



(51) International Patent Classification:

G06F 40/30 (2020.01) G06F 16/35 (2019.01)

(21) International Application Number:

PCT/US2019/060262

(22) International Filing Date:

07 November 2019 (07.11.2019)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

16/182,765 07 November 2018 (07.11.2018) US

(71) Applicant: JPMORGAN CHASE BANK, N.A. [US/US];
383 Madison Avenue, New York, New York 10179 (US).

(72) Inventors: LIN, David D; c/o JPMorgan Chase Bank,
N.A., 383 Madison Avenue, New York, New York 10179
(US). ZHENG, Kent Jiatian; c/o JPMorgan Chase Bank,
N.A., 383 Madison Avenue, New York, New York 10179
(US). SHEN, Kai; c/o JPMorgan Chase Bank, N.A., 383
Madison Avenue, New York, New York 10179 (US).
ROMAHI, Yazann; c/o JPMorgan Chase Bank, N.A., 383
Madison Avenue, New York, New York 10179 (US). LI,
Wei Victor; c/o JPMorgan Chase Bank, N.A., 383 Madi-
son Avenue, New York, New York 10179 (US). SESEN,
Mustafa Berkan; c/o JPMorgan Chase Bank, N.A., 383
Madison Avenue, New York, New York 10179 (US).
STAINES, Joe; c/o JPMorgan Chase Bank, N.A., 383
Madison Avenue, New York, New York 10179 (US).

(74) Agent: KING, Robert A.; c/o Greenberg Traurig, LLP,
Intellectual Property Department, 77 West Wacker Drive,
Suite 3100, Chicago, Illinois 60601 (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.

(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).

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))

(54) Title: SYSTEMS AND METHODS FOR CONTENT FILTERING OF PUBLICATIONS

(57) Abstract: Systems and methods for content filtering based on relevance of publications are disclosed. In one embodiment, in an information processing apparatus comprising at least one computer processor, a method for content filtering of publications may include: (1) receiving a trained neural network, the trained neural network being trained to predict a relevance probability of a publication and comprising a plurality of word vectors, each word vector having a set of trained weights; (2) receiving a publication for evaluation; (3) identifying textual data in the publication; (4) identifying a plurality of sentences in the textual data; (5) identifying a plurality of words in the plurality of sentences; (6) transforming each of the words into word vectors; (7) applying the trained neural network to predict the relevance probability for the publication using the word vectors; and (8) outputting the relevance probability for the publication.



SYSTEMS AND METHODS FOR CONTENT FILTERING OF PUBLICATIONS

RELATED APPLICATIONS

[0001] This application claims the benefit of and priority to U.S. Patent Application Ser. No. 16/182,765, filed November 7, 2018, the disclosure of which is hereby incorporated by reference in its entirety.

BACKGROUND OF THE INVENTION

1. Field of the Invention

[0002] This technology generally relates to content filtering and, more particularly, to systems and methods for content filtering based on relevance of publications.

2. Description of the Related Art

[0003] Publications, such as news publications, newsletters, news websites, blogs, corporate reports and filings, play an important role in both large and small organizations. On a regular day, an organization may receive thousands of news publications from multiple content provider devices. Unfortunately, only a few may be of interest to the business task at hand.

[0004] Organizations often have particular interest in publications of specific topic, such as mergers and acquisitions related news, stock market related news, a company and/or organization specific related news, or product specific news, although other types of topics may be interest. The quality of the results of those searches is based on the keyword search terms and databases used, and, as a result, there is a good chance of missing a publication of interest. Thus, identification of publications of interest using keyboard-based search is often inefficient, time consuming, and prone to error.

SUMMARY OF THE INVENTION

[0005] Systems and methods for content filtering based on relevance of publications are disclosed. In one embodiment, in an information processing apparatus comprising at least one computer processor, a method for content filtering of publications may include: (1) receiving a trained neural network, the trained neural network being trained to predict a relevance probability of a publication and comprising a plurality of word vectors, each word vector having a set of trained weights; (2) receiving a publication for evaluation; (3) identifying textual data in the publication; (4) identifying a plurality of sentences in the textual data; (5) identifying a plurality of words in the plurality of sentences; (6) transforming each of the words into word vectors; (7) applying the trained neural network to predict the relevance probability for the publication using the word vectors; and (8) outputting the relevance probability for the publication.

[0006] In one embodiment, the method may further include tagging each of the words with a part of speech information. In one embodiment, the plurality of words may be tagged with the part of speech using a part-of-speech tagging algorithm. The tagged words may be transformed into word vectors.

