



(51) International Patent Classification:
G06F 16/81 (2019.01)

(21) International Application Number:
PCT/MY2019/050094

(22) International Filing Date:
14 November 2019 (14.11.2019)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
PI 2018001927 14 November 2018 (14.11.2018) MY

(71) Applicant: **MIMOS BERHAD** [MY/MY]; Technology
Park Malaysia, Kuala Lumpur, 57000 (MY).

(72) Inventors: **AMRUDDIN, Amru Yusrin**; Mimos Berhad,
Technology Park Malaysia, Kuala Lumpur, 57000 (MY).

MOHD HELMI, Mohd Marzuq Ikram; Mimos Berhad,
Technology Park Malaysia, Kuala Lumpur, 57000
(MY). **JOHARI, Muhammad Awis Jamaluddin**; Mimos
Berhad, Technology Park Malaysia, Kuala Lumpur, 57000
(MY). **MUSTAFA, Rosnin**; Mimos Berhad, Technology
Park Malaysia, Kuala Lumpur, 57000 (MY). **GOON, Wooi
Kin**; Mimos Berhad, Technology Park Malaysia, Kuala
Lumpur, 57000 (MY).

(74) Agent: **PYPRUS SDN BHD**; Unit 16-08, 16th Floor, Block
A, Damansara Intan, No. 1, Jalan SS20/27, Selangor, Petal-
ing Jaya, 47400 (MY).

(81) Designated States (unless otherwise indicated, for every
kind of national protection available): AE, AG, AL, AM,
AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ,
CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO,
DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN,
HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP,

(54) Title: SYSTEM AND METHOD TO DETECT AND GENERATE RELEVANT CONTENT FROM UNIFORM RESOURCE
LOCATOR (URL)

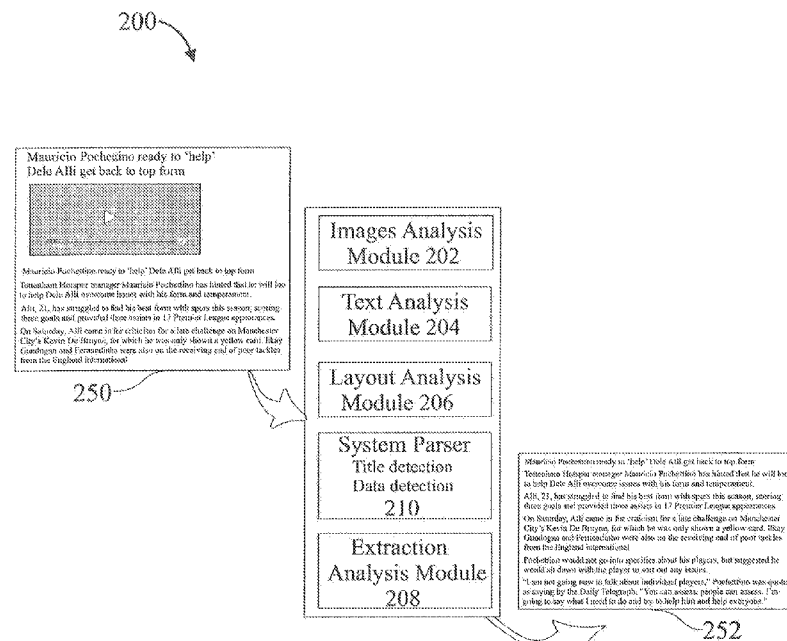


FIG. 2

(57) Abstract: Disclosed is a system and method to detect and generate relevant content from a received Uniform Resource Locator (URL). The system comprises an image analysis module (202), text analysis module (204), layout analysis module (206), and extraction module (208). The image analysis module (202) analyzes a plurality of images by capturing images from the received URL. The image analysis module (202) captures a URL screenshot from the received URL processed by an OCR engine. The text analysis module (204) analyzes the text by reading information about an HTML file from the received URL. The text analysis module (204) utilizes a headless browser to download the HTML file from the received URL and removes a plurality of HTML tags from the HTML file. The layout analysis module (206) analyzes a web layout by dividing the document object model (DOM) element blocks by scoring the height and weight of the web layout; The layout analysis module (206) uses a JavaScript module to mark and sort each DOM element blocks

KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

- (84) Designated States** (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

- *of inventorship (Rule 4.17(iv))*

Published:

- *with international search report (Art. 21(3))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*
- *in black and white; the international application as filed contained color or greyscale and is available for download from PATENTSCOPE*

SYSTEM AND METHOD TO DETECT AND GENERATE RELEVANT CONTENT FROM UNIFORM RESOURCE LOCATOR (URL)

FIELD OF INVENTION

5

[0001] The present invention relates to a system and method for content retrieval, in particular to system and method to detect and generate relevant content from a received Uniform Resource Locator (URL).

BACKGROUND

10

[0002] The continued growth of the data over the internet has great commercial value, and become an important source of intelligence information. Detection and retrieval of relevant, and important data from a uniform resource locator (URL) depend on various parameters. Existing systems and methods capture important data and extract the captured data from the URL if it contains a proper HTML format, does not contain AJAX pages, has sufficient

15

text information, and navigates the DOM structure (by using either XPath or CSS Query).

[0003] In case the criteria mentioned above are not met, the probability of detecting and extracting important data becomes low. Further, the DOM structure is bound to change depending on the website owner. Some websites use dynamic CSS class or dynamic DOM elements making it hard to get the right data.

20

[0004] US patent number 9,442,903 B2 issued to Katic discloses a method for generating preview data for online content. It receives a hyperlink from a user client device and acquires a set of data from a target referenced by the hyperlink. Then it stores a portion of the set of data on a server for analyzing the set of data to automatically generate a set of candidates for each of a summary and title of a preview for the hyperlink. The preview provides a synopsis of content at the target referenced by the hyperlink. Then it ranks the summary and title preview candidates based on relevance to the target referenced by the hyperlink. The relevance is determined via analysis of positive and negative signals for each candidate. It selects a top-ranked summary and title from the set of candidates. It transmits preview candidate with selected summary and title to the user client device. However, the method for generating preview data for online content uses an algorithm to

25

30

provide positive and negative signals for the data elements to select the proper data to use for the link preview. Positive signals indicate that a particular image, name, or sentence is relevant or appropriate to use in describing the linked content. Negative signals indicate that a particular image, or other data element, may not be suitable for use as automatically generated preview data.

