Wrapper Extraction and Integration using GNN

Salman Naseer[a], Muhammad Mudasar Ghafoor[b], Sohaib bin Khalid Alvi[c] Iqra Zafar[d], Ghulam Murtaza[e]

[a]Department of Information Technology, University of the Punjab Gujranwala Campus, Gujranwala Pakistan.

[b]Director Campus, Department of Administrative Sciences, University of the Punjab Jehlum Campus, Jehlum, Pakistan.

[c]Department of Computer Science, Gift University, Gujranwala, Pakistan.

[d]Punjab Group of Colleges, PCW1, Zia ul Haq road, Gujranwala, Pakistan.

[e]Department of Commerce, Islamia University, Bahawalpur, Pakistan.

Corresponding address**:** salman@pugc.edu.pk

Abstract

Extracting data from the web is most prominent and discussing field now days. Extraction of useful semi structured data from the World Wide Web is the main aim. The extraction from the large web normally known as deep web is done by form submission cannot be done by any ordinary search engine. In data mining the automatic detection and extraction of data becomes bulky due to the uncertain structures of websites. Data extraction techniques developed till date are normally dealing with the extraction of text, audio, video etc. but there is a little and bit weak methods regarding the extraction of image data is the concern of recent research. One of the arts of image data extraction is DOM Document Object Model, it is a solution to extract the semi structured data but by the time the HTML documents are getting larger and contain more data. It is found that there is getting lengthy processing time and also emerged with noisy information. In the given research work we have tried to give a graphical representation of for the improvement of Wrapper Extraction of Image using DOM and JSON (WEIDJ). We have proposed the Graph Neural Network (GNN) to be used in wrapper extraction to improve the performance.

Introduction

Knowledge Discovery (KD) is a term used for extraction data from bigger database. While the extraction process of relevant information from the web is known as the data mining, by using this process the data is organized and categorized to determine the links between data. The data on the web could be accessed and categorized in three different types as USD unstructured data, SD structured Data and SSD semi Structured Data.

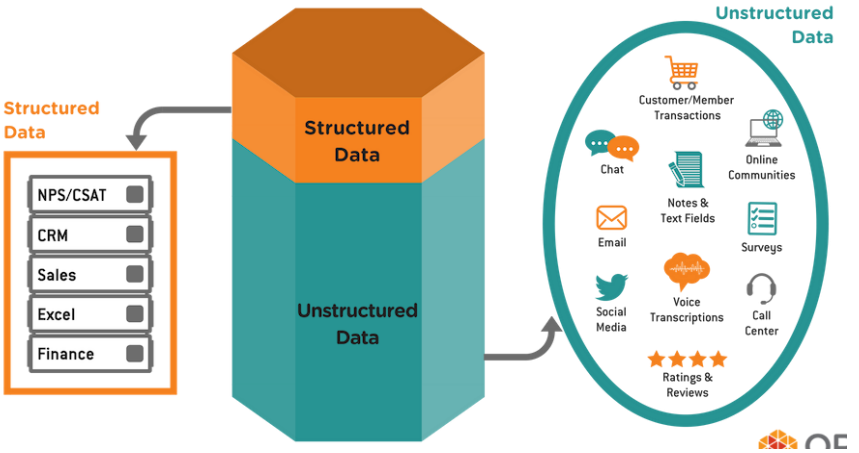


Fig 1. Ratio of Structured data and structured data

For circulation of information to users World Wide Web is considered as the main pool of data. The world stat of internet has described that 4.13 billion (Benkirane et al.) half of the population of the world is using the internet and creating a bulk of information on daily bases. The data regarding that is shown in Fig 2.

While the data integration is also considered as a part of the data extraction where the useful information is extracted from the web the process the called Web Data Extraction that allows the user to have a look at the data and analyze as well as arrange the data in a structured way such as tables format. Analysis and the extraction of data form the web is becoming a popular paradigm of research now a days.

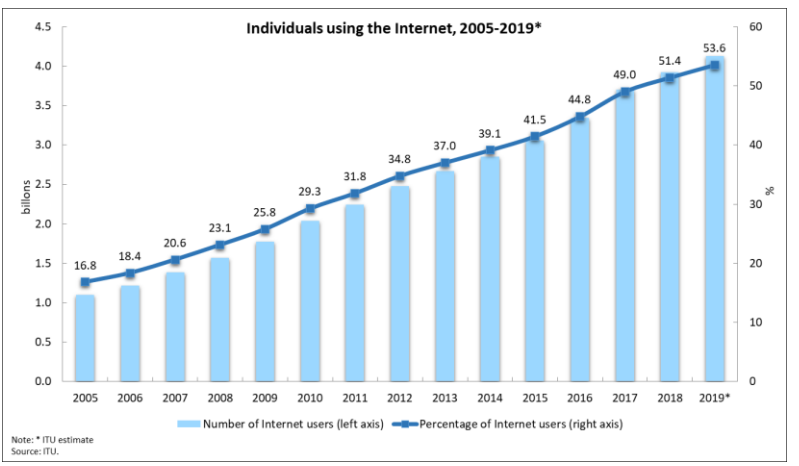


Fig 2.Individuals Using Internet 2015-2019

In early days there is manual way to extract this data by manually developing wrappers that are able to extract any of the given type of data. But in late nineteen

there comes some automatic techniques to develop the wrappers based on Xpath expressions that are able to extract the data from the web. Writing wrappers in Xpath manually is a difficult task due to the restricted rules then here comes some tools like W4F (Abrar et al., 2021) that provides a wizard based on GUI to generate wrapper based on HTML tags and elements and that wrapper is able to extract data from the similar type of websites like having the same structure.

Then there comes a fully programming language for the wrapper extraction by Gottlob and Koch (Riaz et al., 2022) which is base on the DOM tree to handle the semi structured nature of web pages. The extraction of facts using wrapping language is totally depends upon monadic dialogue over tree. This makes the wrapping language appropriate for being included into visual tools, fulfilling the condition that all it constructs can be implemented through corresponding visual primitives. The tree extraction technique followed by two steps one is segmentation and partial tree alignment.

In first step the segments of the web page are made this preprocessing does not extract any data only splits the web page based on DOM and then the next step the data is extracted based on DOM tree and then these trees are again into a new single tree. Partial tree alignment technique is a field of certainty and excluding those which are not aligned. But there is a factor of maintenance in this field in spite of all features. Machine Learning techniques come up with the new features that overcome the issue to an extent.

Machine learning techniques based on supervised and unsupervised machine learning are also plays a vital role in extracting data and also help the wrappers to extract data from the website the main need of this comes when we have to extract data from different sources but these sources should be sum up and categorized before the extraction of the data. Even there is a need of implementing the machine learning ruled to integrate extracted data in tabular format so that the extracted information could be helpful for the user. Some of the techniques used for extraction data from web are WHISK (Shahzad et al., 2021) that is a supervised learning technique that is used to generate rules to extract data from the text documents.

Then further more extraction techniques comes in the research like SRV take into account some documents containing tags and describe them as token and then extracts data on bases of those tokens. SRV was proposed by Freytag (Bokhari et al., 2022) and use Nive Base algorithm in addition with relational Learner. SoftMelay (Faouri et al., 2022) is also one of the Machine learning base extractor. It depends upon the Finite State Automata namely finite-state Transducer. STALKER (Ahmad, Boubakar, et al., 2022) is also a participant of this race it is a supervised learning based algorithm that is used to extract data based on human set tokens on the web page and then extract the data on the bases of those set tokens.

