



(51) International Patent Classification:
G06F 16/34 (2019.01)

(21) International Application Number:
PCT/US2019/037413

(22) International Filing Date:
17 June 2019 (17.06.2019)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
201810714779.1 29 June 2018 (29.06.2018) CN

(71) Applicant: **MICROSOFT TECHNOLOGY LICENSING, LLC** [US/US]; One Microsoft Way, Redmond, Washington 98052-6399 (US).

(72) Inventors: **ZHANG, Xingxing**; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). **HUANG, Shaohan**; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). **CUI, Lei**; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Wash-

ington 98052-6399 (US). **GE, Tao**; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). **WEI, Furru**; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). **ZHOU, Ming**; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US).

(74) Agent: **MINHAS, Sandip S.** et al.; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US).

(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, 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.

(54) Title: GENERATING DOCUMENT SUMMARY

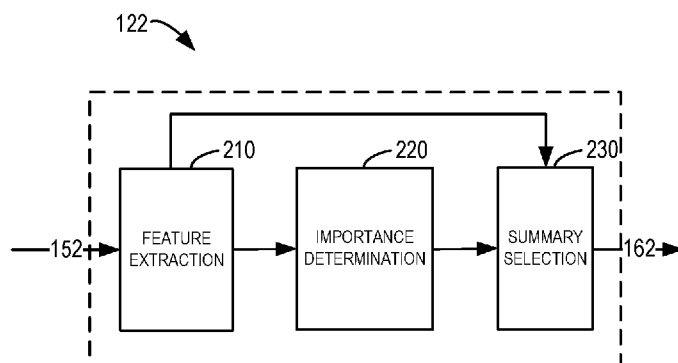


FIG. 2

(57) Abstract: According to implementations of the subject matter described herein, there is provided a solution for generating a summary of a document. In this solution, feature information of pages comprised in a document is extracted, which characterizes at least one type of content contained in each page. Respective importance of the pages is determined at least based on the extracted feature information. A summary of the document is generated for the document by selecting a predetermined number of pages less than the number of the pages based on the respective importance. Through the solution, instead of providing all the pages, pages containing important content may be determined automatically to serve as the summary of the document. This summary allows the user to learn quickly main content of the document, shorten the time consumed in browsing all documents and/or facilitate location of a document of interest as soon as possible.

(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:

- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
- *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

- *with international search report (Art. 21(3))*

GENERATING DOCUMENT SUMMARY

BACKGROUND

[0001] People usually need to search for or review various documents such as presentation slide documents, word processing documents, and the like in activities such as routine work, studies, conferences, and the like. Sometimes one document may contain a lot of pages. The user has to spend massive time to read through all the pages to learn about main content in the document. If there are a large number of documents to be read, more time will be taken. For example, in a scenario of performing search, a search engine returns a great number of documents as search results; and the user may need to browse all pages of each of the documents to determine whether that document includes the desired content. Such inefficient document browsing cannot provide fast and convenient localization of documents and may cause unnecessary time wasting for the user.

SUMMARY

[0002] In accordance with implementations of the subject matter described herein, there is provided a solution for generating a summary of a document. In the solution, feature information of a plurality of pages comprised in a document is extracted, and the feature information of each page characterizes at least one type of content contained in the page. Respective importance of the plurality of pages in the document is determined at least based on the extracted feature information. A summary is generated for the document by selecting a predetermined number of pages from the plurality of pages based on the respective importance, and the predetermined number is less than the number of the plurality of pages. Through the solution, instead of providing all the pages, pages containing important content may be determined automatically from a plurality of pages of the document to serve as the summary of the document. This summary of the document allows the user to learn quickly main content of the document, shorten the time consumed by the user in browsing all documents, and/or facilitate location of a document of interest as soon as possible.

[0003] This Summary is provided to introduce a selection of concepts in a simplified form that are further described below in the Detailed Description. This Summary is not intended to identify key features or essential features of the claimed subject matter, nor is it intended to be used to limit the scope of the claimed subject matter.

BRIEF DESCRIPTION OF THE DRAWINGS

[0004] Fig. 1 illustrates a block diagram of a computing environment in which a plurality of implementations of the subject matter described herein can be implemented;

5 [0005] Fig. 2 illustrates a block diagram of a summary generating module in accordance with an implementation of the subject matter described herein;

[0006] Fig. 3 illustrates a schematic diagram of an implementation of feature extraction and importance determination based on a neural network in accordance with an implementation of the subject matter described herein;

10 [0007] Fig. 4 illustrates a schematic diagram of feature extraction for a page in accordance with an implementation of the subject matter described herein;

[0008] Figs. 5A-5C illustrate schematic diagrams of example user interfaces for presenting a summary of a document in accordance with some implementations of the subject matter described herein; and

15 [0009] Fig. 6 illustrates a flowchart of a process for generating a summary of a document in accordance with an implementation of the subject matter described herein.

[0010] Throughout the drawings, the same or similar reference symbols refer to the same or similar elements.

DETAILED DESCRIPTION OF EMBODIMENTS

20 [0011] The subject matter described herein will now be discussed with reference to several example implementations. It is to be understood these implementations are discussed only for the purpose of enabling those skilled persons in the art to better understand and thus implement the subject matter described herein described herein, rather than suggesting any limitations on the scope of the subject matter.

25 [0012] As used herein, the term “includes” and its variants are to be read as open terms that mean “includes, but is not limited to.” The term “based on” is to be read as “based at least in part on.” The term “an implementation” and “an implementation” are to be read as “at least an implementation.” The term “another implementation” is to be read as “at least one other implementation.” The terms “first,” “second,” and the like may refer to different or same objects. Other definitions, either explicit or implicit, may be included
30 below.

[0013] As mentioned above, when a document includes a plurality of pages, a user needs to spend a lot of time to browse all pages for a document to grasp important contents therefrom. Therefore, it is necessary to generate a summary for a document for presenting main content of the document. To this end, there is proposed a solution for

generating a summary of a document in implementations of the subject matter described herein. The solution involves extracting feature information of a plurality of pages of the document, and the extracted feature information characterizes at least one type of content contained in respective pages. Respective importance of a plurality of pages in the document is determined based on the extracted feature information, and some of the plurality of pages is selected based on the determined importance for generating the summary of the document. By feature extraction and importance analysis based on the features, pages including important content may be determined automatically from the plurality of pages of the document, instead of all pages serving as the summary of the document. The summary of the document allows a user to obtain main content of the document quickly, shorten the time consumed by the user in browsing all documents, and/or facilitate location of a document of interest as soon as possible so as to determine a next operation to be performed on the certain document (for example, browsing, downloading, forwarding, and/or the like).

[0014] Basic principles and various example implementations of the subject matter described herein will now be described with reference to the drawings. Fig. 1 illustrates a block diagram of a computing device 100 in which implementations of the subject matter described herein can be implemented. It would be appreciated that the computing device 100 described in Fig. 1 is merely for purpose of illustration, without suggesting any limitation to the functions and scopes of implementations of the subject matter described herein in any manner. As shown in Fig. 1, the computing device 100 includes a computing device 100 in a form of general-purpose computing device. Components of the computing device 100 may include, but are not limited to, one or more processors or processing units 110, a memory 120, a storage device 130, one or more communication units 140, one or more input devices 150, and one or more output devices 160.

[0015] In some implementations, the computing device 100 may be implemented as various user terminals or service terminals. The service terminal may be any server, large-scale computing device or the like that is provided by various service providers. The user terminal may for example be any type of mobile terminal, fixed terminal, or portable terminal, including a mobile telephone, station, unit, device, multimedia computer, multimedia tablet, Internet node, communicator, desktop computer, laptop computer, notebook computer, netbook computer, tablet computer, personal communication system (PCS) device, personal navigation device, personal digital assistant (PDA), audio/video player, digital camera/video camera, positioning device, television receiver, radio

broadcast receiver, E-book device, gaming device, or any combinations thereof, including accessories and peripherals of these devices, or any combinations thereof. It would be appreciated that the computing device 100 can support any type of interface for a user (such as a “wearable” circuit and the like).

5 **[0016]** The processing unit 110 may be a physical or virtual processor and can perform various processes based on programs stored in the memory 120. In a multi-processor system, multiple processing units execute computer-executable instructions in parallel so as to improve the parallel processing capacity of the computing device 100. The processing unit 110 may also be referred to as a central processing unit (CPU), a
10 microprocessor, a controller or a microcontroller.

[0017] The computing device 100 typically includes a plurality of computer storage medium. Such medium may be any available medium accessible by the computing device 100, including, but not limited to, volatile and non-volatile medium, and removable and non-removable medium. The memory 120 may be a volatile memory (for example,
15 a register, cache, Random Access Memory (RAM)), non-volatile memory (for example, a Read-Only Memory (ROM), Electrically Erasable Programmable Read-Only Memory (EEPROM), flash memory), or any combination thereof. The storage device 130 may be any removable or non-removable medium and may include machine-readable medium, such as a RAM, flash drive, disk or any other medium, which can be used for storing
20 information and/or data and accessed in the computing device 100.