[0005] US patent application number 2015/0066895 A1 issued to Komissarchik; Julia; et al. discloses a system and method of creating a domain-specific facts network. It includes a set of document acquisition servers to collect information from the internet and uses surface and deep web crawling mechanisms. It also includes a document repository database that stores all the collected documents. Then it includes a set of knowledge agent servers to process the documents stored in the database and extract candidate facts from these documents. The candidate facts are stored in the candidate database. It also includes inference and verification servers to integrate and verify candidate facts from the database and stores the results in the knowledge database. The knowledge database can be used as a source for data feeds and can be copied to a database server for an internet application, such as a business information search, job search or travel search. However, the creation of a domain-specific facts network converts unstructured and semi-structured information into a structured format to be used as a knowledge repository for different search applications and is not able to accurately detect and extract the important and high fidelity data when sufficient textual information is not provided by the internet and websites.

[0006] Therefore, there is a need for a reliable, efficient and cost-effective system and method to detect and generate relevant content from a received Uniform Resource Locator (URL).

SUMMARY

[0007] The present invention discloses a system and method to detect and generate relevant content from a received Uniform Resource Locator (URL). Furthermore, there is also a need for a system and method to perform high fidelity content detection on targeted URL by at least one of analyzing images (still image), and text (text extraction) of the web content, vision-based content extraction, and analysing block web-layout (size of web-layout) of the web content to extract useful information therefrom.

[0008] In one aspect of the present invention, there is provided a system to detect and generate relevant content from a received Uniform Resource Locator, URL. The system comprises a memory to store machine-readable instructions pertaining to generation of relevant content; and a processor coupled to the memory and operable to execute the machine-readable instructions stored in the memory. The processor comprises an image analysis module for analyzing a plurality of images by capturing images from the received URL, wherein the image analysis module captures a webpage of the received URL in a form of image file for processing by an Optical Character Recognition, OCR, engine; a text analysis module to analyze the text by reading information pertaining to an Hypertext Markup Language, HTML, file from the received URL, wherein the text analysis module utilizes a headless browser to download the HTML file from the received URL and removes a plurality of HTML tags from the HTML file; a layout analysis module to analyze a web layout by dividing a plurality of document object model, DOM, element blocks by scoring height and width of the web-layout, wherein the layout analysis module uses a JavaScript module to mark and sort each DOM element blocks by height and width; and an extraction module utilizes a system parser to retrieve a title, and a date to classify an array of text, wherein the extraction module uses HTML metadata received by the text analysis module and generates the relevant content in a textual format.

[0009] In one embodiment, the OCR engine provides an array of text and transmits the array of text to the text analysis module. Further, the text analysis module removes the HTML tags from the HTML file and processes the remaining array of text.

[0010] In another embodiment, the layout analysis module is adapted to further analyses number of words in each DOM element block to determine an article content and processes an array of text blocks sorted base on size of the web layout. In a further embodiment, the system parser retrieves the title by using DOM extraction for a plurality of <title></title> tags and retrieves the date using a regular expression on text data received from the text analysis module, the layout analysis module, and the extraction module.

[0011] In another aspect of the present invention, there is provided a method for detecting and generating relevant content from a received Uniform Resource Locator, URL. The method comprises steps of receiving, by an image analysis, a URL; analyzing a plurality of images by capturing images from the received URL, wherein a webpage is

captured in a form of image file from the received URL and process by an Optical Character Recognition, OCR, engine; analyzing, by a text analysis module, texts by reading information pertaining to an Hypertext Markup Language, HTML, file from the received URL, wherein the text analysis module utilizing a headless browser to download the HTML file and remove HTML tags from the HTML file; analyzing, by a layout analysis module, a web-layout by dividing a plurality of document object model, DOM, element blocks by scoring height and width of the web layout; retrieving by an extraction module, a title and a date of the URL through a system parser, wherein the extraction module uses HTML metadata received by the text analysis module; and analyzing and generating detected relevant content in a textual format.

[0012] In one embodiment, the OCR engine provides an array of text and transmitting the array of text to the text analysis module. In another embodiment, it further comprises removing the HTML tags from the HTML file and processing the remaining array of text by the text analysis module.

[0013] In another embodiment, The method according to claim 6, wherein the system parser retrieves the title by using DOM extraction for a plurality of <title></title> tags and retrieves the date using a regular expression on text data received from the text analysis module, the layout analysis module, and the extraction module.

[0014] In a further embodiment, each array of text data from the text analysis module, the layout analysis module (206), and the extraction module (208) are compared by using an intersection of a plurality of predefined sets of a matrix.

[0015] Accordingly, one advantage of the present invention is that it captures the important data and extracts the important data from URL if it contains dynamic HTMLS, and AJAX pages, and further extracts the important data when the URL contains less sufficient textual information.

[0016] Accordingly, one advantage of the present invention is that it analyses text by reading HTML tags from the URL received from a user computing device.

[0017] Accordingly, one advantage of the present invention is that it analyses web-layout by dividing the DOM element block by height and width for scoring.

[0018] Accordingly, one advantage of the present invention is that it detects title and date by using a system parser.

5 [0019] Accordingly, one advantage of the present invention is that it analyses content by returning final content in text format.

[0020] Other features of embodiments of the present invention will be apparent from accompanying drawings and from the detailed description that follows.

10

BRIEF DESCRIPTION OF THE DRAWINGS

[0021] In the figures, similar components and/or features may have the same reference label. Further, various components of the same type may be distinguished by following the reference label with a second label that distinguishes among the similar components. If
15 only the first reference label is used in the specification, the description applies to any one of the similar components having the same first reference label irrespective of the second reference label.

[0022] FIG. 1 illustrates a flowchart of the method for detecting and generating relevant content from a received Uniform Resource Locator (URL), in accordance with an
20 embodiment of the present invention.

[0023] FIG. 2 illustrates a system for detecting and generating relevant content from a received Uniform Resource Locator (URL), in accordance with an embodiment of the present invention.

[0024] FIG. 3 illustrates a flowchart of the URL inserted into the system, in accordance
25 with an embodiment of the present invention.