These some techniques are efficient and work worth but by the time there comes a lot of advancement in the era of internet the social websites like face book twitter and snap chat and other blogs brings a revolution in the studies of web extraction. When we comes toward the extraction of images from the web we got a lot of

problems while amerging that data and integrate into useful information. Because of the the image data have the same structure but there exists a different type of information in every image for that purpose WEID wrapper extraction Image Data is a technique developed to extract images.

Related Work

(Ishfaq et al., 2022), in this research they have tried to develop a tool to extract the semi structure data based on DOM and JSON this mediator tool is normally called wrapper used to extract data from the heterogeneous websites. Experiments will be conducted on Setiu wetlands web site and biodiversity web pages dataset for test bed. The extraction is focused on the image data and main aim is how to arrange that data in tabular from after extraction. It is propose that the model is based on DOM tree to mine data regions in webpage.

They describe the sequence of the working in a way like that first of all the web pages works as the input on which the DOM tree is constructed and then web segmentation is performed from segmentation the classification of the data is performed then JSON applied to the classified data and extract images finally arrange the extracted images in tabular form.

After input of URL the DOM is formed on the HTML pages to convert them in a pattern tree they can change the contents in the structure tree to get the data regions. Importance of this act is due to the data regions contains the useful information. Because every web page is developed using several use full content and these contents resides in tags so tags could be considered as data regions.

The data that is extracted from the web is divided in four categories in the given research these categories are image, text, audio, video and can be indicated by the parser key word “src=” and the extension of the data item like , JPG, .mp4, .avi etc. while the text could be accessed using simple tags.

They aim to find every appropriate graphic block present in the modern web page on this level. In essence, each node in the DOM tree can be supplied as a visible block, but nodes,< table>, and <p> should not be combined to form a single visual block. This is because they are frequently utilized for organizational purposes. Many rules are taken into account so that the visible block can be removed, as seen below: Tags are cued by the horizontal rule <hr> that is frequently presented in visual browsers. We may divide the section if this tag appears on every DOM node. If one of a DOM node's children has a very distinctive legacy hue. It won't be split any longer. Valid nodes are relaxed while the proper blocks are retrieved, going undetected. Separators can be used as markers to separate distinct sections on a website. This visual block segmentation is done to evaluate every multimedia component, allowing for the retrieval of all necessary data.

To avoid the noise information, they have selected the images for extraction that only have more than 70 x 70 size and all those images that have 50 x 50 or below are ignored from the extraction.

Results: The wrapper developed can extract multiple type of images like .jpg, .gif, .bmp they have taken 5 web pages and extract the images based on DOM, JSON and WEIDJ by using the DOM the number extracted images is less and also consuming more time while JASON also extract less images but consume a little less time as compare to DOM but when it comes to WEIDJ the number of images extracted is more than both of the techniques and takes a least time as compare to the both techniques. (Naseer, Liu, & Sarkar, 2019), there are multiple type of noise images that could be in following types local noise these could be advertisements, privacy notice, links for navigation, copyright . Global noise includes the advertisement, links of navigation, copyright. Prominent noise reflects the duplicate websites, reuse websites. Navigational noise could be of three types fixed noise navigational noise and web service noise. WEIDJ is a model that helps the user to extract the images from the websites when a user input the URL it will extract the data from semi structured format of the web base on the DOM tree.

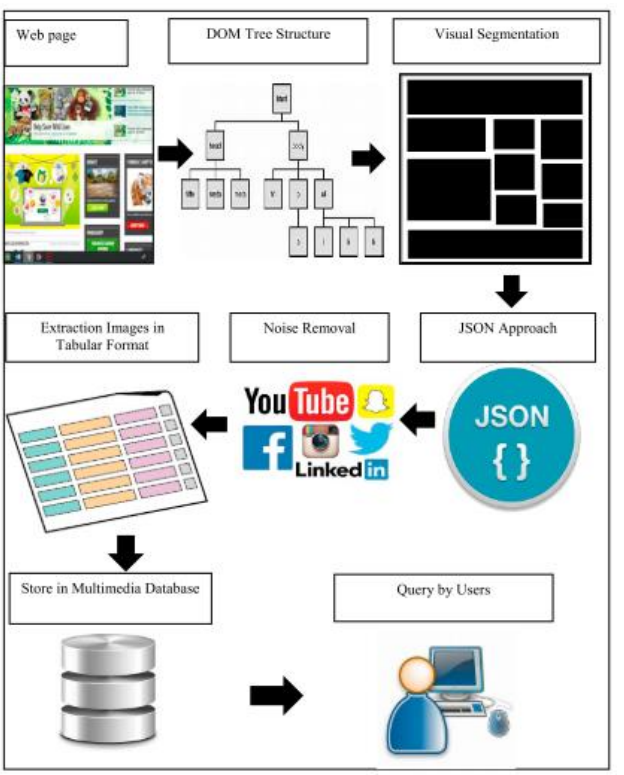


Fig 3. Workflow for Extraction Images in WEIDJ

WEIDJ use AJAX and XML to extract data the AJAX is used to reduce the processing on the client in client server model. It is difficult to describe the extraction rules, in this technique they are using JASON and DOM in a combination to get the accurate results from the web and arrange that data in the tabular form. There are many application that are focused on extracting data from the web pages (Ahmad, Manzoor, Naseer, Ghaffar, & Hussein, 2021; Khan et al., 2023) as the contact in web pages varies from web page to web page like contacts, lists, advertisements, additional information. In this research they are trying to explain that which technique is best to extract data from different web pages. Extracting data from the data regions is an easy task because these regions are stuffed with data. Previously the DOM is used as structure of web sites.

They have selected the FangJia web site to conduct the experiment using WEIDJ there exist multiple algorithms VIBS, MDR, DEPTA and VIDE those are selected to target and extract images from the web. Time analysis and precision and recall is considered in this study.

The single machine is used for the experiment using all the algorithms so that the time should be calculated in a proper way i.e. HP-UNOJQE6, Intel(R) Core(TM) i7-6500U CPU @250 GHz with 12.0 GB RAM in a Win10 64bit operating system platform (Naseer et al., 2019).

The analysis on this the time slices are taken with the difference of 5 pages at each reading WEIDJ performs first five pages in 12.69 and reaches to 48.84 till the 40 pages then VIBS starts at a good speed first 5 pages in 7 but delayed to 62.69 then here comes MDR performs 5 pages at 19.19 and last at the 164.16 running forward towards the DEPTA at first 5 pages 20.98 and when reaches to 40 pages it takes 180.71 at the end VIDE algorithm starts for the first 5 pages at 53.13 and ends at 40 pages with a score of 389.52.

Fig 3. Comparison Performance Existing Method

(Ahmad et al., 2021) they have proposes the web extraction method using DOM and JSON and built a model using PHP and XML and multimedia database layers. To retrieve data they have made multiple layers first layer is the user layer represent the user who implement the system second layer is the interface layer provides interaction between the user and the data source and user can identify the point of data to be extracted. Third layer is the source layer which contain the data that is to be extracted by the user and here the content could be classified whether text, image, video or other media. XML and JSON is the fourth layer at which the results are placed at XML and JSON document. Last layer is the storage layer that is a multimedia database where the extracted data is stored.

Now the implementation on PHP bases is that is the major step after the classification of multimedia data. They have defined an architecture that consists of web, adapter, meta data, repository and multimedia database. Also consist of three layers client tire, application tire, data tire. The web is a collection of semi-structured data in different fields and adapter is a place where the classification of the data is placed and because semi-structure and self defining characteristics of data extraction JASON and XML are employed as testing algorithms. Meta data is the short term storage for data extracted from web. To extract the data from the web all the items are placed in XHTML and HTML content and this could be easily transformed into DOM tree.