[0018] The computing device 100 may further include additional removable/non-removable, volatile/non-volatile memory medium. Although not shown in Fig. 1, a disk drive is provided for reading and writing a removable and non-volatile disk and a disc drive is provided for reading and writing a removable non-volatile disc.
25 In these cases, each drive is connected to the bus (not shown) via one or more data medium interfaces.

[0019] The communication unit 140 communicates with a further computing device via communication medium. Additionally, functions of components in the computing device 100 may be implemented by a single computing cluster or multiple computing machines
30 that can communicate via communication connections. Therefore, the computing device 100 may operate in a networked environment using a logical connection with one or more other servers, network personal computers (PCs) or further general network nodes.

[0020] The input device 150 may include one or more of a variety of input devices, such as a mouse, a keyboard, a tracking ball, a voice-input device, and the like. The output

device 160 may be one or more of a variety of output devices, such as a display, a loudspeaker, a printer and the like. By means of the communication unit 140, the computing device 100 may further communicate with one or more external devices (not shown) such as the storage device, display device and the like, with one or more devices
5 that enable users to interact with the computing device 100, or any devices (for example, a network card, modem, and the like) that enable the computing device 100 to communicate with one or more other computing devices, if required. Such communication may be performed via an input/output (I/O) interface (not shown).

[0021] In some implementations, as an alternative of being integrated on a single device,
10 some or all components of the computing device 100 may also be arranged in form of cloud computing architecture. In the cloud computing architecture, these components may be provided remotely, and may work together to implement the functions described herein. In some implementations, the cloud computing provides computing, software, data access and storage service, which will not require end users to be aware of the
15 physical location or configuration of the systems or hardware that provides these services. In various implementations, the cloud computing provides the services via a wide area network (for example, Internet) using an appropriate protocol. For example, a cloud computing provider provides applications over a wide area network, which can be accessed through a web browser or any other computing components. Software or
20 components of the cloud computing architecture and respective data may be stored on a server at a remote location. Computing resources in a cloud computing environment may be merged or distributed at a location in a remote data center. Cloud computing infrastructures may provide the services through a shared data center, though they behave as a single access point for the user. Hence, the cloud computing architecture may be
25 used to provide the components and functionalities as described herein from a service provider at a remote location. Alternatively, they may be provided from conventional servers or installed directly or otherwise on a client device.

[0022] The computing device 100 may be used to implement generation of document summary in a plurality of implementations of the subject matter described herein. The
30 memory 120 may include one or more modules, each of which includes one or more program instructions, and these modules can be accessed and operated by the processing unit 110, so as to implement functionalities of various implementations described herein. For example, the memory 120 may include a summary generating module 122 for performing generating a summary of a document.

[0023] When generating the summary is performed, the computing device 100 can receive a document 152 via an input device 150. The document 152 includes a plurality of pages, for example, page 102-1 to page 102-9, which are collectively referred to as a page 102 herein. The document 152 is provided to a summary generating module 122, to
5 cause the summary generating module 122 to select some pages 102 from the document 152 as a summary 162 of the document 152 output by the output device 160. In the example as shown in Fig. 1, the summary generating module 122 selects the page 102-1, page 102-3, page 102-4, and page 102-8 from the document 152 as a summary 162 of the document 152 for output.

[0024] It would be appreciated that the form of the page, the number of pages included in the document, and the number of the pages included in the summary as shown in Fig. 1 are provided merely for purpose of illustration. In other examples, a document may include any number of pages containing any content, and the generated summary may also include any number of pages. Pages selected as a summary from a document will also
15 vary depending on the technique used by the summary generating module.

[0025] How a summary 162 of a document 152 is generated by a summary generating module 122 will be described below in detail with reference to Figs. 2 to 5C. Fig. 2 illustrates a summary generating module 122 for generating a summary of a document in accordance with some implementations of the subject matter described herein. As shown,
20 the summary generating module 122 includes a feature extraction module 210, an importance determination module 220 and a summary selection module 230.

[0026] The feature extraction module 210 is configured to receive the document 152 including a plurality of pages 102. The feature extraction module 210 is configured to extract feature information of a plurality of pages, and the feature information of each
25 page characterizes at least one type of content contained in the page. The features extracted by the feature extraction module 210 are provided to the importance determination module 220 which is configured to determine respective importance of the plurality of pages 102 in the document 152 based on the plurality of extracted features. The importance of a page indicates whether the page includes main or important content
30 of the document 152.

[0027] The summary selection module 230 is configured to select a predetermined number of pages (for example, K pages) from the plurality of pages 102 of the document 152 based on the importance determined by the importance determination module 220, to generate a summary for the document. The predetermined number of pages selected is

less than the number of pages included in the document 152. It is supposed that the number of pages included in the document 152 is N (N is an integer greater than 1) and the predetermined number of pages included in the summary is K , then $K < N$. The predetermine number may be preconfigured by a user or system and is fixed for a plurality of documents. Alternatively, the predetermined number may vary dynamically, for example, vary with the values of the determined importance or the total number of pages included in the document. For example, one or more pages with their importance greater than a certain threshold importance may be selected from the document 152 for generating the summary 162. For another example, if the number of pages included in a document 152 is relatively large, a larger predetermined number may be determined, or vice versa.

[0028] In accordance with the implementations of the subject matter described herein, the generated summary includes a smaller number of pages, facilitating a user to browse important content of the document quickly so as to make a decision of a next step, for example, whether to browse all pages, download the document, forward the document, utilize the document for other purposes, and/or the like.

[0029] In some implementations, the document 152 may include various formats of document, such as presentation slide documents, word processing documents, webpages, printable document format (PDF) documents and the like. For these types of documents, corresponding applications (for example, presentation applications, word processing applications, webpage browsing application, PDF readers and the like) may be used to present pages 102 included in the document. The pages 102 may include different types of content. For example, a page of a presentation slide document may typically include a combination of any of delicately arranged texts, images (such as inserted pictures, clipart, icons, background images, charts, and the like), and inserted multimedia (such as video, audio, and the like). A document in other format may also include one type or a combination of different types of contents. For example, in a word processing document, some pages may include a combination of words with pictures, symbols and charts, in addition to pages only including text, to enrich presentation of content of the pages.

[0030] Typically, in one document, whether a page is important is related to the content included in the page. In order to identify important pages from a plurality of pages, feature information of each page may be determined in a feature extracting manner for characterizing the content included therein. The extracted feature information may be used to measure whether each page is important in the page. In some implementations, the feature extraction module 210 may be configured to extract different feature

information of the page based on different types of content included in each page 102.

[0031] In an implementation, if it is determined that the content contained in the page 102 includes a text, the semantic meaning and the length of the text may be used to characterize the content of the page 102. Therefore, the feature extraction module 210
5 may be configured to extract feature information related to the text of the page 102, including one or more of the semantic feature of the text and feature information indicating the number of words included in the text. Fig. 3 illustrates an example of a page 102, in which the page 102 includes a text 310 including a title in the form of text and a body in the form of text. The feature extraction module 210 may be configured to
10 extract a semantic feature of the text 310 and the number of words included in the text 310.

[0032] In an implementation, if it is determined that one or more images are included in the page 102, the locations of the images in the page, the occupied areas of the images in the page (which are related to the sizes and the number of the images), and the features of
15 the image contents included therein may be used to characterize the content of the page 102. Therefore, the feature extraction module 210 may be configured to extract one or more of feature information related to images of the page 102, including a feature map of the images, feature information indicating the occupied area of the images in the page, and feature information indicating the locations of the images in the page. For example, in
20 the example of Fig. 3, the page 102 includes an image portion 320 which includes two images. The feature extraction module 210 may be configured to extract a feature map of these images, extracting a total occupied area of these images in the page and/or an area of each image (which may be indicated by the length and the width of the image), and/or respective locations of these images in the page 102 (which are indicated by the
25 two-dimensional coordinates, for example).

[0033] In some implementations, a page 102 may include a mixture of a plurality of types of contents or a plurality of content sections of the same type, and these contents or content sections may be arranged in the page 102 in a certain particular manner (which is particularly embodied in presentation slides or pages). In order to explore feature
30 information included in organization of different contents or content sections, the page 102 may be converted into an image, and a feature map of the converted image are extracted by the feature extraction module 210 as hybrid feature information of the page 102. In the example of Fig. 3, the page 102 may be considered as an image, and a feature map of this image may be extracted by the feature extraction module 210.

[0034] Different feature information (for example, text-related feature information, image-related feature information and/or hybrid feature information) extracted for the page 102 may be combined as total feature information of the page 102 for determining the importance of the page at the importance determination module 220. It would be appreciated that other types of feature information of the page 102 may also be extracted.