[0025] FIG. 4 illustrates a flowchart of the present system for capturing images and analyzing the inserted URL, in accordance with an embodiment of the present invention.

[0026] FIG. 5 illustrates a flowchart of the present system for reading HTML tag and analyzing the inserted URL, in accordance with an embodiment of the present invention.

[0027] FIG. 6 illustrates a flowchart of the present system for analyzing the web-layout block of the content from HTML page, in accordance with an embodiment of the present invention.

[0028] FIG. 7 illustrates a flowchart of the system parser for obtaining title and date, in accordance with an embodiment of the present invention.

DETAILED DESCRIPTION

[0029] Systems and methods are disclosed for detecting and generating relevant content from a received Uniform Resource Locator (URL). Embodiments of the present invention include various steps, which will be described below. The steps may be performed by hardware components or may be embodied in machine-executable instructions, which may be used to cause a general-purpose or special-purpose processor programmed with the instructions to perform the steps. Alternatively, steps may be performed by a combination of hardware, software, firmware, and/or by human operators.

[0030] Various methods described herein may be practised by combining one or more machine-readable storage media containing the code according to the present invention with appropriate standard computer hardware to execute the code contained therein. An apparatus for practicing various embodiments of the present invention may involve one or more computers (or one or more processors within a single computer) and storage systems containing or having network access to computer program(s) coded in accordance with various methods described herein, and the method steps of the invention could be accomplished by modules, routines, subroutines, or subparts of a computer program product.

[0031] The present invention discloses a system and method whereby the relevant, important and high fidelity data is automatically detected and generated from the Uniform Resource Locator (URL). The system and method include an image analysis module, a text analysis module, a layout analysis module, and an extraction module. The image analysis module analyzes images by capturing images from the received URL. The image analysis module captures a URL screenshot from the received URL processed by the OCR engine.

The text analysis module analyzes the text by reading information about an HTML file from the received URL. The text analysis module utilizes a headless browser to download the HTML file from the received URL and removes a plurality of HTML tags from the HTML file. The layout analysis module analyzes a web layout by presenting the webpage through Document Object Model (DOM) Application Programming Interface (API). Through the DOM API, the website is divided into DOM element blocks. Each of the DOM element block is scored with height and width of the web layout. In one embodiment, the DOM element block is scored based on an area of the element block. The layout analysis module may use a JavaScript module to mark and sort each DOM element blocks accordingly. The extraction module utilizes a system parser to retrieve a title, and a date to classify an array of text. The extraction module uses HTML metadata received by the text analysis module and generates the relevant content in a textual format.

[0032] Although the present invention has been described with the purpose of detecting and generating relevant content from a received Uniform Resource Locator (URL), it should be appreciated that the same has been done merely to illustrate the invention in an exemplary manner and any other purpose or function for which explained structures or configurations could be used, is covered within the scope of the present invention.

[0033] Exemplary embodiments will now be described more fully hereinafter with reference to the accompanying drawings, in which exemplary embodiments are shown. This invention may, however, be embodied in many different forms and should not be construed as limited to the embodiments set forth herein. These embodiments are provided so that this invention will be thorough and complete and will fully convey the scope of the invention to those of ordinary skill in the art. Moreover, all statements herein reciting embodiments of the invention, as well as specific examples thereof, are intended to encompass both structural and functional equivalents thereof. Additionally, it is intended that such equivalents include both currently known equivalents as well as equivalents developed in the future (i.e., any elements developed that perform the same function, regardless of structure).

[0034] Thus, for example, it will be appreciated by those of ordinary skill in the art that the diagrams, schematics, illustrations, and the like represent conceptual views or processes illustrating systems and methods embodying this invention. The functions of the various elements shown in the figures may be provided through the use of dedicated hardware as

well as hardware capable of executing associated software. Similarly, any switches shown in the figures are conceptual only. Their function may be carried out through the operation of program logic, through dedicated logic, through the interaction of program control and dedicated logic, or even manually, the particular technique being selectable by the entity
5 implementing this invention. Those of ordinary skill in the art further understand that the exemplary hardware, software, processes, methods, and/or operating systems described herein are for illustrative purposes and, thus, are not intended to be limited to any particular name.

[0035] Specific details are given in the following description to provide a thorough
10 understanding of the embodiments. However, it will be understood by one of ordinary skill in the art that the embodiments may be practised without these specific details. For example, circuits, systems, networks, processes, and other components may be shown as components in block diagram form in order not to obscure the embodiments in unnecessary detail. In other instances, well-known circuits, processes, algorithms, structures, and
15 techniques may be shown without unnecessary detail to avoid obscuring the embodiments.

[0036] FIG. 1 illustrates a process 100 of high fidelity content detection in accordance with an embodiment of the present invention. The process extracts important information from, for example, an Uniform Resource Locator (URL) of a website, and based on the webpage contents, the system processes in terms of images, texts, layout, title and date to output the
20 extracted information in text format. The process is initiated with a step 102 of receiving the URL. In an embodiment, the received URL is encoded in a predefined format. In an embodiment, the predefined format is posted in an image analysis module 202, a text analysis module 204, and a layout analysis module 206. The image analysis module 202 analyzes a plurality of images by capturing images from the received URL. The image
25 analysis module 204 captured an image of webpage of the given URL and processes by an OCR engine. In an embodiment, the OCR engine provides an array of text and transmits the array of text to the images analysis module 202.

[0037] The method then includes a step 104 of analyzing a plurality of images by capturing images from the received URL. Further, the method includes the step 106 of analyzing texts
30 by reading a plurality of HTML tags from the received URL. The method then includes the step 108 of analyzing a web-layout through Document Object Model (DOM) Application Programming Interface (API). The web-layout is divided into a plurality of

DOM element blocks. Each DOM element block is further scored by their respective height and width. In one embodiment, the element block can be scored by area size. Generally, the DOM is a cross-platform and language-independent application programming interface that treats an HTML, XHTML, or XML document as a tree structure wherein each node is an object representing a part of the document. The objects can be manipulated programmatically, and any visible changes are occurring, as a result, may then be reflected in the display of the document.