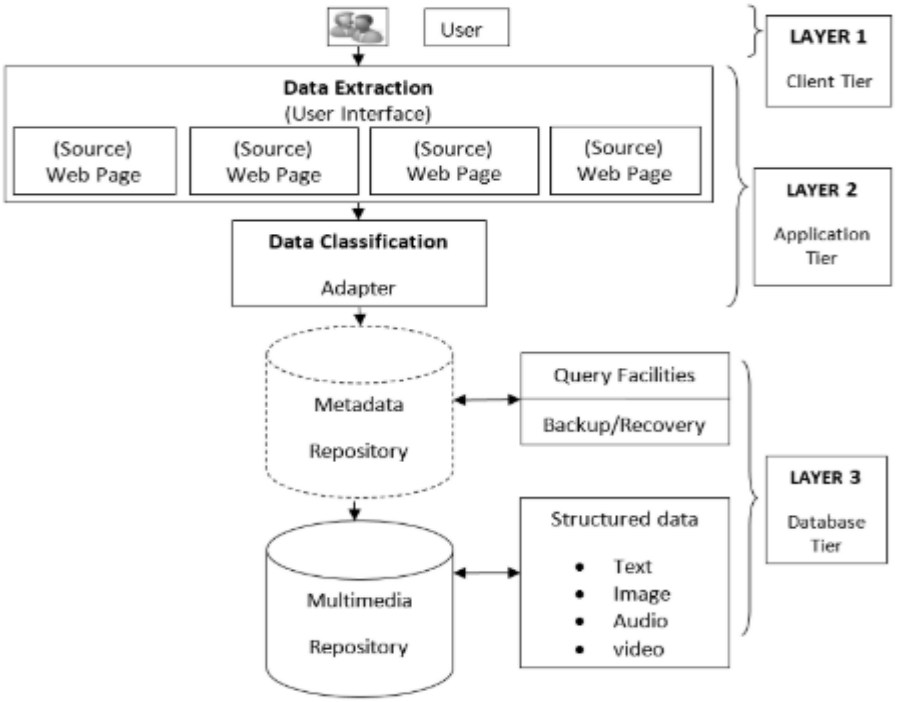


Fig 4. Multiple types of semi-structured data integration architecture (Ahmad, Ijaz, et al., 2022)

One algorithm based on the extraction technique of data from HTML pages and delete unnecessary data from extraction to reduce the processing time. Define list of pages classify data and represent the extracted data.

It is described that the JSON is more effective in extracting image data from the web as compare to the DOM.

(Naseer & Chaudhry, 2011) they have proposed the image extraction technique based on rules i.e. basic principle, visual analogy, principal of legend, principal of form the basic principal states that the image contain multiple items but one building block should be there, there color of the product in the image should be different product should be completely visible, target the important area, the object should be represented in environment. A picture chose ought to have a close by setting or then again a legend.

The setting of the picture ought to be fascinating for the item data search, the setting of the picture is intriguing if the catchphrases discovered relate to terms in the thesaurus. Form says that the image should be in the size of a particular photo graph.

All the above principals are applied in image extraction like legend is for the text and analogy is used on automatic form recognition. Then particular questions should

also be considered while extracting data from the web is the web page in proper distance, as worked on French so it is also considered that the page is in French then the occurrence of the image in page. The form of image should also b considered as on some parameters height and width should be between 60 to 160 pixels and file size should be less than 3.12 kilobyte. At the end the image should also contain the interested content. here the idea is to index the page with the content that it contains specifically using the n-grams explained in (Ali, Khan, & Naseer, 2022). The common words are then stored in the dictionary in form of single word and then these words are further used in N-grams.

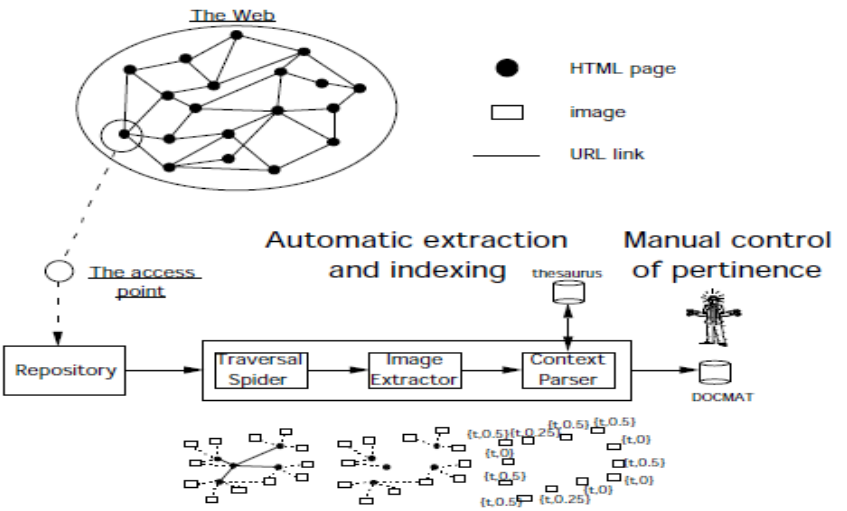


Fig 5. The principle of Wimex-Bot (Naseer et al., 2012)

The distance of the page from the main HTML page selected is calculated and this analysis is help full for the next image extraction. It also confirmed that the image selector only selects the image that is according to the above defined principals. Maxim Bakaev (Naseer et al., 2021)

Their visual page analysis algorithm that was found on : that algorithm takes the screenshot as input and collects UI by which the interface is formed. Additionally the DOM and HTML/CSS source code can be extracted. For preprocessing they improve the quality of input by making that in black and white and improve the edges. Then a threshold is set to produce binary images. Because of the moderately low measure of shading in UIs contrasted with general pictures, separate twofold pictures from various shading channels are not giving critical enhancements to ensuing preparing. This is finished utilizing OpenCV's edge identification for flat and vertical lines on the paired framework. The subsequent rundown of vectors is then checked for square shapes by looking for curved shapes with 4 corners over a base size. To distinguish text in the UI, we utilize a blend of OpenCV's nearby edge recognition. The jumping square shapes are clarified with the printed substance and

added to the past rundown of square shapes. Composite structures are recognized dependent on the untyped square shapes and text pieces distinguished in past advances utilizing choice tree rules. five on a level plane neighboring words with equivalent vertical arrangement exist, and text of min. 2 lines in vertical closeness.

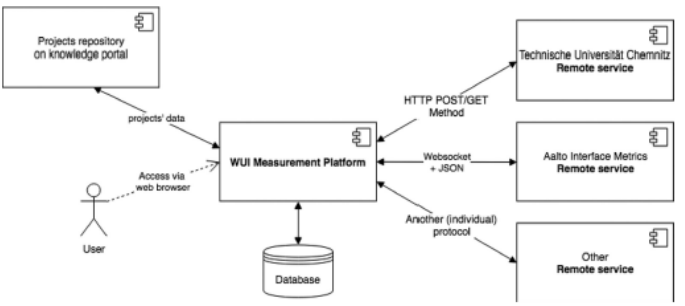


Fig 6. The scheme for the WUI measurement platform (Satti et al.)