[0035] In some implementation, extraction of feature information and measurement of an importance may be implemented based on a neural network. Generally, the neural network is constructed to include one or more network layers, each of which includes one or more neurons. Each neuron processes an input according to a parameter obtained through training, and generates an output. The output of the preceding layer is transmitted to the next layer as an input for further processing. Parameters of all the neurons of the neural network form a set of parameters of the neural network. When the set of parameters of the neural network is determined, the neural network may be operated to perform respective functions. As used herein, the neural network may also be referred as a “learning network” or a “neural network model.” The terms “learning network,” “neural network,” “neural network model,” “model” and “network” may be used interchangeably below.

[0036] Fig. 4 illustrates an example of feature extraction and importance determination based on a neural network 400 in accordance with an implementation of the subject matter described herein. The neural network 400 is used to process each page 102 of the document 152. As shown in Fig. 4, the neural network 400 includes a plurality of sub-networks 412, 414 and 416 for implementing the feature extraction module 220, and these neural sub-networks are provided for text-related feature information, image-related feature information and hybrid feature information of each page 102, respectively.

[0037] In some implementations, when the content contained in the page 102 includes a text, the sub-network 412 may be configured to extract a semantic feature of the text and/or feature information indicating the number of words included in the text, as feature information related to the text of the page 102. When extracting the semantic feature, the sub-network 412 may implement semantic feature extraction of the text using a plurality of natural language processing methods. In some implementations, the sub-network 412 may extract a vectorized representation of each word of the text, and vectorized representations of all words are combined as the semantic feature of the text. A vectorized representation of a word may also be referred to as vector code of the word. The vectorized representation of a word may be comprised of a plurality of numerical

values included therein. The dimension of vectorized representation of different words may be the same, but the numerical value included therein may be different. A similarity between vectorized representations of words may also indicate a semantic similarity between different words. In other words, there is a small difference between vectorized representations of similar words, while there is a great difference between vectorized representations of words that are completely unrelated.

[0038] In some implementations, a vectorized representation of a word may be selected, for example, from a library including pre-trained word vectorized representations, and such library of word vectorized representations may include, for example, Glove vectors, Word2Vec vectors, FastText vectors and the like, of different words. In some other implementations, the sub-network 412 may be trained using different words included in a plurality of documents, such that the sub-network 412 can determine a respective vectorized representation for each word, to indicate a semantic similarity between these words.

[0039] In some implementations, all words included in the text of the page 102 are extracted as a sequence words. A symbol visually missing from the text of the page 102, such as a line break, may be replaced with a special token and be included in the sequence of words, facilitating representing an interval between different lines in a text. Since the text of the page 102 may include lots of words and there may be different numbers of words included in texts of different pages 102, in order to enable semantic features of texts of different pages to be comparable, vectorized representations of words may undergo dimension reduction process, and the vectorized representations may be averaged across different words. Specifically, it is supposed that the sequence of words of the text of the page 102 is represented as $S = (w_1, w_2, \dots, w_n)$, where n is a length of the sequence S and represents the number of words (including the substitution of particular tokens) in the sequence. It is supposed that w_i represents the vectorized representation of the i^{th} word, and w_i represents a one-hot vector of the vectorized representation of the i^{th} word for dimension reduction. The semantic feature (which is denoted as s) of the text of the page 102 may be represented as an average of vectorized representations of different words in the text, which may be expressed as follows:

$$s = \frac{1}{n} \sum_i w_i \in (w_i) \quad (1)$$

[0040] In some implementations, the sub-network 412 may also be configured to extract

feature information indicating the number of words contained in the text of the page 102. For example, the number n of words contained in the text of the page 102 may be mapped into a vectorized representation $\mathbf{w}_t$. The sub-network 412 may concatenate the vectorized representation of the text of the page 102 and the vectorized representation of the number of words contained in the text, as text-related feature information (which is represented as $\mathbf{v}$) of the page 102, which may be expressed, for example, as follows:

$$\mathbf{v} = \begin{bmatrix} \mathbf{s} \\ \mathbf{w}_t n \end{bmatrix} \quad (2)$$

[0041] Still referring to Fig. 4, when the content contained in the page 102 includes an image, the sub-network 414 may be configured to extract image-related feature information of each page 102, such as a feature map of the image, feature information indicating the occupied area of the image in the page and feature information indicating the location of the image in the page, as mentioned above. When extracting a feature map of the image, the sub-network 414 may include various neural network layers for image processing, to extract, from the image, feature information characterizing various aspects (for example, image color, texture, profile, edge and the like) of the image. For example, the sub-network 414 may include a convolutional neural network (CNN), long short-term memory (LSTM) neural network and the like, which perform well in the image processing field. In some implementations, the sub-network 414 may map the area of each image (or the total image area) in the page 102 into a vectorized representation, or may map the location of each image in the page 102 (for example, two-dimensional coordinates of a center pixel of the page 102) into a vectorized representation. The feature map, area and/or location-related feature information of the image may be concatenated as image-related feature information of the page 102.

[0042] Further, the sub-network 416 is configured to extract a feature map of the image as hybrid feature information of the page 102 in a case of converting the page 102 into an image. In some implementations, regardless of the content contained in the sub-network 416 including a text, an image, other types of contents or a mixture thereof, the sub-network 416 can extract hybrid feature information of the page 102. The method of extracting a feature map of the whole image of the page 102 may be similar to the method of extracting a feature map of the image contained in the page 102. For example, the sub-network 416 may include neural networks, such as CNN, LSTM and the like, for extracting a feature map. In some implementations, the sub-network 414 may only extract feature information related to the area and/or location related to the image

contained in the page 102, and a feature map of the whole page 102 may be extracted by the sub-network 416.

[0043] In some implementations, in order to extract a feature map of the image, the sub-network 414 and/or 416 may use a network portion for executing feature extraction in a pre-trained neural network for image processing. For example, a network layer for performing feature extraction in a neural network for image classifying tasks having been trained in advance may serve as a sub-network 414 or 416.

[0044] The feature information extracted by sub-networks 412, 414 and/or 416 may be concatenated and provided to the importance determination module 220. The importance determination module 220 may include one or more network layers 422 for further processing various feature information after being concatenated. The network layer 422 may include, for example, a fully connected layer. Typically, the fully connected layer includes activation neurons with a linear rectified function (ReLU). In some examples, the network layer 422 has a drop function, i.e., during processing process, the network layer 422 drops some neurons temporarily according to a certain probability (i.e., even though these neurons are not involved in processing). The importance determination module 220 may further include a regression unit 424 for determining an importance based on the processed feature information. The regression unit 424 may map the processed feature information into an importance based on various regression functions. Depending on the designed regression function, the feature information may be mapped to a scalar value for indicating the importance.

[0045] In some implementations, the sub-networks 412, 414 and/or 416 for feature extraction may employ neural networks for text feature extraction and image feature extraction, which have been trained in advance. In order to further optimize feature extraction of these sub-networks, in some implementations, the pre-trained sub-networks 412, 414 and/or 416 may also be trained together with the network layers 422, 424 for importance determination, so as to further optimize these sub-networks and parameters of the network layers, to enable a cooperation there between for accomplishing the purpose of importance determination. Of course, the sub-networks 412, 414 and/or 416 may also be initialized as random values and trained together with the network layers 422, 424.

[0046] During a training process, one or more training documents may be used to enable neural networks in the feature extraction module 210 and the importance determination module 220 to learn the capabilities of feature extraction and importance determination. Each training document may include a plurality of training pages, and a predetermined

number of training pages of the training pages are labeled as the summary of the training document. Monitored with the true value information, the parameters of the neural networks in the extraction module 210 and the importance determination module 220 may be adjusted constantly until a target of optimization is fulfilled.

5 **[0047]** As the summary of the training document, the predetermined number of training pages may be labeled in various manners. One possible manner is a manual label related to whether each training page is used as a summary of a corresponding training document, i.e., the document is reviewed by human being and the following are labeled: whether each training page is important for the training document, how important it is or whether it can
10 serve as the summary of the document. In order to reduce the expensive cost resulting from the manual label, whether each training page is able to serve as the summary of the training document may be labeled automatically, in an implementation. Typically, if a page is important for a document, the page will be presented for a longer time during a process of presenting the entire document. For example, during a process of presenting a
15 presentation slide document, if one slide is of great importance, the speaker will spend more time to present the slide, thus the slide will be presented for a longer time. When browsing a document, such as a presentation slide document, word processing document, webpage and the like, a user probably lingers at an important page of that the user is interested in for a longer time before switching to the next page.

20 **[0048]** Therefore, in order to obtain the summary of the training document, whether the training page is labeled as a summary may be determined based on duration of presenting each training page during a process of presenting the training document. For example, for each training document, a predetermined number of pages being presented for longer duration in the document may be labeled as the summary of the training document.