[0038] The method then includes the step 110 of obtaining a title and date of the URL through a system parser 210 (shown in FIG. 2). Furthermore, the method includes the step 112 of analyzing and generating detected relevant content in a textual format. In an embodiment, the system parser 210 retrieves the title by using DOM extraction for a plurality of <title></title> tags and retrieves the date using a regular expression on text data received from the text analysis module 204, the layout analysis module 206, and the extraction module 208.

[0039] In an embodiment, each array of text data from the text analysis module 204, the layout analysis module 206, and the extraction module 208 are compared by using an intersection of predefined sets of a matrix. Following is an example of the matrix. Each array of text data from each of the modules are compared by using an intersection of sets method in the table matrix.

	Analyze Images (I)	Analyze Text (T)	Analyze Layout (L)
Analyze Images (I)	X	$A1=I \cap T$	$A2=I \cap L$
Analyze Text (T)	$A3=T \cap I$	X	$A4=T \cap L$
Analyze Layout (L)	$A5=L \cap I$	$A6=L \cap T$	X

Table 1: Matrices of array of text data rendered by the text analysis module, the layout analysis module and the extraction module.

[0040] The final content is obtained in textual format by using the below formula:

[0041] Final Content = $A1 \cap A2 \cap A3 \cap A4 \cap A5 \cap A6$

[0042] With the above scheme, it captured the final content of the important data of the webpage in text form, whereby non-important or less-important data will be filtered out. Based on the formula, the system is able to extract the important data based on the texts that appears the most from the various modules of the present invention. These non-
5 important data include banners, advertisements, images and videos, etc.

[0043] FIG. 2 illustrates a system 200 for detecting and generating relevant content from a received Uniform Resource Locator (URL), in accordance with an embodiment of the present invention. The system 200 includes an image analysis module 202, a text analysis module 204, a layout analysis module 206, and an extraction module 210. The image
10 analysis module 202 is configured to analyze a plurality of images by capturing images from the received URL. The image analysis module 202 captures an image file of a webpage of the received URL processed by an OCR engine. In an embodiment, the OCR engine provides an array of text and transmits the array of text to the images analysis module 202.

[0044] The text analysis module 204 is configured to analyze the text by reading information about an HTML file from the received URL. The text analysis module 204 utilizes a headless browser, i.e. a web browser without Graphical User Interface (GUI), to download the HTML file from the received URL and removes a plurality of HTML tags from the HTML file. In an embodiment, the text analysis module 204 removes the HTML
15 tags from the HTML file and processes the remaining array of text.

[0045] The layout analysis module 206 is configured to analyze a web layout by dividing a plurality of document object model (DOM) element blocks by scoring, based for example, the height and width of the web layout. The layout analysis module 206 may use a JavaScript module to mark and sort each DOM element blocks by height and width. In an
20 embodiment, the layout analysis module 206 further analyses number of words in each DOM element block to determine an article content and processes an array of text blocks sorted base on size of the web layout.

[0046] The extraction module 208 is configured to utilize a system parser 210 to retrieve a title, and a date to classify an array of text. The extraction module 208 uses HTML metadata
30 received by the text analysis module 204 and generates the relevant content in a textual format. In an embodiment, the system parser 210 retrieves the title by using DOM

extraction for a plurality of `<title></title>` tags and retrieves the date using a regular expression on text data received from the text analysis module 204, the layout analysis module 206, and the extraction module 208.

[0047] Through the system 200, the URL screenshot 250 is processed and output with a texture layout 252 of the URL screenshot 250.

[0048] FIG. 3 exemplifies a process 300 of text array extraction from the URL in accordance with an embodiment of the present invention. In this example, the YAHOO! main page is used to illustrate the process 300 only, not limitation. At step 302, the URL such as `http://www.yahoo.com` is inserted into the system 200. At step 304, the URL is encoded in a unified format to conserve the URL's integrity. Base64 scheme, for example may be used, to encode the URL, and in this case, the `http://www.yahoo.com` is encoded as `aHR0cDovL3d3dy55YWVhby5JB20`. At step 306, the encoded URL is posted to the image analysis module 202, the text analysis module 204, and the layout analysis module 206 for processing. Through the image analysis module 202, the system captures an image of the URL and analyses to extract text with at least an OCR module accordingly. The text analysis module 204 is adapted to extract the textual content from the HTML through headless browser. The layout analysis module 206 extracts text blocks sorted based on the layout sizing. In one embodiment, the extraction of the layout analysis may be carried out with JavaScript. At step 308, the system receives the response from the image analysis module 202, the text analysis module 204, and the layout analysis module 206.

[0049] FIG. 4 illustrates a process for image analysis 400 in accordance with an embodiment of the present invention. At step 402, the image analysis module captures the entire webpage of the given URL as image file. The image file can be in a compressed (lossless/lossy), uncompressed, vector or compound formats. At step 404, an Optical Character Recognition (OCR) engine is utilised to process the image file. At step 406, the OCR engine extracts and returns the array of text on the capture image file of the given URL's page. Then at step 408, the system receives an array of text from image analysis module 202. In one embodiment, the image analysis may adapt readability process to filter and eliminate unwanted text.

[0050] FIG. 5 illustrates a process of text analysis 500 of in accordance with an embodiment of the present invention. At step 502, a HTML file from the targeted URL is

downloaded by using a headless browser. At step 504, HTML tags are removed from the HTML file. At step 506, the remaining array of text is obtained. At step 508, the system receives an array of text from the text analysis module 204. In the text analysis process, the meta data of the HTML file is also extracted. During the text analysis, in one
5 embodiment, readability method or process can be used for filtering out unwanted texts.

[0051] FIG. 6 illustrates a process 600 of layout analysis in accordance with an embodiment of the present invention. At step 602, the HTML file from targeted URL downloaded using the headless browser. At step 604, it marks and sorts each DOM element block by height and width by using JavaScript. At step 606, the layout analysis module
10 performs an analysis to compute a number of words in each block to determine article content. The system receives 608 an array of text blocks sorted base on layout sizing (big to small layout) from layout analysis module 206. In one embodiment, the layout sizing comprises area size.

15 [0052] FIG. 7 illustrates a process 700 of text parsing in accordance with an embodiment of the present invention. At step 702, the system 200 uses the HTML metadata obtained from the text analyzing process. In step 704, a title is extracted using DOM Extraction for <title></title> tags. At step 706, a date is extracted using a regular expression of the text block data. At step 708, an array of text classified by title and date is obtained.