In extension to this all UI metrics they have created a platform that is able to work with various remote services. Metric group entity is helpful in specifying the types like complexity related, color-related, the remote service entity class name use to reference the implementation of class for working with along with certain service. They given platform architecture is helpful in integrating the metric supplied by WUI.

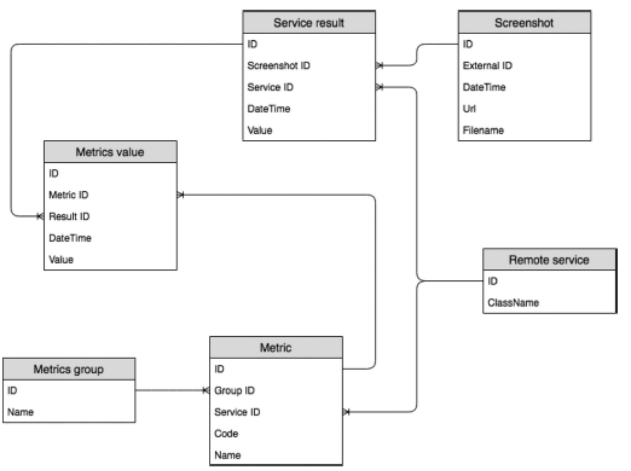


Fig 7. The scheme for the WUI measurement platform (Satti et al.)

(H. Ali et al., 2022)The website has been divided into sections that can be arranged in different categories, such as "boilerplate" vs. "not boilerplate" or something more specific like "heading," "article content," "table of contents," etc. The DOM tree,

render tree, or graphic block are just a few examples of the categories that are created based on units. Determining the level of the might balance the complexity and accuracy is crucial. Moreover, there are various methods, such as the visual page blocking method utilized by VIPS (Malik, Ghafoor, & Naseer, 2011) and diffbot (Jabeen, Zia, & Naseer, 2021). This method involves removing the web page's whole screen before applying a classifier to categorize the images based on their pixel content. The entire internet page is considered as one location in a top-down approach, and the non-coherent portions are then divided at regular intervals. In a different way, each pixel is treated as a separate region, and the comprehensible areas are then combined in each improvement. An appropriate number of regions are required. Moreover, one may simulate some page smoothing and define (or fit a capacity for) a Markov irregular field or other type of field that spans the full web page. Here, it is believed that because the squares on the web page seem to cluster together, names of nearby squares would serve as a reliable approximation of the square's mark. The last option uses a full-screen web browser to create a component vector that includes the location and other highlights of each HTML component. Finally, their methodology is reviewed, taking into account various approaches and balancing accuracy and unpredictability.

The tags are classified as relevant or irrelevant. The suggested tool was constructed in JavaScript and ran as an add-on for the internet browser. It formed an overlay over the DIV component that the mouse moved over and included text and image components. The client then decided whether or not the component was relevant. As with the first and labelled website page, it at that time spared the JSON design results. If the accuracy has been sufficiently enhanced, we want to switch to an arrangement using a web administration in the future. We also intend to expand the proving ground dataset to include a larger sample size of website

pages.

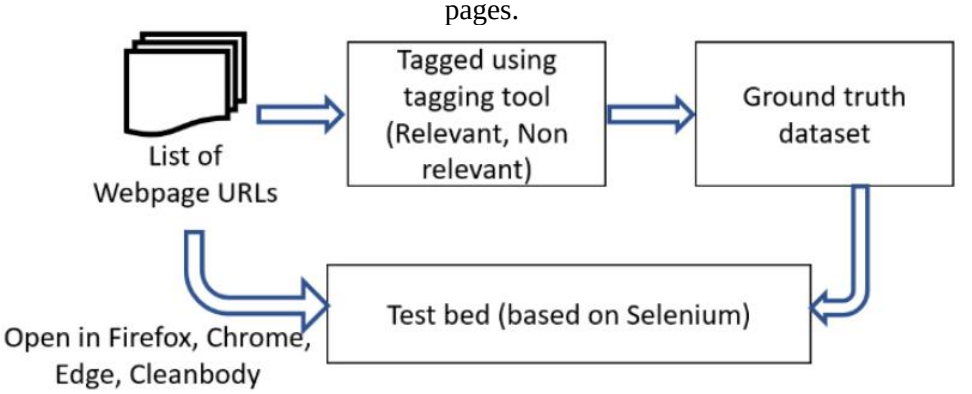


Fig 7. Architecture of the testing framework

"Intelligence Extraction Using Machine Learning Technics"

Authors in the research discuss that everyday millions and large amount of the data is uploaded and processed, for handling this much data we need a system that can handle, efficient and user friendly through which we can get meaningful and structured data (Sandhu, Haider, Naseer, & Ateeb, 2011).

Researchers discusses the modules of intelligence-based data extractions i.e CSV Extraction, Web Page text and images-based extraction, email, url and table extraction, video, image and PDF extractions. According to researcher's intelligence data extraction is nothing until we extract structured and meaningful data.

In the research they conclude that using csv-based data extraction user must provide name of the file with extension to program using the Pandas, and CSV libraries.

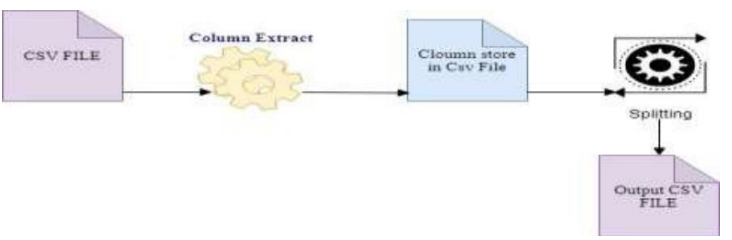


Figure 1: CSV Extraction

Web page extraction extracts the contents from the web pages and store them in file. It can be categorized into text and images. Text extraction extracts the data from the static websites, and it uses mainly 3 libraries i.e Sys, Beautiful soup, Request. similarly, images extraction uses four libraries which are: Request, Re, Beautiful Soup, and Urllib. Email extraction uses 3 libraries which are: Re, Request, and Beautiful Soup.

Urlextraction module, users provide URL to this module, it will read the website and return the url against it. 2 libraries used for this which are:Similarly, table extraction extracts the tables from the webpage and store data in CSV file

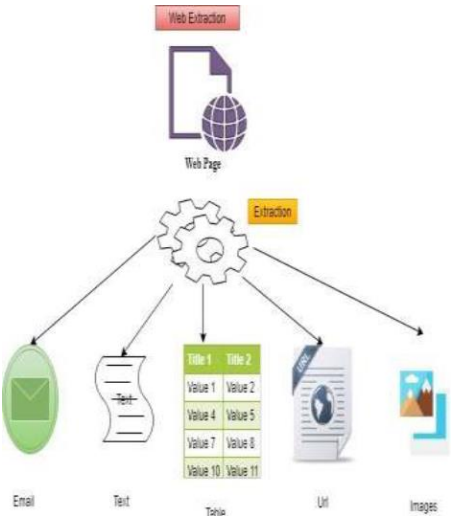


Figure 2: Url Extraction

Video extraction extracts every frame of a video user want to extract. Frame extracted will be in image format. Figure 3 explains module uses the 2 libraries CV2 and Os.