25 **[0049]** In some other implementations, for some types of documents, such as presentation slide documents, the documents may include some remark information. The remark information is information that is used for further interpreting content of a page in during the process of editing the documents. If there is such information, it may be regarded that the page is important and requires more information to interpret. Hence,
30 when determining whether a training page may be identified as a summary, it may be performed based on whether the page includes the remark information. In some implementations, the presence of the remark information in the training page may be used in combination with the duration of presenting the training page to determine whether the page can be labeled as the summary of the training document.

[0050] How to extract feature information of each page 102 and how to determine an importance based on the feature information have been discussed above. As mentioned above, after respective importance of all pages 102 of the document 152 are determined, the summary selection module 230 selects a predetermined number of pages based on the importance, and then generates a summary 162 of the document 152 from the selected pages. Specifically, the summary selection module 230 may rank the pages 102 of the document 152 based on the importance, and then selects a predetermined number of pages ranked at the top.

[0051] In some implementations, in addition to the importance determined based on the feature information, the summary selection module 230 may determine whether a page 102 may be selected to generate the summary 162 by further considering other information. In an implementation, it may be determined whether a plurality of pages 102 includes attention information. The attention information may be a type of information indicating the probability of important information of the document is high.

[0052] For example, an opening or ending portion of a series of pages 102 probably includes pages about an overview or summary of the document. The pages may include some keywords for indicating overview or summary information involved in the content contained in the document, and such information may be viewed as attention information. Consequently, in some implementations, the summary selection module 230 may determine whether the pages 102 include a keyword which may be, for example, "overview," "summary," or the like. A selection of keyword may be set based on different languages. In some implementations, whether the keyword is included may be determined only through the title portion of the text included in the page 102, such as a header line, a title of a chapter and the like. If a certain page 102 includes such keyword, the summary selection module 230 may directly select the page as a part of the summary 162, regardless of the importance of the page. The summary selection module 230 may also adjust the importance of the page (for example, adjusting the importance to a value indicating greater importance by weighting or increasing a certain value) in a case that the page 102 includes the keyword.

[0053] In some other implementations, if the document contains a reference to other contents, for example, a reference to a multimedia file (including video, audio, image file and the like) and/or a reference to another document, which means that the content probably included in the page requires further interpretation with extended information of the reference, and thus is important for the entire document. Therefore, the attention

information may further include a reference to a multimedia file and/or another document. The summary selection module 230 may further determine whether a page 102 includes a reference to a multimedia file and/or another document. The summary selection module 230 may determine whether the page 102 includes a reference to a multimedia file and/or another document, and then determine whether the page is selected as a part of the summary 162 based on the determined result. For example, if it is determined that a certain page 102 includes a reference to a multimedia file and/or another document, the summary selection module 230 may directly select the page as a part of the summary 162, regardless of the important score of the page. The summary selection module 230 may further adjust the importance of the page (for example, adjusting to a value indicating greater importance by weighting or increasing a certain value) in a case that the page 102 includes a reference to a multimedia file and/or another document.

[0054] The summary 162 generated by the summary selection module 230 may be presented to a user in various manners. For example, it is presented via a display of the computing device 100. In some implementations, the predetermined number of selected pages is highlighted in the presentation of the plurality of pages 102. In this way, the page as the summary of the document may be displayed distinctively from other pages, thus the user can distinguish these pages from a great number of pages rapidly and thus may choose to browse these pages first to obtain the main or important content of the document. In some implementations, the predetermined number of selected pages or preview pages corresponding to the predetermined number of pages selected may be presented independently from the document. This is advantageous for a user to capture the main or important content of the document rapidly in a case of needing document preview.

[0055] Figs. 5A to 5C illustrate diagrams of example user interfaces for presenting a summary of a document in accordance with some implementations of the subject matter described herein. In the examples as illustrated in Figs. 5A to 5C, the document 152 is shown as a presentation slide document. Fig. 5A illustrates a user interface 501 in a presentation slide application, at which a plurality of pages 102-1 to 102-9 of the document 152 are presented in a “Slide Sorter” mode of slide viewing. The pages 102-1, 102-3, 102-4 and 102-8 selected from these pages as the summary of the document 162 are highlighted.

[0056] Fig. 5B illustrates a user interface 502 of a search application in a search scenario. The user interface 502 displays search results related to a search term “Deep learning

tutorial” returned from a search application, where one search result 510 includes a presentation slide document related to the search term “Deep learning tutorial”, which includes a plurality of slides. Together with the search results, preview pages 512 corresponding to a small number of pages of the presentation slide document are displayed, and these pages are selected as the pages for the summary of the document in accordance with implementations of the subject matter described herein. In this way, the user may determine rapidly, from a plurality of search results, whether the search result 510 is the desired search result, and thus may determine whether the corresponding link of the search result 510 should be clicked, so as to further view all pages of the document and/or download the document.

[0057] Fig. 5C illustrates a user interface 503 of an application of receiving and sending email in a scenario of receiving and sending email. The user interface 503 displays an email received through the application of receiving and sending email, the email includes a presentation slide document 152 as an attachment. The user interface 503 also displays preview pages 512 corresponding to a small number of pages of the presentation slide document 152, and these pages are selected as pages of the summary of the document in accordance with the implementations of the subject matter described herein. As such, the user viewing the email may determine the main or important content of the attachment of the email rapidly, and thus determine the next action, for example, whether the attachment is to be further viewed, how the email is replied, and the like.

[0058] Fig. 5A-5C illustrates some examples for presenting of a summary of a document. According to requirements in the actual applications, the summary of the document may be displayed in various manners. The subject matter described herein is not limited in this regard.

[0059] Fig. 6 illustrates a flowchart of a process 600 of generating a summary of a document in accordance with some implementations of the subject matter described herein. The process 600 may be implemented by the computing device 100, for example implemented at the summary generating module 122 in the memory 120 of the computing device 100. At block 610, the computing device 100 extracts feature information of a plurality of pages comprised in the document, the feature information of each page characterizing at least one type of content contained in the page. At block 620, the computing device 100 determines respective importance of the plurality of pages in the document at least based on the extracted feature information. At block 630, the computing device 100 generates a summary for the document by selecting a

predetermined number of pages from the plurality of pages based on the respective importance, the predetermined number being less than the number of the plurality of pages.

5 [0060] In some implementations, extracting the feature information of the plurality of pages comprises: for each page, determining whether content contained in the page comprises a text; and in response to the content contained in the page comprising the text, extracting at least one of a semantic feature of the text and feature information indicating the number of words comprised in the text, as text-related feature information of the page.

10 [0061] In some implementations, extracting the feature information of the plurality of pages comprises: for each page, determining whether content contained in the page comprises an image; and in response to the content contained in the page comprising the image, extracting at least one of a feature map of the image, feature information indicating an occupied area of the image in the page, and feature information indicating a location of the image in the page, as image-related feature information of the page.

15 [0062] In some implementations, extracting the feature information of the plurality of pages comprises: for each page, converting the page into an image; and extracting a feature map from the image as hybrid feature information of the page.

20 [0063] In some implementations, selecting the predetermined number of pages from the plurality of pages further comprises: determining whether the plurality of pages comprises attention information, the attention information comprising at least one of the following: a keyword, a reference to a multimedia file and a reference to a further document; and selecting the predetermined number of pages from the plurality of pages further based on presence of the attention information in the plurality of pages.

25 [0064] In some implementations, extracting the feature information of the plurality of pages comprises: extracting the feature information of the plurality of pages using a neural network, and determining the respective importance of the plurality of pages comprises: determining the respective importance of the plurality of pages using the neural network.

30 [0065] In some implementations, the neural network is trained using a training document comprising a plurality of training pages, the predetermined number of training pages of the plurality of training pages being labeled as a summary of the training document based on at least one of the following: a duration of presenting the training pages during a presentation of the training document, remark information contained in the training pages, and a manual label related to whether the training pages are used as the summary of the training document.

[0066] In some implementations, the process 600 further comprises: highlighting the predetermined number of selected pages in a presentation of the plurality of pages.

5 [0067] In some implementations, the process 600 further comprises: presenting the predetermined number of selected pages or preview pages corresponding to the predetermined number of pages independently from the document.

[0068] In some implementations, the document is selected from a group consisting of: a presentation slide document, a word processing document, a webpage and a printable document format (PDF) document.

10 [0069] Some example implementations of the subject matter described herein are given below.

[0070] In an aspect, the subject matter described herein provides a computer-implemented method. The method comprises: extracting feature information of a plurality of pages comprised in a document, the feature information of each page characterizing at least one type of content contained in the page; determining respective
15 importance of the plurality of pages in the document at least based on the extracted feature information; and generating a summary for the document by selecting a predetermined number of pages from the plurality of pages based on the respective importance, the predetermined number being less than the number of the plurality of pages.

[0071] In some implementations, extracting the feature information of the plurality of
20 pages comprises: for each page, determining whether content contained in the page comprises a text; and in response to the content contained in the page comprising the text, extracting at least one of a semantic feature of the text and feature information indicating the number of words comprised in the text, as text-related feature information of the page.

