Yunqiang, Xiang Sean Zhou, and Thomas S. Huang. "One-class SVM for learning in image retrieval." Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205). Vol. 1. IEEE, 2001

[10] Mayershofer, Christopher, et al. "LOCO: Logistics objects in context." 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2020.

[11] Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Kämäräinen, J. K., ... & Ma, Z. (2020, August). The eighth visual object tracking VOT2020 challenge results. In European Conference on Computer Vision (pp. 547-601). Springer, Cham.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 63 - 68
https://publishoa.com
ISSN: 1309-3452

[12] Raggett, Dave, Arnaud Le Hors, and Ian Jacobs. "HTML 4.01 Specification." W3C recommendation 24 (1999).

[13] Shithil, Shaekh Mohammad, et al. "Container ISO Code Recognition System Using Multiple View Based on Google LSTM Tesseract." Computational Intelligence in Machine Learning. Springer, Singapore, 2022. 433-440.