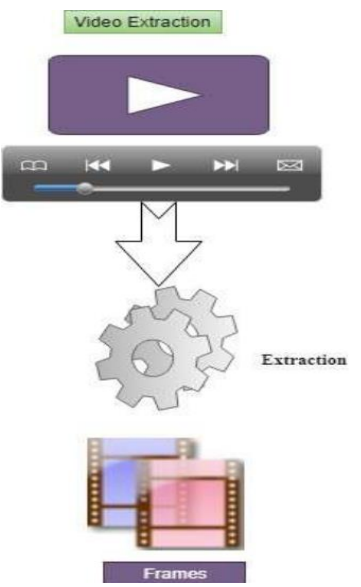


Figure 3: Video Extraction

image extraction is for extracting text from the images and it uses 2 libraries, which are explained in the Figure 4. Libaries are: Pytesseract and PIL.

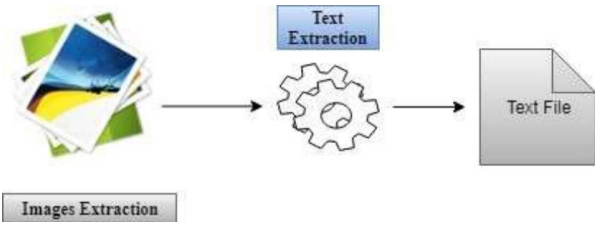


Figure 4: Image Extraction

PDF extraction module extracts the text from pdf files, researchers used only one library. Which is PyPDF2. This is also explained in Figure 5.

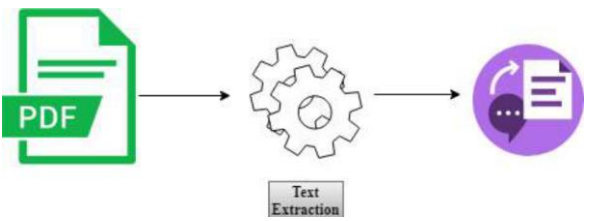


Figure 5: PDF Extraction

Results

Researchers from this research concluded that they system they proposed worked on different set of data collected from various sources, such as companies. System was able to extract the data which is structured and was meaningful for the users.

Figure 6 explains the system researchers implemented. Intelligence extraction system for the structured data.

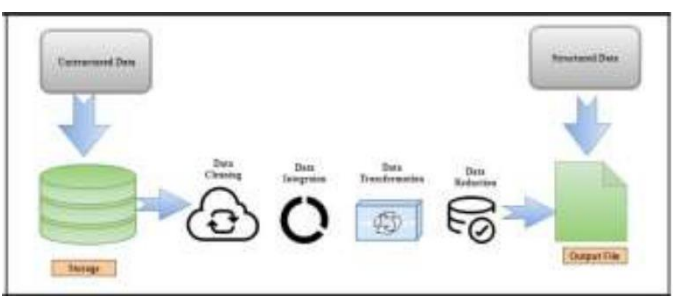


Figure 6: Intelligence Extraction

"A Vision-Based Approach for Deep Web Form Extraction"

WWW A big source of knowledge is the World Wide Web containing either in deep web or Surface web. Information in deep web, specially pages are generated through dynamic data quarries from the backed data base. It can't be indexed or scraped by

search engines. Data in the deep web no doubt in well structured and has good characteristics but it is difficult to use them as pages created by developers follows different structures. So this research focused on the ways to extract data from Deep web effectively and quickly ignoring the difference in the structures of the pages (Naseer et al., 2018).

Researchers used different approaches for the data extraction. These approaches are vision based, HTML based, Machine learning based, NLP based, and Ontology based. Researchers like Crescenzi (Naseer et al., 2017) analyzed the HTML code and data extraction rules are used and then He suggested RoadRunner, which means matching tags to extract html-based pages with info. Chang (Naseer et al., 2023) used the string pattern matching to extract the data, his model offers expressiveness and easy to understand the matching string. (Zaman-ul-Haq et al., 2022) used the DOM Tree analysis to extract the data, he set the certain rules for the DOM tree MDR, grouping the plurality of similar nodes to separate the data field, was then proposed, in which each node corresponds to the data records.

(Velusamy et al., 2021) suggested the VIPS, which means visual semantic, the DOM tree and visual features are combined to create a DOM tree with visual characteristics. This is used for the data extraction. This research also used visual based data extraction. The data is from the deep web is presented into table form. So, this research combined DOM tree with convolution neural network than find out the form from web page.

Researchers used the from recognition algorithm to understand the pattern of the page in tabular form and then match it with DOM tree to extract the data. Figure 7 explains the exact structure researchers used to extract the data (Ghafoor, Nawaz, Munir, & Saleem, 2022).

Algorithm 1. Form Recognition Algorithm

1. **Input:** The address of a Web page that contains a table **URI**.
2. **Execute Form Recognition**
   - Get the picture of the web page using the **URI**
   - Get the HTML source code using the **URI**
   - Build the DOM tree using the HTML source
   - Get the coordinate of the point related to the form in the web page
   - Get the DOM element of the form using Algorithm 2
   - Get the HTML source of the DOM element of the form
   - Generate the Excel spreadsheet file **table** using the HTML source
3. **Output:** A Microsoft Excel spreadsheet file that contains a Web form **table.**

Figure 7: Visual Algorithm

Results

Researchers proposed the VBF to extract the data automatically from using the visual based information. Which is combined with Visual information and DOM tree. Compared with VIDE, it is not needed to obtain visual information of all nodes. And using the method proposed data extraction was done efficiently and effectively from the deep web. Figure also explains that proposed method gave better results in less time, and gave more results.

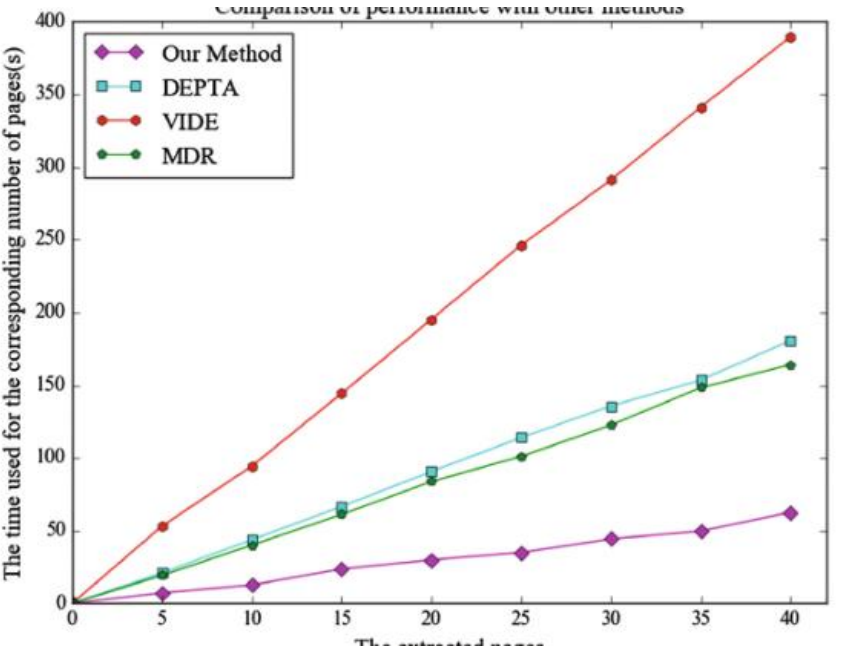


Figure 9: Results

"Content-based Title Extraction from Web Page"

NajlahGali and PasiFränti

Internet is the main source of information for the users. Web based applications and bots are used to collect information from websites foe the users. But the main issues here are the information on the pages are not well structured and meaningful for the users. There are a lot of irrelevant information such as advertisement, headers, footers. Search engines heavily rely on different methods of data extraction, which are semi-automatic, fully automated and manual approaches. This research focuses on automatic extraction of the data of the website titles. It is said that titles are the most obvious description of the website (Ahmad, Cherif, et al., 2022)

Titles are important as it gives users basic information about the website and user can understand the nature of website and can define objectives of websites. Based on the titles user also decides if they want to open the site or not (Khan et al., 2023).