[0072] In some implementations, extracting the feature information of the plurality of
25 pages comprises: for each page, determining whether content contained in the page comprises an image; and in response to the content contained in the page comprising the image, extracting at least one of a feature map of the image, feature information indicating an occupied area of the image in the page, and feature information indicating a location of the image in the page, as image-related feature information of the page.

30 [0073] In some implementations, extracting the feature information of the plurality of pages comprises: for each page, converting the page into an image; and extracting a feature map from the image as hybrid feature information of the page.

[0074] In some implementations, selecting the predetermined number of pages from the plurality of pages further comprises: determining whether the plurality of pages comprises

attention information, the attention information comprising at least one of the following: a keyword, a reference to a multimedia file and a reference to a further document; and selecting the predetermined number of pages from the plurality of pages further based on presence of the attention information in the plurality of pages.

5 [0075] In some implementations, extracting the feature information of the plurality of pages comprises: extracting the feature information of the plurality of pages using a neural network, and determining the respective importance of the plurality of pages comprises: determining the respective importance of the plurality of pages using the neural network.

10 [0076] In some implementations, the neural network is trained using a training document comprising a plurality of training pages, the predetermined number of training pages of the plurality of training pages being labeled as a summary of the training document based on at least one of the following: a duration of presenting the training pages during a presentation of the training document, remark information contained in the training pages, and a manual label related to whether the training pages are used as the summary of the
15 training document.

[0077] In some implementations, the method further comprises: highlighting the predetermined number of selected pages in a presentation of the plurality of pages.

20 [0078] In some implementations, the method further comprises: presenting the predetermined number of selected pages or preview pages corresponding to the predetermined number of pages independently from the document.

[0079] In some implementations, the document is selected from a group consisting of: a presentation slide document, a word processing document, a webpage and a printable document format (PDF) document.

25 [0080] In a further aspect, the subject matter described herein provides an electronic device. The electronic device comprises: a processing unit; and a memory coupled to the processing unit and having instructions stored thereon which, when executed by the processing unit, cause the device to perform acts comprising: extracting feature information of a plurality of pages comprised in a document, the feature information of each page characterizing at least one type of content contained in the page; determining
30 respective importance of the plurality of pages in the document at least based on the extracted feature information; and generating a summary for the document by selecting a predetermined number of pages from the plurality of pages based on the respective importance, the predetermined number being less than the number of the plurality of pages.

[0081] In some implementations, extracting the feature information of the plurality of pages comprises: for each page, determining whether content contained in the page comprises a text; and in response to the content contained in the page comprising the text, extracting at least one of a semantic feature of the text and feature information indicating the number of words comprised in the text, as text-related feature information of the page.

[0082] In some implementations, extracting the feature information of the plurality of pages comprises: for each page, determining whether content contained in the page comprises an image; and in response to the content contained in the page comprising the image, extracting at least one of a feature map of the image, feature information indicating an occupied area of the image in the page, and feature information indicating a location of the image in the page, as image-related feature information of the page.

[0083] In some implementations, extracting the feature information of the plurality of pages comprises: for each page, converting the page into an image; and extracting a feature map from the image as hybrid feature information of the page.

[0084] In some implementations, selecting the predetermined number of pages from the plurality of pages further comprises: determining whether the plurality of pages comprises attention information, the attention information comprising at least one of the following: a keyword, a reference to a multimedia file and a reference to a further document; and selecting the predetermined number of pages from the plurality of pages further based on presence of the attention information in the plurality of pages.

[0085] In some implementations, extracting the feature information of the plurality of pages comprises: extracting the feature information of the plurality of pages using a neural network, and wherein determining the respective importance of the plurality of pages comprises: determining the respective importance of the plurality of pages using the neural network.

[0086] In some implementations, the neural network is trained using a training document comprising a plurality of training pages, the predetermined number of training pages of the plurality of training pages being labeled as a summary of the training document based on at least one of the following: a duration of presenting the training pages during a presentation of the training document, remark information contained in the training pages, and a manual label related to whether the training pages are used as the summary of the training document.

[0087] In some implementations, the acts further comprise: highlighting the predetermined number of selected pages in a presentation of the plurality of pages.

[0088] In some implementations, the acts further comprise: presenting the predetermined number of selected pages or preview pages corresponding to the predetermined number of pages independently from the document.

5 [0089] In some implementations, the document is selected from a group consisting of: a presentation slide document, a word processing document, a webpage and a printable document format (PDF) document.

[0090] In another one aspect, the subject matter described herein provides a computer program product, being tangibly stored on a non-transient computer storage medium and comprising machine-executable instructions which, when executed by a device, cause the
10 device to perform acts comprising: extracting feature information of a plurality of pages comprised in a document, the feature information of each page characterizing at least one type of content contained in the page; determining respective importance of the plurality of pages in the document at least based on the extracted feature information; and generating a summary for the document by selecting a predetermined number of pages
15 from the plurality of pages based on the respective importance, the predetermined number being less than the number of the plurality of pages.

[0091] In some implementations, extracting the feature information of the plurality of pages comprises: for each page, determining whether content contained in the page comprises a text; and in response to the content contained in the page comprising the text,
20 extracting at least one of a semantic feature of the text and feature information indicating the number of words comprised in the text, as text-related feature information of the page.

[0092] In some implementations, extracting the feature information of the plurality of pages comprises: for each page, determining whether content contained in the page comprises an image; and in response to the content contained in the page comprising the
25 image, extracting at least one of a feature map of the image, feature information indicating an occupied area of the image in the page, and feature information indicating a location of the image in the page, as image-related feature information of the page.

[0093] In some implementations, extracting the feature information of the plurality of pages comprises: for each page, converting the page into an image; and extracting a
30 feature map from the image as hybrid feature information of the page.

[0094] In some implementations, selecting the predetermined number of pages from the plurality of pages further comprises: determining whether the plurality of pages comprises attention information, the attention information comprising at least one of the following: a keyword, a reference to a multimedia file and a reference to a further document; and

selecting the predetermined number of pages from the plurality of pages further based on presence of the attention information in the plurality of pages.

5 [0095] In some implementations, extracting the feature information of the plurality of pages comprises: extracting the feature information of the plurality of pages using a neural network, and determining the respective importance of the plurality of pages comprises: determining the respective importance of the plurality of pages using the neural network.

10 [0096] In some implementations, the neural network is trained using a training document comprising a plurality of training pages, the predetermined number of training pages of the plurality of training pages being labeled as a summary of the training document based on at least one of the following: a duration of presenting the training pages during a presentation of the training document, remark information contained in the training pages, and a manual label related to whether the training pages are used as the summary of the training document.

15 [0097] In some implementations, the machine executable instruction, when executed by a device, causes the device to: highlight the predetermined number of selected pages in a presentation of the plurality of pages.

[0098] In some implementations, the machine executable instruction, when executed by a device, causes the device to: present the predetermined number of selected pages or preview pages corresponding to the predetermined number of pages independently from the document.

20 [0099] In some implementations, the document is selected from a group consisting of: a presentation slide document, a word processing document, a webpage and a printable document format (PDF) document.

25 [00100] In a still further aspect, the subject matter described herein provides a computer-readable medium, which is tangibly stored on a non-transient computer storage medium and comprises machine-executable instructions which, when executed by a device, causes the device to perform the method in accordance with the above aspects.

30 [00101] The functionally described herein can be performed, at least in part, by one or more hardware logic components. For example, and without limitation, illustrative types of hardware logic components that can be used include Field-Programmable Gate Arrays (FPGAs), Application-specific Integrated Circuits (ASICs), Application-specific Standard Products (ASSPs), System-on-a-chip systems (SOCs), Complex Programmable Logic Devices (CPLDs), and the like.

[00102] Program code for carrying out methods of the subject matter described herein

may be written in any combination of one or more programming languages. These program codes may be provided to a processor or controller of a general purpose computer, special purpose computer, or other programmable data processing apparatus, such that the program codes, when executed by the processor or controller, cause the functions/operations specified in the flowcharts and/or block diagrams to be implemented. The program code may be executed entirely or partly on a machine, executed as a stand-alone software package partly on the machine and partly on a remote machine or entirely on the remote machine or server.

[00103] In the context of this disclosure, a machine-readable medium may be any tangible medium that may contain, or store a program for use by or in connection with an instruction execution system, apparatus, or device. The machine-readable medium may be a machine-readable signal medium or a machine readable storage medium. A machine-readable medium may include but not limited to an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus, or device, or any suitable combination of the foregoing. More specific examples of the machine-readable storage medium would include an electrical connection having one or more wires, a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), an optical fiber, a portable compact disc read-only memory (CD-ROM), an optical storage device, a magnetic storage device, or any suitable combination of the foregoing.