[0007] In one embodiment, the publication for evaluation may include a press release, a news publication, a website, a corporate report, a corporate filing, an electronic book, etc.

[0008] In one embodiment, the textual data in the publication may be identified using a character recognition algorithm.

[0009] In one embodiment, the plurality of sentences may be identified using punctuation in the textual data.

[0010] In one embodiment, the tagged words may be transformed into word vectors by creating a vocabulary of tagged words, and assigning each of the tagged words a vector of floating point numbers.

[0011] In one embodiment, the method may further include receiving an override of the relevance probability for the publication; labeling the relevance probability based on the override; and re-training the trained neural network using the publication and the override.

[0012] In one embodiment, the trained neural network may be trained by receiving a plurality of historical publications; identifying textual data in each of the historical publications; identifying a plurality of sentences in the textual data; identifying a plurality of words in the plurality of sentences; transforming each of the words into word vectors; generating the trained neural network using the word vectors; and storing the trained neural network. The trained neural network may be further trained by tagging each of the words with a part of speech information. The tagged words may be transformed into word vectors.

[0013] In one embodiment, a system for content filtering of publications may include a content provider device providing a publication to be evaluated for a relevance probability; a client device that receives the publication and the relevance probability; and a content filtering device in communication with the content provider device and the client device and comprising at least one computer processor. The content filtering device may perform the following: receive a trained neural network, the trained neural network being trained to predict a relevance probability of a publication and comprising a plurality of word vectors, each word vector having a set of trained weights; receive a publication for evaluation; identify textual data in the publication; identify a plurality of sentences in the textual data; identify a plurality of words in the

plurality of sentences; transform each of the words into word vectors; apply the trained neural network to predict the relevance probability for the publication using the word vectors; and output the relevance probability for the publication.

[0014] In one embodiment, the content filtering device may tag each of the words with a part of speech information, and the tagged words may be transformed into word vectors. In one embodiment, the plurality of words may be tagged with the part of speech using a part-of-speech tagging algorithm.

[0015] In one embodiment, the publication for evaluation may include a press release, a news publication, a website, a corporate report, a corporate filing, an electronic book, etc.

[0016] In one embodiment, the textual data in the publication may be identified using a character recognition algorithm.

[0017] In one embodiment, the plurality of sentences may be identified using punctuation in the textual data.

[0018] In one embodiment, the tagged words may be transformed into word vectors by creating a vocabulary of tagged words, and assigning each of the tagged words a floating point number.

[0019] In one embodiment, the content filtering device may also receive an override of the relevance probability for the publication; labels the relevance probability based on the override; and re-train the trained neural network using the publication and the override.

[0020] A method for content filtering is implemented by one or more content filtering devices to identify textual data and non-textual data within a publication. A vector embedding is created based on the textual data. A relevance probability associated with the publication is predicted based on the

trained neural network algorithm with a vector embedding layer. The publication is classified based on an indicator received in response to transmitting the predicted relevance probability.

[0021] A content filtering device includes memory comprising programmed instructions stored thereon and one or more processors configured to be capable of executing the stored programmed instructions to identify textual data and non-textual data within a publication. A vector embedding is created based on the textual data. A relevance probability associated with the publication is predicted based at least on the created vector embedding and a trained neural network algorithm. The publication is classified based on an indicator received in response to transmitting the predicted relevance probability.

[0022] A non-transitory computer readable medium has stored thereon instructions for content filtering including executable code which when executed by one or more processors causes the processors to identify textual data and non-textual data within a publication. A vector embedding is created based on the textual data. A relevance probability associated with the publication is predicted based at least on the created vector embedding and a trained neural network algorithm. The publication is classified based on an indicator received in response to transmitting the predicted relevance probability.

[0023] This technology provides a number of advantages including methods, content filtering devices and non-transitory computer readable media that more efficiently and/or effectively filter news content. With this technology, identification of relevant publications is enhanced based on previously stored mappings in an optimized process to continually improve the accuracy of identifying relevant publications. This technology also provides an

efficient and streamlined process of filtering news data to identify relevant news data of interest to the user quickly and accurately. Accordingly, this technology provides substantially improved content filtering and more efficient and accurate identification and classification of publications.