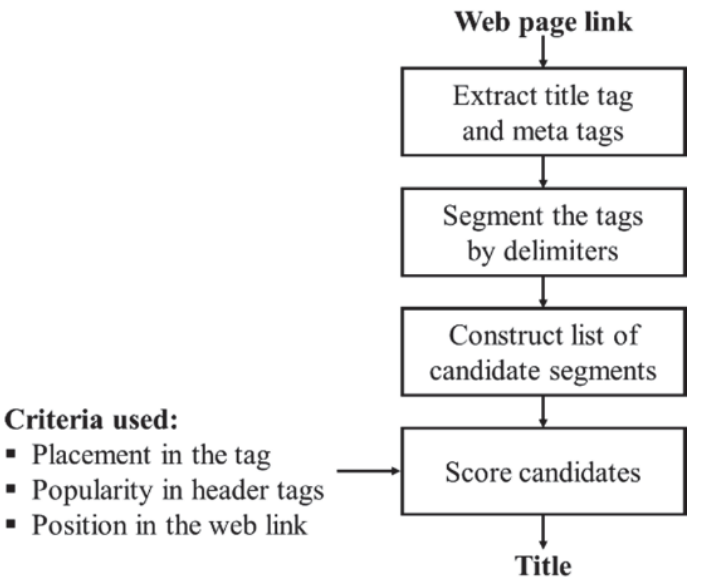


Figure 10: Model of Extraction

Figure 10 explains the model of data extraction used in the research. In the step 1 download the webpage and phrase it with the DOM tree, DOM helps in understanding the overall structure of the website and dynamically access and handle all the elements present in the website. Navigation through DOM will identify the meta tags such as title tags, og tags, and keywords. Keywords contains the phrases and words related to the website content. After the extraction of title and meta tags they used regular expressions to create set of phrases and words using the set criteria and rules. Basically, for the comparison purposes they used the following evaluation methods:

1. TTA title tag identifier
2. Title Tag
3. Title Finder

Results

In this research paper researchers proposed fully automated approach for extracting the website titles from the web pages. The method is called title tag analyzer. Proposed method fully and significantlyoutperforms and gave baseline from 0.62 – 0.84 in the average of similarity.

Title and meta tags in the websites usually contain the correct information about website but sometimes they also contain irrelevant information that need to be filtered out. Words on the website has highest impact on the selecting the correct title for the website.

"Extraction of Relevant Images of Boilerplate Removal In Web Browsers"

Introduction:

This paper discusses the development of a labelling tool, a testing system, and a dataset's consistency for the examination of web pages from a reader's perspective. They implement a method to include DOM-related functionality by exploiting a headless browser's rendering output.

Problem Statement and Approaches:

The challenge is segmenting the web page into small portions so that it is possible to define and categorize each individual unit. Any of the following may be the unit of the web page: One way to address the issue is to use a segmentation computer vision strategy, like VIPS and diff-bot. This includes, the way of scattering image on the basis of pixel likeness. The biggest change faced is to recollect the blocks of the image into the original DOM so the observer can see the image accurately. Obtaining a specific measure, such as text density, is another option. Any page smoothing can be applied, and a Markov random field or other form of field can be described that covers the full webpage (or suit a function for). Another alternative is to provide a headless browser that renders the webpage and generates a function vector. This is the strategy used (Naseer, Ghafoor, bin Khalid Alvi, & ul Islam, 2022).

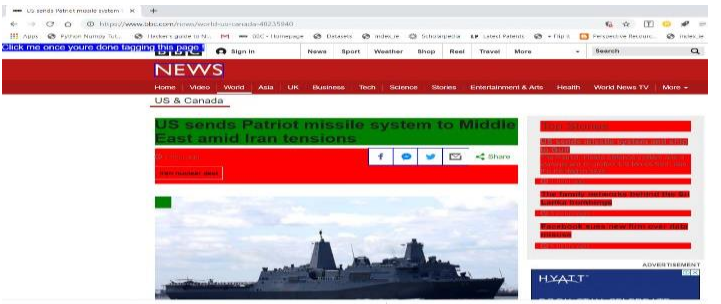
Testing Framework:

They constructed a forum and tools to evaluate algorithm performance and compare accuracy. With the help of telemetry data, they blended variables like the frequency of use from a dataset of approximately 1000 high impact URLs. They decided to focused on news websites initially since, according to the top Alexa sites, their usage is often strong.

Tool for Tagging and labeling:

They created an approach for labelling and marking individual web page elements as relevant or not relevant. The tool was designed in JavaScript and deployed as a web browser extension.

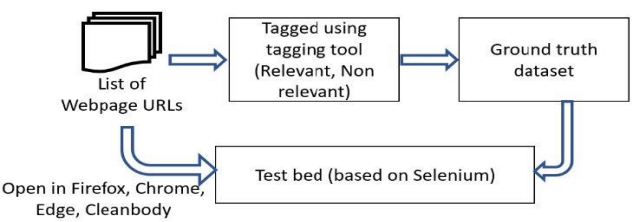
(a): SS of tagging tool:



(b): Tagged elements shown in output:


```
{
  "label": "R",
  "tag": "H1",
  "id": "240",
  "content": "<h1 class=\"story-body__h1 relevant\" id=\"240\">US sends Patriot missile
},
{
  "label": "NR",
  "tag": "DIV",
  "id": "243",
  "content": "<div class=\"story-body__mini-info-list-and-share-row noise\" id=\"243\">
},
```


Architecture of the testing framework:



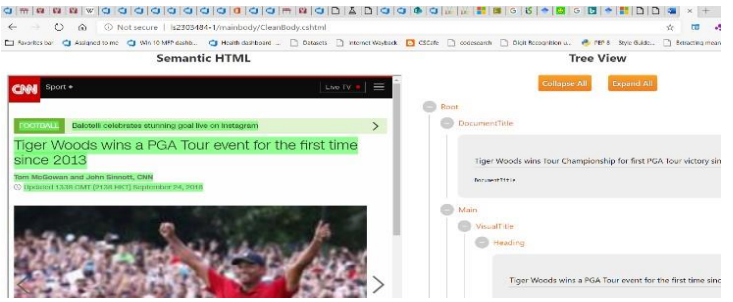
Building HTML tagged pages Database:

Using a tagging tool, we manually annotated a collection of roughly 1000 high-importance Websites. For every text and visual component, they classified it as relevant or not. It was kept as a JSON file.

Accuracy measurement using a test bed

They developed a test bed to compute and evaluate the reading view implementation's accuracy across different browsers. The testing method automated the opening of several save webpages using Selenium. The supported browsers included Chrome, Firefox, and Edge. The original framework was created just for text. We altered the testing framework utilizing the headless browser approach to obtain the output results for pictures.

The predictions that the headless browser tool generates for a given URL. The related content is indicated by Green:



The Extraction using the Headless browser solution:

They chose a customized headless browser solution because it is readily configurable to recognize pictures and is flexible. The rendering engine approach operated by rendering the webpage and generating a vector of traits for each webpage feature. With the help of the answer, we created a unique picture classifier. In order to collect information from the produced webpage, such as the placement and size of the pictures, the classifier initially employed the headless browser approach. Our classifier utilized a version of the MART gradient boosting technique. The default gradient boosting parameters are learning rate=0.4, shrinkage=0.53, leaves number=92, iterations count=50.For webpage dataset solution they use a 70:30 training: testing ratio.