[00104] Further, while operations are depicted in a particular order, this should not be understood as requiring that such operations be performed in the particular order shown or in sequential order, or that all illustrated operations be performed, to achieve desirable results. In certain circumstances, multitasking and parallel processing may be advantageous. Likewise, while several specific implementation details are contained in the above discussions, these should not be construed as limitations on the scope of the subject matter described herein, but rather as descriptions of features that may be specific to particular implementations. Certain features that are described in the context of separate implementations may also be implemented in combination in a single implementation. Rather, various features described in the context of a single implementation may also be implemented in multiple implementations separately or in any suitable sub-combination.

[00105] Although the subject matter has been described in language specific to structural features and/or methodological acts, it is to be understood that the subject matter specified

in the appended claims is not necessarily limited to the specific features or acts described above. Rather, the specific features and acts described above are disclosed as example forms of implementing the claims.

CLAIMS

1. A computer-implemented method, comprising:
 - extracting feature information of a plurality of pages comprised in a document, the feature information of each page characterizing at least one type of content contained in the page;
 - determining respective importance of the plurality of pages in the document at least based on the extracted feature information; and
 - generating a summary for the document by selecting a predetermined number of pages from the plurality of pages based on the respective importance, the predetermined number being less than the number of the plurality of pages.
2. The method of claim 1, wherein extracting the feature information of the plurality of pages comprises: for each page,
 - determining whether content contained in the page comprises a text; and
 - in response to the content contained in the page comprising the text, extracting at least one of a semantic feature of the text and feature information indicating the number of words comprised in the text, as text-related feature information of the page.
3. The method of claim 1, wherein extracting the feature information of the plurality of pages comprises: for each page,
 - determining whether content contained in the page comprises an image; and
 - in response to the content contained in the page comprising the image, extracting at least one of a feature map of the image, feature information indicating an occupied area of the image in the page, and feature information indicating a location of the image in the page, as image-related feature information of the page.
4. The method of claim 1, wherein extracting the feature information of the plurality of pages comprises: for each page,
 - converting the page into an image; and
 - extracting a feature map from the image as hybrid feature information of the page.
5. The method of claim 1, wherein selecting the predetermined number of pages from the plurality of pages further comprises:
 - determining whether the plurality of pages comprises attention information, the attention information comprising at least one of the following: a keyword, a reference to a multimedia file, and a reference to a further document; and
 - selecting the predetermined number of pages from the plurality of pages further based on presence of the attention information in the plurality of pages.

6. The method of claim 1, wherein extracting the feature information of the plurality of pages comprises: extracting the feature information of the plurality of pages using a neural network, and

wherein determining the respective importance of the plurality of pages comprises: determining the respective importance of the plurality of pages using the neural network.

7. The method of claim 6, wherein the neural network is trained using a training document comprising a plurality of training pages, the predetermined number of training pages of the plurality of training pages being labeled as a summary of the training document based on at least one of the following:

a duration of presenting the training pages during a presentation of the training document,

remark information contained in the training pages, and

a manual label related to whether the training pages are used as the summary of the training document.

8. The method of claim 1, further comprising:

highlighting the predetermined number of selected pages in a presentation of the plurality of pages.

9. The method of claim 1, further comprising:

presenting the predetermined number of selected pages or preview pages corresponding to the predetermined number of pages independently from the document.

10. The method of claim 1, wherein the document is selected from a group consisting of: a presentation slide document, a word processing document, a webpage, and a printable document format (PDF) document.

11. An electronic device, comprising:

a processing unit; and

a memory coupled to the processing unit and having instructions stored thereon which, when executed by the processing unit, cause the device to perform acts comprising:

extracting feature information of a plurality of pages comprised in a document, the feature information of each page characterizing at least one type of content contained in the page;

determining respective importance of the plurality of pages in the document at least based on the extracted feature information; and

generating a summary for the document by selecting a predetermined number of pages from the plurality of pages based on the respective importance, the

predetermined number being less than the number of the plurality of pages.

12. The device of claim 11, wherein extracting the feature information of the plurality of pages comprises: for each page,

determining whether content contained in the page comprises a text; and

in response to the content contained in the page comprising the text, extracting at least one of a semantic feature of the text and feature information indicating the number of words comprised in the text, as text-related feature information of the page.

13. The device of claim 11, wherein extracting the feature information of the plurality of pages comprises: for each page,

determining whether content contained in the page comprises an image; and

in response to the content contained in the page comprising the image, extracting at least one of a feature map of the image, feature information indicating an occupied area of the image in the page, and feature information indicating a location of the image in the page, as image-related feature information of the page.

14. The device of claim 11, wherein extracting the feature information of the plurality of pages comprises: for each page,

converting the page into an image; and

extracting a feature map from the image as hybrid feature information of the page.

15. A computer program product being tangibly stored on a computer storage medium and comprising machine-executable instructions which, when executed by a device, cause the device to perform the method of any of claims 1-10.