20 [0053] Thus the present system and method automatically captures the important data and extracts the important data from the URL if it contains dynamic HTML5 and AJAX pages. Further, the present invention can also extract the important data when the URL contains less textual information. The present invention analyses text by reading HTML tags from the URL received from a user computing device. Additionally, the present invention
25 analyses web-layout by dividing the DOM element block by height and width for scoring. Furthermore, the present invention analyses content by returning final content in text format.

[0054] While embodiments of the present invention have been illustrated and described, it will be clear that the invention is not limited to these embodiments only. Numerous
30 modifications, changes, variations, substitutions, and equivalents will be apparent to those

skilled in the art, without departing from the scope of the invention, as described in the claims.

CLAIMS:

1. A system to detect and generate relevant content from a received Uniform Resource Locator, URL, characterized in that, the system comprising:
5 a memory to store machine-readable instructions pertaining to generation of relevant content; and
a processor coupled to the memory and operable to execute the machine-readable instructions stored in the memory, wherein the processor comprises:
an image analysis module (202) for analyzing a plurality of images by capturing images
10 from the received URL, wherein the image analysis module (202) captures a webpage of the received URL in a form of image file for processing by an Optical Character Recognition, OCR, engine;
a text analysis module (204) to analyze the text by reading information pertaining to an Hypertext Markup Language, HTML, file from the received URL, wherein the text
15 analysis module (204) utilizes a headless browser to download the HTML file from the received URL and removes a plurality of HTML tags from the HTML file;
a layout analysis module (206) to analyze a web layout by dividing a plurality of document object model, DOM, element blocks by scoring height and width of the web-layout, wherein the layout analysis module (206) uses a JavaScript module to mark and
20 sort each DOM element blocks by height and width; and
an extraction module (208) utilizes a system parser (210) to retrieve a title, and a date to classify an array of text, wherein the extraction module (208) uses HTML metadata received by the text analysis module (204) and generates the relevant content in a
textual format.
25
2. The system according to claim 1, wherein the OCR engine provides an array of text and transmits the array of text to the text analysis module (204).
3. The system according to claim 1, wherein the text analysis module (204) removes the
30 HTML tags from the HTML file and processes the remaining array of text.
4. The system according to claim 1, wherein the layout analysis module (206) further analyses number of words in each DOM element block to determine an article content and processes an array of text blocks sorted base on size of the web layout.

5. The system according to claim 1, wherein the system parser (210) retrieves the title by using DOM extraction for a plurality of <title></title> tags and retrieves the date using a regular expression on text data received from the text analysis module (204), the layout analysis module (206), and the extraction module (208).
6. A method for detecting and generating relevant content from a received Uniform Resource Locator, URL, characterized in that the method comprising steps of:
- receiving, by an image analysis (202), a URL;
- analyzing a plurality of images by capturing images from the received URL, wherein a webpage is captured in a form of image file from the received URL and process by an Optical Character Recognition, OCR, engine;
- analyzing, by a text analysis module (204), texts by reading information pertaining to an Hypertext Markup Language, HTML, file from the received URL, wherein the text analysis module utilizing a headless browser to download the HTML file and remove HTML tags from the HTML file;
- analyzing, by a layout analysis module (206), a web-layout by dividing a plurality of document object model, DOM, element blocks by scoring height and width of the web layout;
- retrieving by an extraction module (208), a title and a date of the URL through a system parser (210), wherein the extraction module (208) uses HTML metadata received by the text analysis module (204); and
- analyzing and generating detected relevant content in a textual format.
7. The method according to claim 6, wherein the OCR engine provides an array of text and transmitting the array of text to the text analysis module (204).
8. The method according to claim 6, further comprising removing the HTML tags from the HTML file and processing the remaining array of text by the text analysis module (204).
9. The method according to claim 6, wherein the system parser (210) retrieves the title by using DOM extraction for a plurality of <title></title> tags and retrieves the date using a regular expression on text data received from the text analysis module (204), the layout analysis module (206), and the extraction module (208).

10. The method according to claim 9, wherein each array of text data from the text analysis module (204), the layout analysis module (206), and the extraction module (208) are compared by using an intersection of a plurality of predefined sets of a matrix.