Results and future work:

It is evident that the main navigators' reading mode solutions work well for text while producing varying outcomes for images. For a few particular dynamic web pages, our headless image browser approach performs admirably, but not for all of them. It is still in progress and is expected to boost its accuracy by offering more training and tweaking parameters.

- Precise classification for extracting relevant content from web pages on various browsers:

Browser	Precision	Recall	F1 Score
Chrome (text)	0.91	0.88	0.88
Firefox (text)	0.91	0.87	0.87
Edge (text)	0.94	0.85	0.88
Headless browser DOM solution (text)	0.68	0.82	0.72
Chrome (images)	0.31	0.5	0.34
Firefox (images)	0.6	0.65	0.61
Edge (images)	0.94	0.76	0.79
Headless browser DOM solution (images)	0.5	0.6	0.55

We plan to move to a solution using a web service in the future. We also plan to extend the dataset of test beds to include the wider selection of webpages.

A Survey of Feature Extraction for Content-Based Image Retrieval System

In this research paper author discusses the Content-Based Image Retrieval system (CBIR) which is a very challenging domain and used/part of many research fields now a days. Today images have an important role in media, communication and data transmission. Images made communication more reliable and user-friendly (Arsalan, Burhan, Naseer, & Rehman, 2022). But for communication and information sharing via images there is a need extract them and then they should process them according to the need and query demanded.

CBIR is a technique that enables a user to get an image based on a query from huge datasets that holds a massive number of images. CBIR system follows two steps.

- Feature Extraction
- Processing

1: Feature Extraction:

The image characteristics in the CBIR system are predominantly grouped into three main classes: color, texture and shape.

When an input or query image is given to CBIR it firstly extracts it's features and then a comparison is done between the feature vector of query and the feature vector of targeted image. But if the input information is too massive in size that it is difficult to process it then it is divided and converted into a set of main features called "Feature Vector"

The features that are extracted contains all that information that is mainly required for the query image/input data. So, further desired processing is done using this main set of feature vector instead of given input feature data. But when an image query or input data comes to the whole/overall system then a comparison is done between feature space and feature database and most relevant and appropriate images are returned in the basis of gap measure. And the relevancy degree of the images that are extracted can be calculated by some methods that gives satisfaction assessment called relevance feedback.

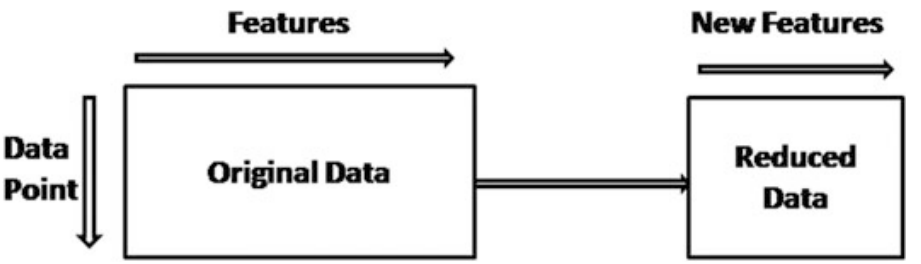
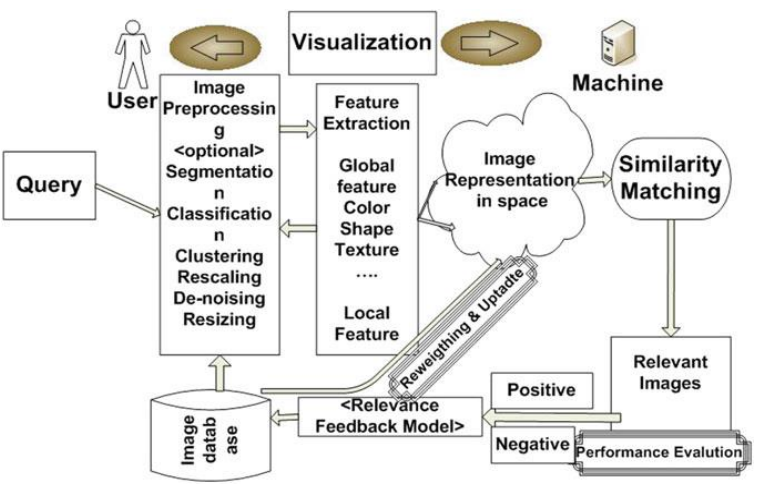


Image of feature extraction.

2: Processing:

Image preprocessing is an image enhancement approach previous to computational analysis. "Image pre-processing aims to enhance image visual appearance and database manipulation.

Visual representation of content-based image retrieval:



Conclusion and future work:

CBIR is very common among communication areas for massive datasets. The main aspects of CBIR are performance time and accuracy. But when we deal with massive datasets the searching process takes much time to complete which effects the system and make it less effective. There is a need of effective, relevant feature extraction system that gives most appropriate results. The solution is to divide and minimize the query related massive datasets, so that the overall consumption or usage of database is decreased and in return we will get relevant and appropriate results of an image retrieval process. So, in future the need is to design a new better system for better result optimization of content-based image retrieval system.

Analysis: From the above study we have made a concrete analysis and we can see from the figure 8 showing the use different tools in

Literature	HTML	DOM	XPath	XML	JSON	AJAX	Programming	Screen Sort	Artificial Intelligence
Sabri									
Sabri									
Mustafaman							**PHP**		
Walaiporn Nakapan							**n-gram**		
Maxim Bakaev							**OpenCv**		
ERI									**classifier**
Joy Bose									
Neha Ghosh							**CBIR (Image Processing)**		
Prof. Harish Patil									**Machine learning algorithm**

Jiachen Pu									**Convolutional neural network**
Najlah Gali									

Fig 8. Use of different techniques in different studies

Future Direction

The above study is all about the extraction of the images from the web pages. All the writers try to address the problem of image extraction and discussed the different ways that could be helpful. sabri tries to increase the efficiency of its developed model WIEDJ web image extraction based on DOM and JASON then the Mustafa tries to address the problem and also compare the multiple ways of extraction images in the last one Walaiporn Nakapan also describe a way to extract images based on n-grams and arrange those images based on the context written on the images. When we have a look at the ERI that have used some sort of classifier to classfy the images extracted. But we are here losing the use of Artificial intelligence to be used in the Web image extraction; we have here proposed the methodology that the GNN could be used to classify the images to integrate in the repository that is to be presented to the user. Furthermore the previous work is on the selection of web pages from the specific fields the data set could be selected from the different fields and the study could be generalized. Artificial amune system could also be used to reduce the noice from the image detection as the sabri adress a problem that they use the image size greater than 50 x 50 to reduce the noise in image extraction.

References

Abrar, U., Yousaf, A., Jaffri, N. R., Rehman, A. U., Ahmad, A., Gardezi, A. A., . . . Choi, J.-G. (2021). Analysis of Complex Solid-Gas Flow under the Influence of Gravity through Inclined Channel and Comparison with Real-Time Dual-Sensor System. Electronics, 10(22), 2849.

Ahmad, S., Boubakar, H., Naseer, S., Alim, M. E., Sheikh, Y. A., Ghaffar, A., . . . Parchin, N. O. (2022). Design of a Tri-Band Wearable Antenna for Millimeter-Wave 5G Applications. Sensors, 22(20), 8012.