1/7

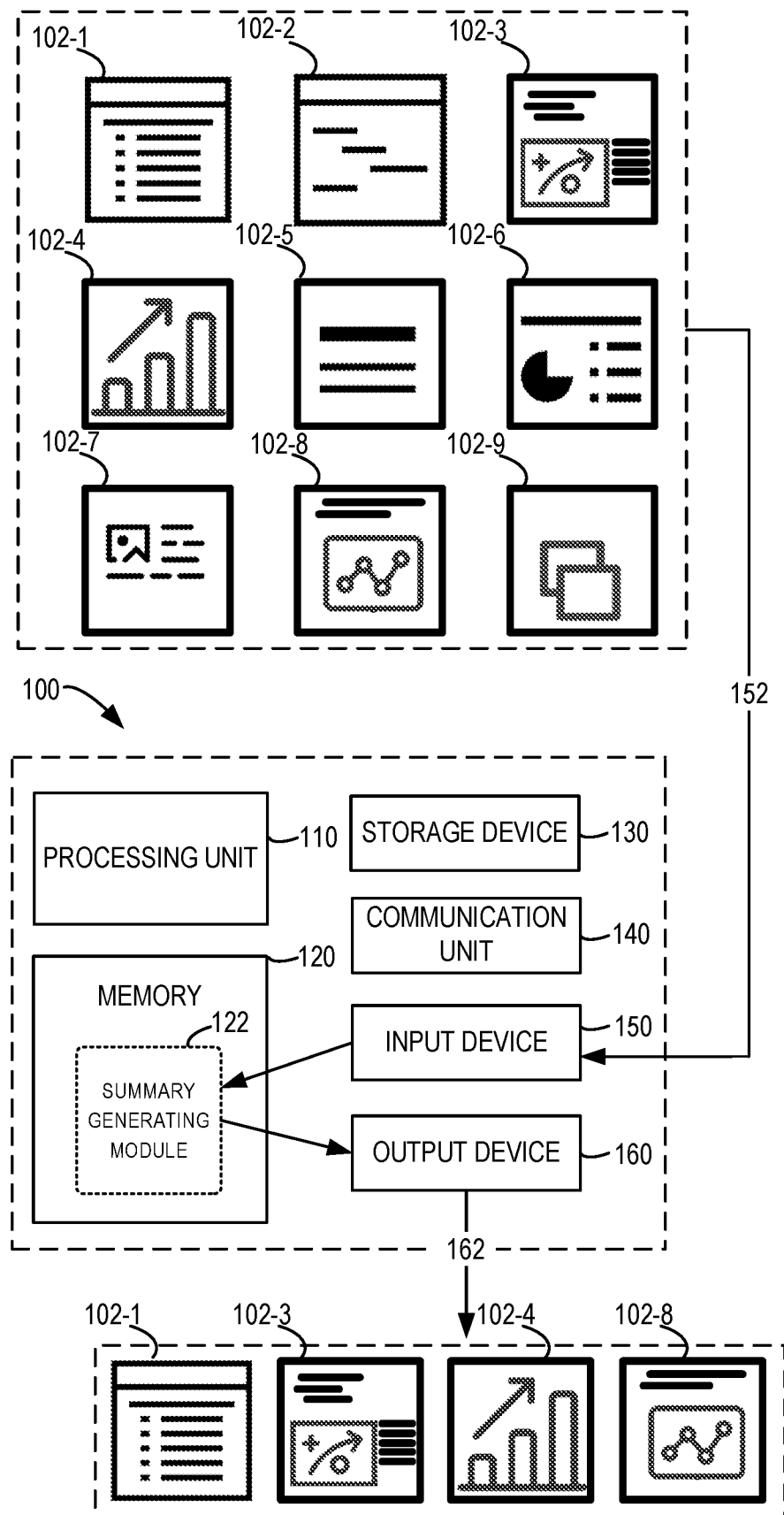


FIG.1

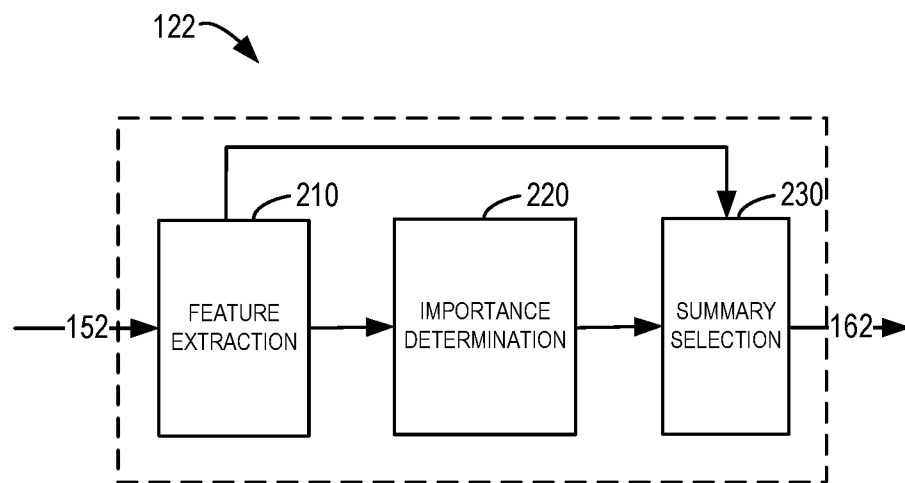


FIG. 2

3/7

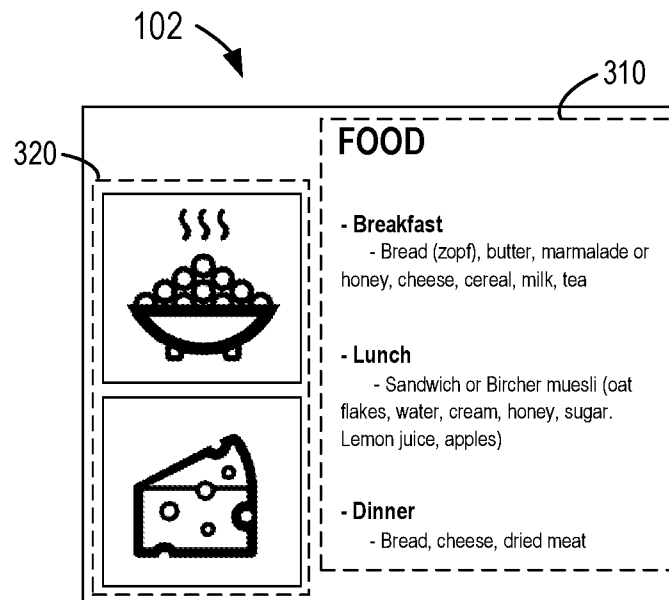


FIG. 3

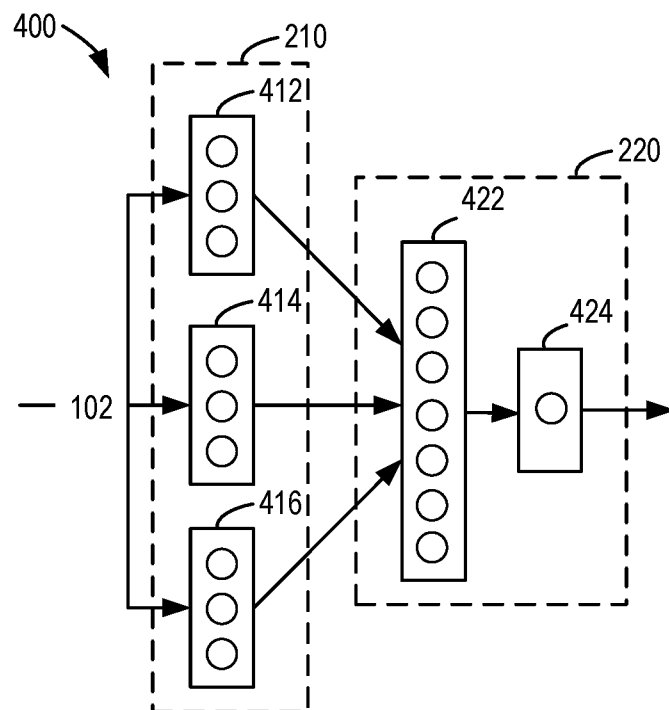


FIG. 4

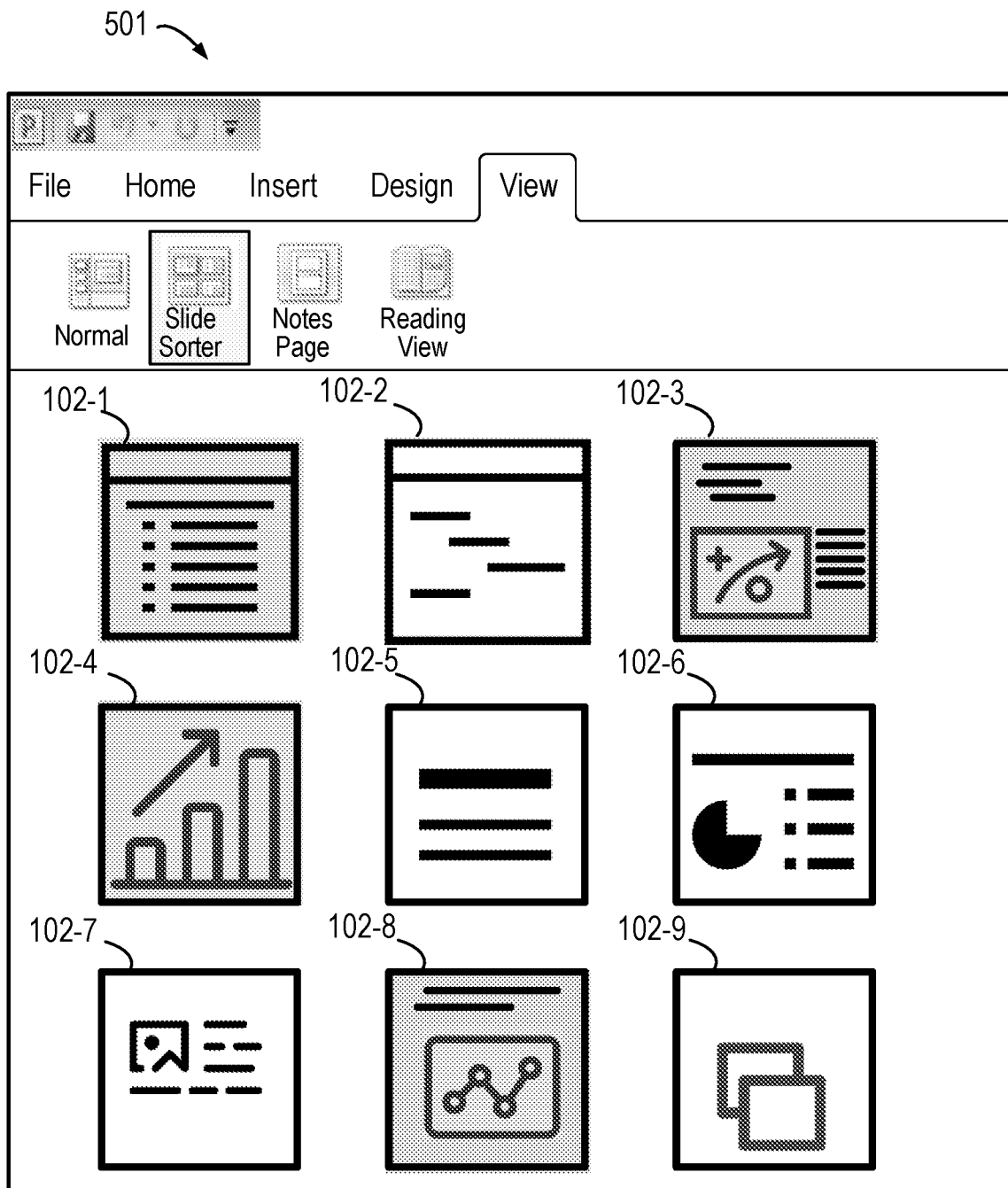


FIG. 5A

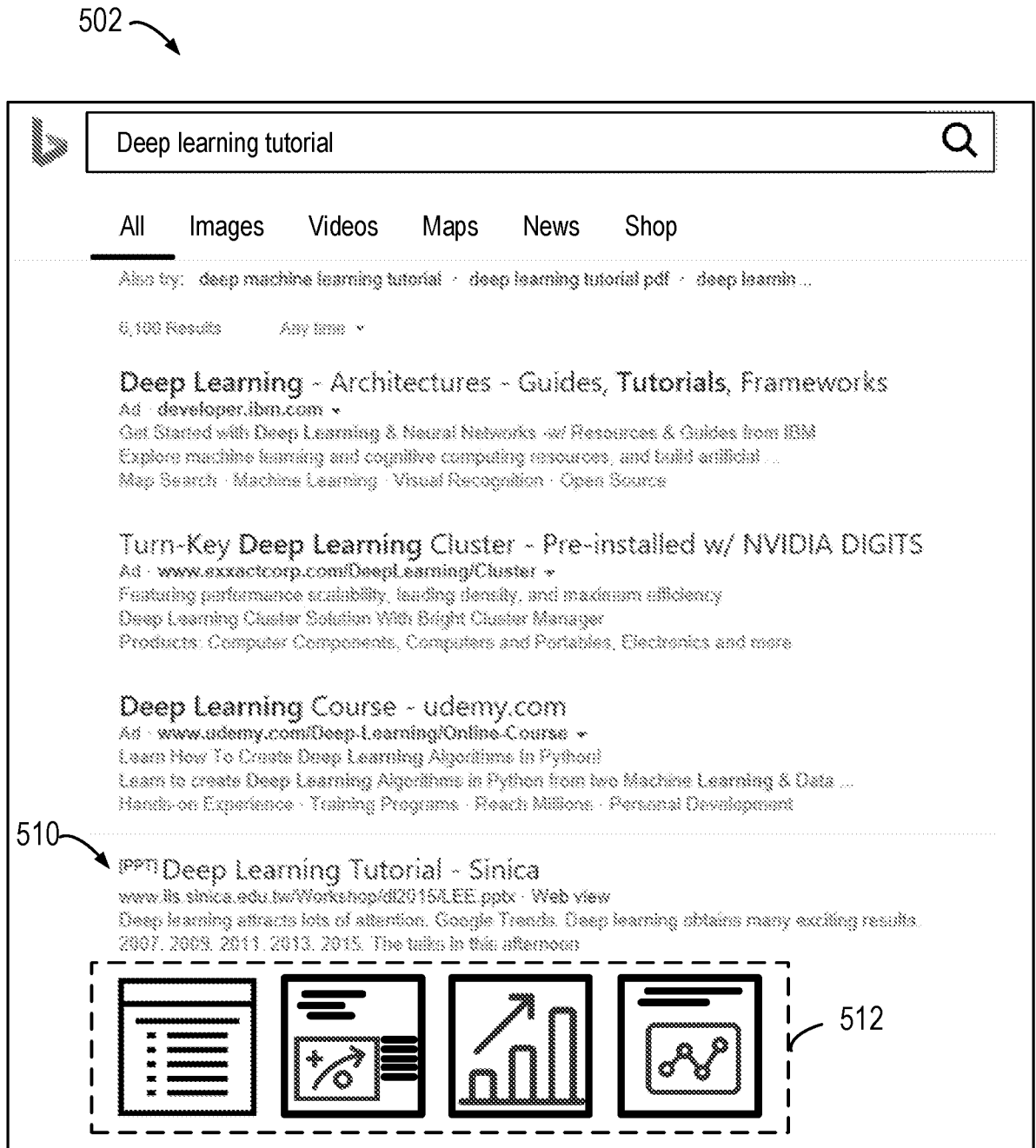


FIG. 5B

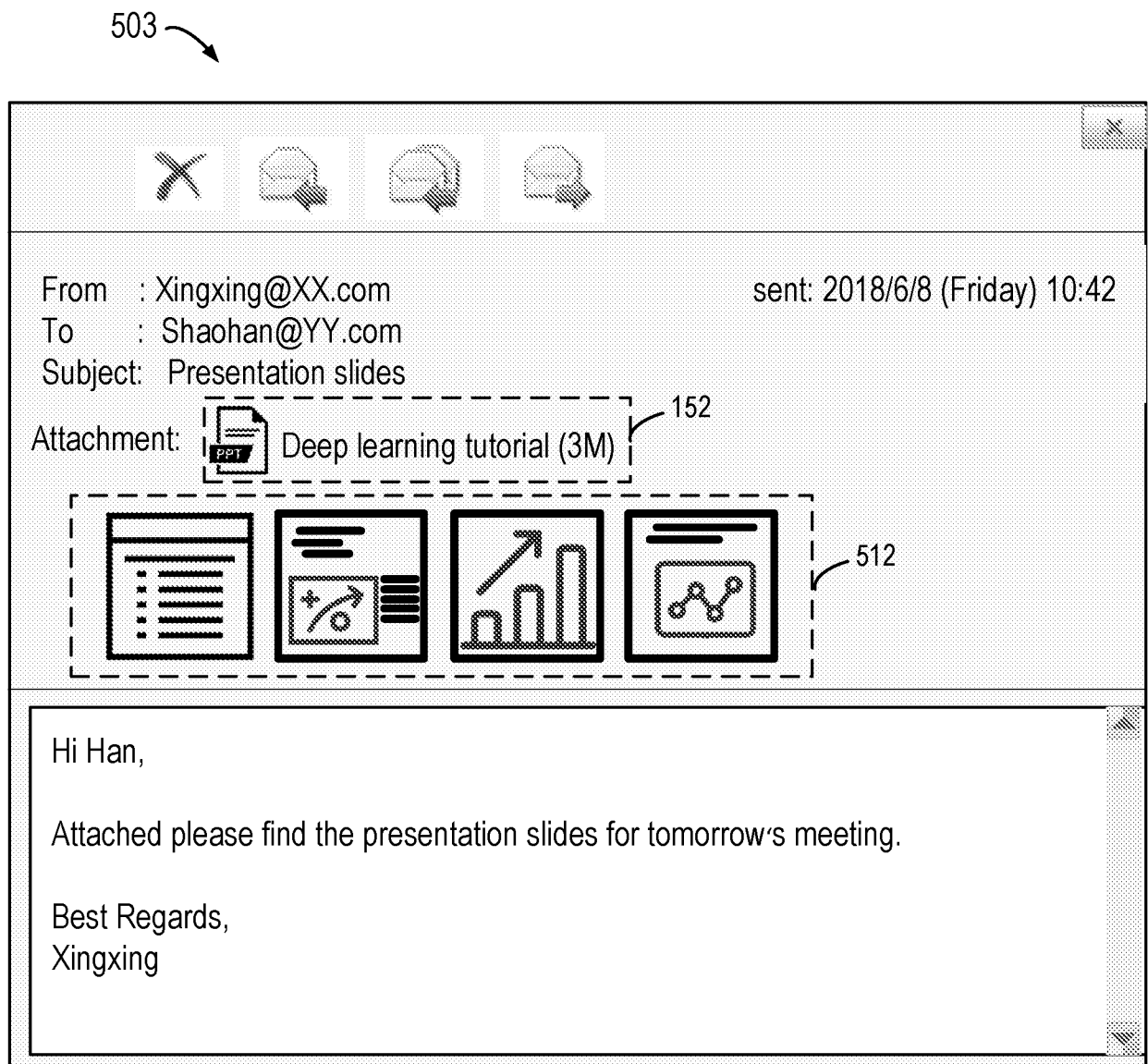
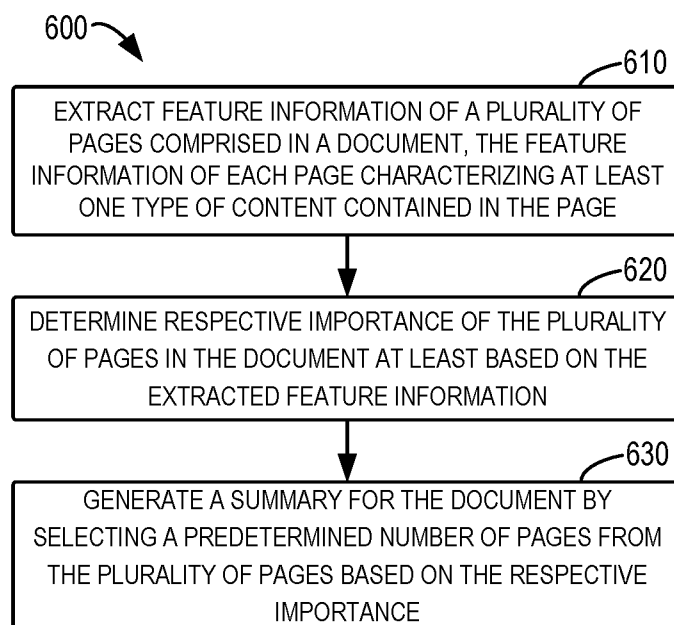


FIG. 5C

**FIG. 6**

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2019/037413

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F16/34
ADD.

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)
G06F

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

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

EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2005/160369 A1 (BALABANOVIC MARKO [GB] ET AL) 21 July 2005 (2005-07-21) abstract; figures 1A, 1B, 2A, 2B, 3 paragraphs [0007], [0018] - [0029], [0042], [0043], [0051], [0060] -----	1-15



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

21 August 2019

Date of mailing of the international search report

27/08/2019

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Hackelbusch, Richard

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2019/037413

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2005160369 A1	21-07-2005	JP 2002063215 A	28-02-2002
		US 6895552 B1	17-05-2005
		US 2005160369 A1	21-07-2005