1/7

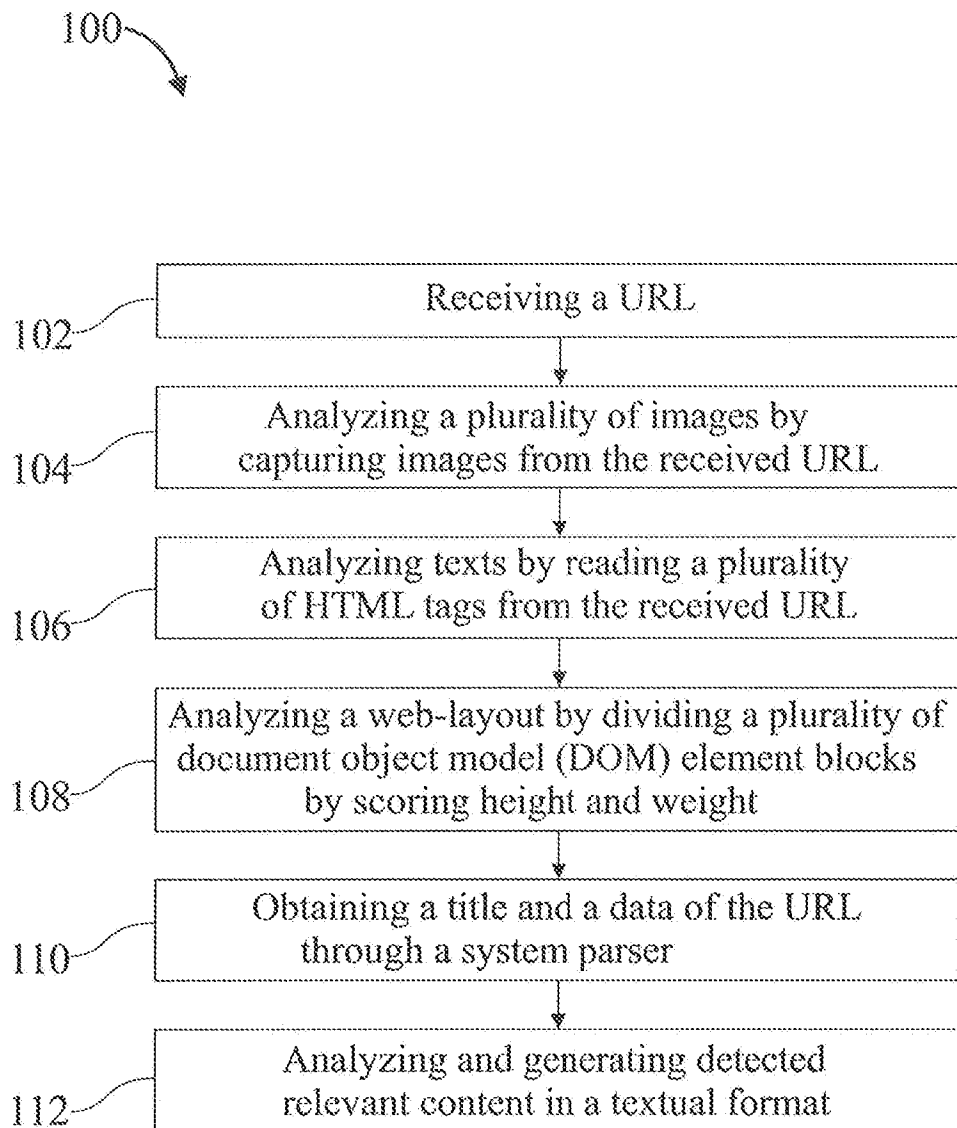
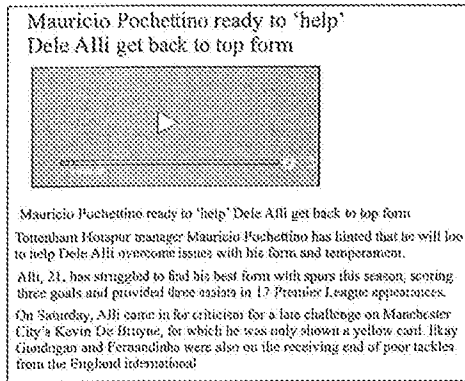


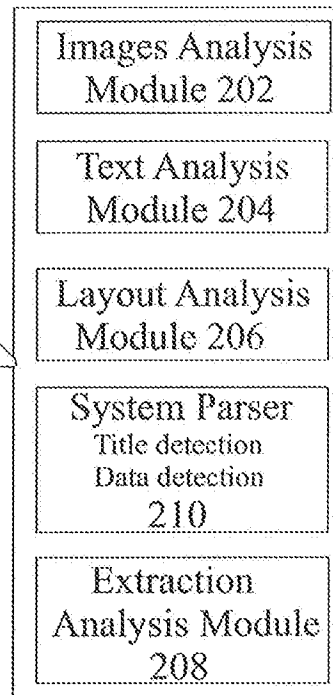
FIG. 1

2/7

200



250



Mauricio Pochettino ready to 'help' Dele Alli get back to top form

Tottenham Hotspur manager Mauricio Pochettino has hinted that he will look to help Dele Alli overcome issues with his form and temperament.

Alli, 21, has struggled to find his best form with Spurs this season, scoring three goals and providing three assists in 17 Premier League appearances.

On Saturday, Alli came in for criticism for a late challenge on Manchester City's Kevin De Bruyne, for which he was only shown a yellow card. Ilkay Gundogan and Fernandinho were also on the receiving end of poor tackles from the England international.

Pochettino would not go into specifics about his players, but suggested he would sit down with the player to sort out any issues.

"I am not going now to talk about individual players," Pochettino was quoted as saying by the Daily Telegraph. "You can assess, people can assess. I'm going to say what I need to do and try to help him and help everyone."