Ahmad, S., Cherif, N., Naseer, S., Ijaz, U., Faouri, Y. S., Ghaffar, A., & Hussein, M. (2022). A wideband circularly polarized CPW-fed substrate integrated waveguide based antenna array for ISM band applications. Heliyon, 8(8), e10058.

Ahmad, S., Ijaz, U., Naseer, S., Ghaffar, A., Qasim, M. A., Abrar, F., . . . Abd-Alhameed, R. (2022). A jug-shaped CPW-fed ultra-wideband printed

monopole antenna for wireless communications networks. Applied Sciences, 12(2), 821.

Ahmad, S., Manzoor, B., Naseer, S., Ghaffar, A., & Hussein, M. (2021). A Flexible Broadband CPW-Fed Circularly Polarized Biomedical Implantable Antenna With Enhanced Axial Ratio Bandwidth.

Ali, H., Batool, K., Yousaf, M., Islam Satti, M., Naseer, S., Zahid, S., . . . Choi, J.-G. (2022). Security Hardened and Privacy Preserved Android Malware Detection Using Fuzzy Hash of Reverse Engineered Source Code. Security & Communication Networks.

Ali, N., Khan, K. I., & Naseer, S. (2022). Islamic Bank: A Bank of Ethics in Compliance with Corporate Social Responsibility. Sustainable Business and Society in Emerging Economies, 4(2), 295-302.

Arsalan, A., Burhan, M., Naseer, S., & Rehman, R. A. (2022). Efficient Interest Packet Forwarding Solution for NDN enabled Internet of Underwater Things. Paper presented at the 2022 17th International Conference on Emerging Technologies (ICET).

Benkirane, S., Guezzaz, A., Azrour, M., Gardezi, A. A., Ahmad, S., Sayed, A. E., . . . Shafiq, M. Adapted Speed System in a Road Bend Situation in VANET Environment.

Bokhari, S. A., Saqib, Z., Amir, S., Naseer, S., Shafiq, M., Ali, A., . . . Hamam, H. (2022). Assessing Land Cover Transformation for Urban Environmental Sustainability through Satellite Sensing. Sustainability, 14(5), 2810.

Faouri, Y., Ahmad, S., Naseer, S., Alhammami, K., Awad, N., Ghaffar, A., & Hussein, M. I. (2022). Compact Super Wideband Frequency Diversity Hexagonal Shaped Monopole Antenna with Switchable Rejection Band. IEEE Access, 10, 42321-42333.

Ghafoor, M. M., Nawaz, S., Munir, A., & Saleem, R. (2022). Lifestyle of Youth in Pakistan: Impact of Lifestyle Factors on the Depressive Behavior of Youth. Journal of Management and Administrative Sciences (JMAS), 2(2), 30-44.

Ishfaq, U., Shabbir, D., Khan, J., Khan, H. U., Naseer, S., Irshad, A., . . . Hamam, H. (2022). Empirical Analysis of Machine Learning Algorithms for Multiclass Prediction. Wireless Communications and Mobile Computing, 2022.

Jabeen, T., Zia, M. H., & Naseer, S. (2021). Right of Privacy: The Lacuna in Pakistan and Indian legal Framework.

Khan, M., Ali, I., Khurram, S., Naseer, S., Ahmad, S., Soliman, A.-T., . . . Choi, J.-G. (2023). ETL Maturity Model for Data Warehouse Systems: A CMMI Compliant Framework. Computers, Materials \& Continua, 74(2), 3849--3863. Retrieved from http://www.techscience.com/cmc/v74n2/50194

Malik, M. E., Ghafoor, M. M., & Naseer, S. (2011). Organizational effectiveness: A case study of telecommunication and banking sector of Pakistan. Far east journal of psychology and business, 2(1), 37-48.

Naseer, S., & Chaudhry, S. (2011). Lr-wpan formation topologies using ieee 802.15. 4. International Journal of Computer Science Issues (IJCSI), 8(6), 39.

Naseer, S., Ghafoor, M., bin Khalid Alvi, S., & ul Islam, H. S. (2022). Denial of Services (DoS) Attack: Implementation in Wireless LAN and Countermeasures. Pakistan Journal of Multidisciplinary Research, 3(2), 1-13.

Naseer, S., Ghafoor, M. M., bin Khalid Alvi, S., Kiran, A., Rahmand, S. U., Murtazae, G., & Murtaza, G. (2021). Named Entity Recognition (NER) in NLP Techniques, Tools Accuracy and Performance. Pakistan Journal of Multidisciplinary Research, 2(2), 293-308.

Naseer, S., Hussain, S., Raza, I., Chaudry, S., Mirza, J., & Raza, M. (2012). Mobile ad-hoc network routing protocols: A simulation and performance analysis using multimedia traffic. Journal of Basic and Applied Scientific Research, 2(10), 9925-9930.

Naseer, S., Liu, W., & Sarkar, N. I. (2019). Energy-efficient massive data dissemination through vehicle mobility in smart cities. Sensors, 19(21), 4735.

Naseer, S., Liu, W., Sarkar, N. I., Chong, P. H. J., Lai, E., Ma, M., . . . Qadir, J. (2018). A sustainable marriage of telcos and transp in the era of big data: Are we ready? Paper presented at the International Conference on Smart Grid Inspired Future Technologies.

Naseer, S., Liu, W., Sarkar, N. I., Chong, P. H. J., Lai, E., & Prasad, R. V. (2017). A sustainable vehicular based energy efficient data dissemination approach. Paper presented at the 2017 27th international telecommunication networks and applications conference (ITNAC).

Naseer, S., Saleem, R., Ghafoor, M.-M., Khurram, S., Ahmad, S., Sayed, A.-E., . . . Choi, J.-G. (2023). Temporal Preferences-Based Utility Control for Smart Homes. Intelligent Automation \& Soft Computing, 36(2), 1699--1714. Retrieved from http://www.techscience.com/iasc/v36n2/51161

Riaz, A. R., Gilani, S. M. M., Naseer, S., Alshmrany, S., Shafiq, M., & Choi, J.-G. (2022). Applying Adaptive Security Techniques for Risk Analysis of Internet of Things (IoT)-Based Smart Agriculture. Sustainability, 14(17), 10964.

Sandhu, U. A., Haider, S., Naseer, S., & Ateeb, O. U. (2011). A survey of intrusion detection & prevention techniques. Paper presented at the 2011 International Conference on Information Communication and Management, IPCSIT.

Satti, M. I., Ahmed, J., Muslim, H. S. M., Gardezi, A. A., Ahmad, S., Sayed, A. E., . . . Shafiq, M. Ontology-Based News Linking for Semantic Temporal Queries.

Shahzad, M. A., Paracha, K. N., Naseer, S., Ahmad, S., Malik, M., Farhan, M., . . . Sharif, A. B. (2021). An Artificial Magnetic Conductor-Backed Compact Wearable Antenna for Smart Watch IoT Applications. Electronics, 10(23), 2908.

Velusamy, P., Rajendran, S., Mahendran, R. K., Naseer, S., Shafiq, M., & Choi, J.-G. (2021). Unmanned Aerial Vehicles (UAV) in precision agriculture: applications and challenges. Energies, 15(1), 217.

Zaman-ul-Haq, M., Saqib, Z., Kanwal, A., Naseer, S., Shafiq, M., Akhtar, N., . . . Hamam, H. (2022). The Trajectories, Trends, and Opportunities for Assessing Urban Ecosystem Services: A Systematic Review of Geospatial Methods. Sustainability, 14(3), 1471.