252

FIG. 2

3/7

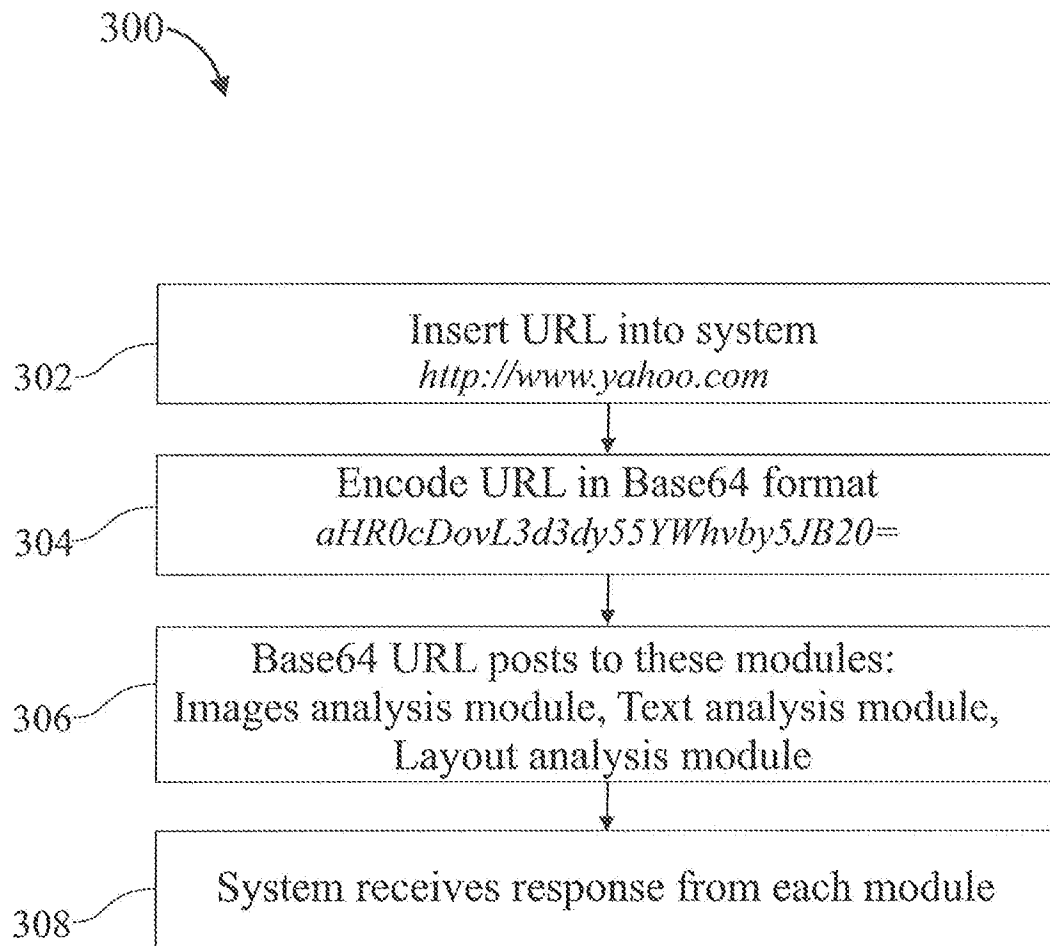


FIG. 3

4/7

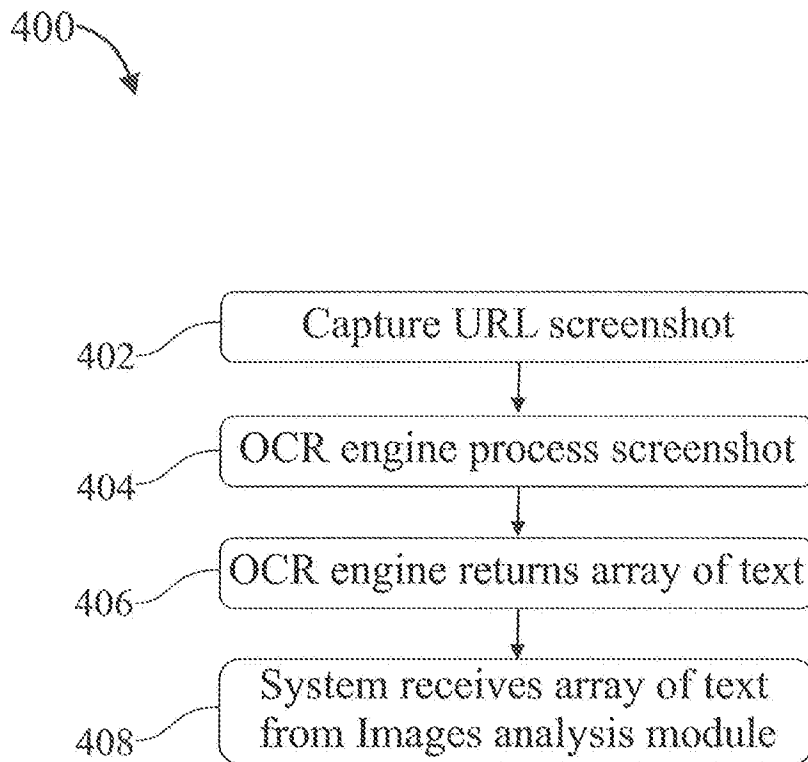


FIG. 4

5/7

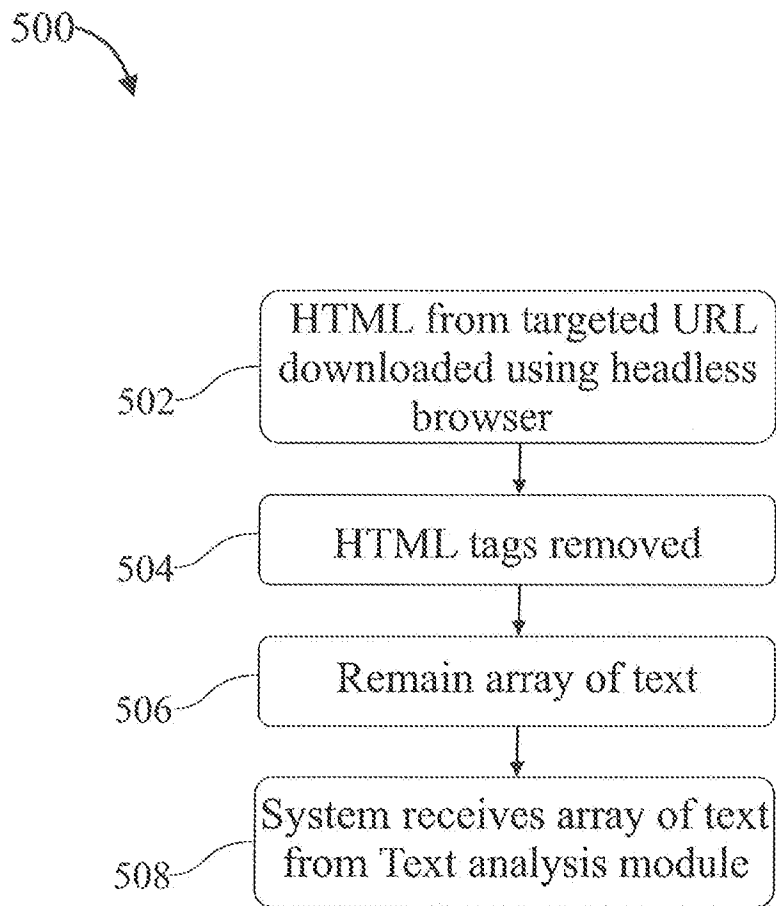


FIG. 5

6/7

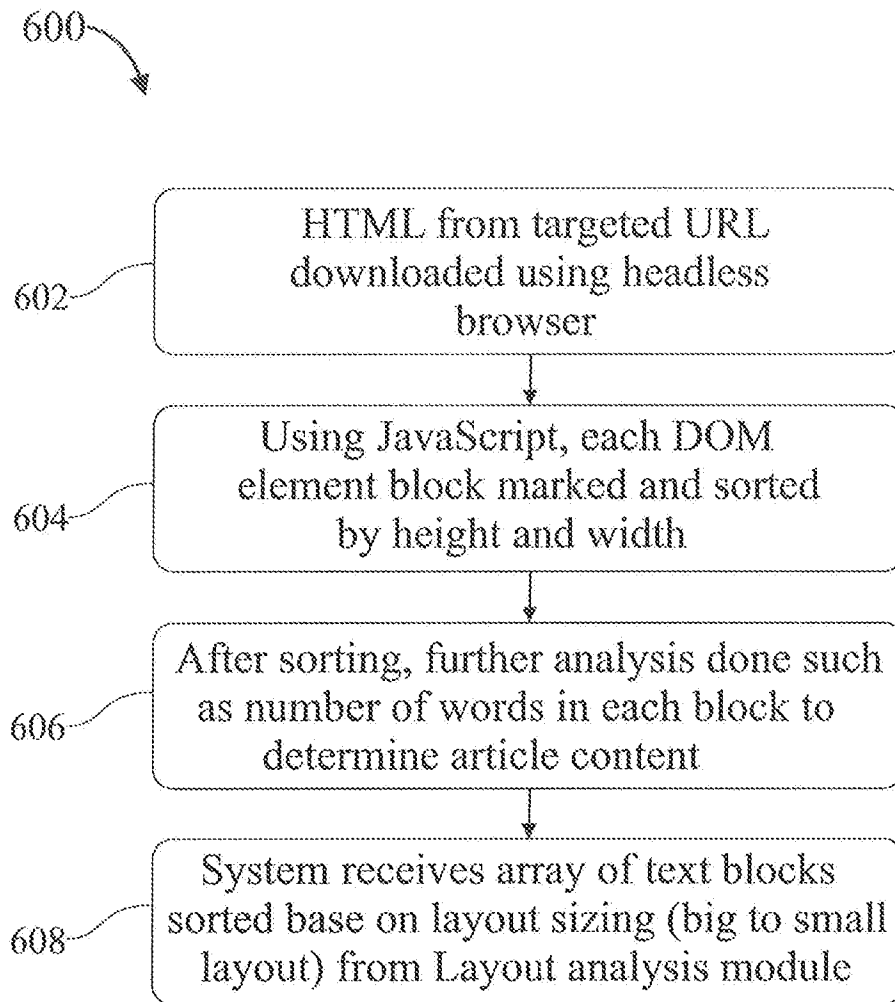


FIG. 6

7/7

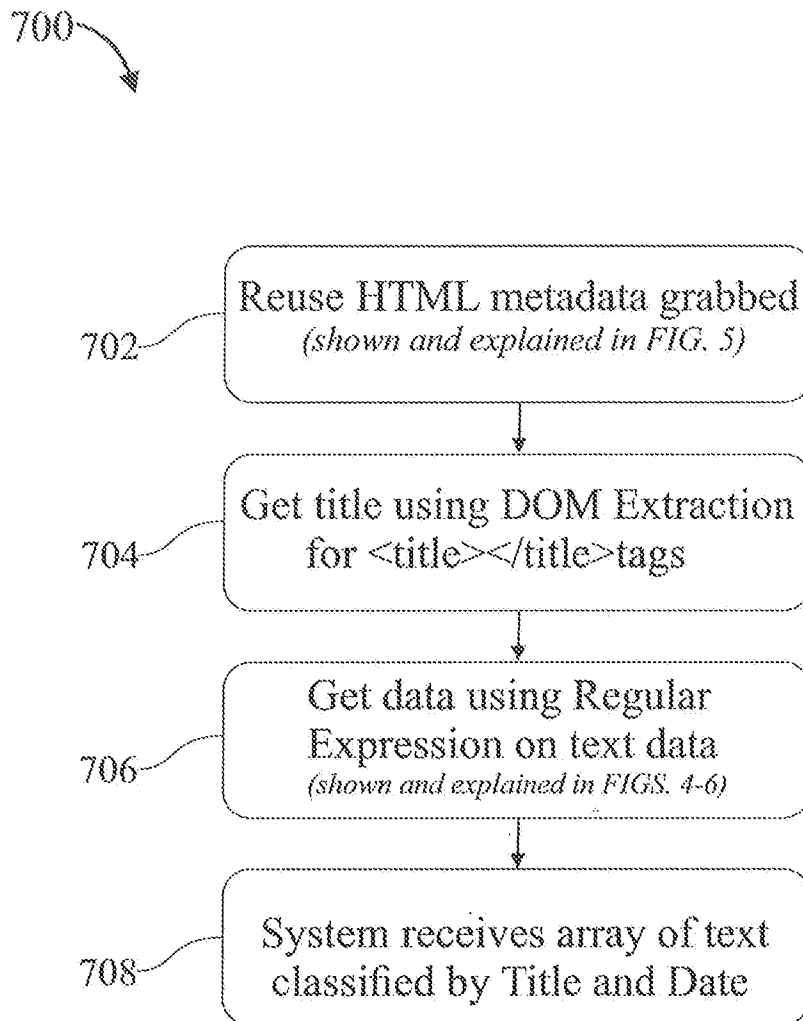


FIG. 7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/MY2019/050094

A. CLASSIFICATION OF SUBJECT MATTER**G06F 16/81**(2019.01)i

FI: G06F16/81

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F16/81

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Published examined utility model applications of Japan 1922-1996

Published unexamined utility model applications of Japan 1971-2020

Registered utility model specifications of Japan 1996-2020

Published registered utility model applications of Japan 1994-2020

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	JP 2012-118737 A (FUJITSU LIMITED) 21 June 2012 (2012-06-21) [0015]-[0040], figures 1, 3	1-10
Y	JP 11-120202 A (RICOH COMPANY, LTD.) 30 April 1999 (1999-04-30) [0076]-[0079]	1-10
A	JP 2011-123740 A (FUJIFILM CORPORATION) 23 June 2011 (2011-06-23) the whole document, figures 1, 5	1-10
A	WO 2015/122620 A1 (MWSTORY CO.,LTD) 20 August 2015 (2015-08-20) the whole document	1-10

☐ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

28 February 2020

Date of mailing of the international search report

10 March 2020

Name and mailing address of the ISA/JP

Japan Patent Office
3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo
100-8915, Japan

Authorized officer

NISHIMURA, Naofumi 5N 5088

Telephone No. +81-3-3581-1101 Ext. 3586

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/MY2019/050094

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
JP	2012-118737	A	21 June 2012	(Family: none)	
JP	11-120202	A	30 April 1999	(Family: none)	
JP	2011-123740	A	23 June 2011	US 2011/0142344 A1 the whole document, figures 1, 5	
WO	2015/122620	A1	20 August 2015	JP 2017-504858 A the whole document KR 10-2015-0096211 A the whole document	