BRIEF DESCRIPTION OF THE DRAWINGS

[0024] FIG. 1 is a block diagram depicting an exemplary network environment with an exemplary content filtering device according to one embodiment;

[0025] FIG. 2 is a block diagram depicting the exemplary content filtering device of FIG. 1;

[0026] FIG. 3 is a flowchart depicting an exemplary method for training the exemplary content filtering device of FIG. 1;

[0027] FIGS. 4A and 4B depict a flowchart depicting an exemplary method for predicting a relevance of publications based on the trained exemplary content filtering device of FIG. 1.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

[0028] Embodiments are directed to systems and methods for content filtering of publication data. Embodiments may efficiently and effectively filter publications based on the relevance of content associated with the publications.

[0029] Referring to FIG. 1, an exemplary network environment 10 with an exemplary content filtering device 12 is illustrated. In one embodiment, the content filtering device 12 may be coupled, via communication network(s) 18, to client devices 16(1)-16(n) and to content provider devices 14(1). The content filtering device 12, content provider devices 14(1)-14(n), and/or client devices

16(1)-16(n) may be coupled together via other topologies as is necessary and/or desired.

[0030] Referring to FIGS. 1-2, the content filtering device 12 may include one or more processors 24, a memory 26, a communication interface 28, which may be coupled by a bus 30 or other communication link, and a graphical user interface 32. The content filtering device 12 may include other types and/or numbers of elements in other configurations. The processor(s) 24 of the content filtering device 12 may execute programmed instructions stored in the memory 26 for the functions described and illustrated herein. The processor(s) 24 of the content filtering device 12 may include one or more CPUs or general purpose processors with one or more processing cores; for example, although other types of processor(s) may also be used.

[0031] The memory 26 may store programmed instructions for one or more aspects of the present technology as described and illustrated herein, although some or all of the programmed instructions may be stored elsewhere. A variety of different types of storage devices, such as random access memory (RAM), read only memory (ROM), hard disk, solid state drives, flash memory, or other computer readable medium which is read from and written to by a magnetic, optical, or other reading and writing system that is coupled to the processor(s) 24, may be used for the memory 26.

[0032] The memory 26 may store one or more computer programs or applications that may include executable instructions that, when executed by the content filtering device 12, may cause the content filtering device 12 to perform actions, such as to transmit, receive, or otherwise identify textual and non-textual data, create vector embedding and perform classifications of publications, for example, and to perform other actions described and illustrated below with reference to FIGS. 3-5. The applications may be implemented as

modules or components of other applications. Further, the applications may be implemented as operating system extensions, modules, plugins, or the like.

[0033] The computer program/application may execute in a cloud-based computing environment, within or as one or more virtual machines or virtual servers that may be managed, for example, in a cloud-based computing environment. The applications and/or the content filtering device 12 itself may be located in one or more virtual servers running in a cloud-based computing environment rather than being tied to one or more specific physical network computing devices. The applications may be running in one or more virtual machines (VMs) executing on the content filtering device 12.

[0034] In one embodiment, the memory 26 may include a Java Servicing module 34, a content filter module 36, a learning database 38, and data storage 40. The memory 26 may include other policies, modules, databases, or applications, as is necessary and/or desired.

[0035] The graphical user interface (GUI) 32 may couple the client devices 16(1)-16(n) and the Java Servicing module 34, and may generate a user interface to control and manage publications. For example, the GUI 32 may generate and present a user interface at the client devices 16(1)-16(n) and/or at the content filtering device 12 that may display, control and manage classifications of publications. Further, the GUI 32 may provide controls that may be used to manage a classification associated with each publications. For example, GUI 32 may receive an invalidation identifier and may change the classification of publications from “relevant” to “not relevant,” or may receive validation identifier change the classification of the publications from “not relevant” to “relevant.” Any other suitable classification or classification(s) may be used as is necessary and/or desired. Other number and/or types of controls may also be provided as is necessary and/or desired. These

classifications may be used for future training purposes to improve the accuracy of the model.

[0036] In one embodiment, the Java Servicing (JS) module 34 may receive publications from the content provider devices 14(1)-14(n) and may manage the communication of publications among the GUI 32, the content filter module 36, the vector database 38, the classification database 40, and the data storage 42. For example, the JS module 34 may be implemented using a JavaScript algorithm or script, although other algorithms or scripts may be used.

[0037] The JS module 34 may also receive classified publications from the content filter module 36, and may transmit the received classified publications to the GUI 32 and the data storage 40. The JS module 34 may store a list of entries corresponding to the received classified publications in the data storage 40 upon receiving the classified publications from the content filter module 36.

[0038] The JS module 34 may receive vector embedding including word vectors corresponding to each tagged word in the publications from the content filter module 36 and may store the received vector embedding in the learning database 38.

[0039] Other types of communications may also be managed by the JS module 34.

[0040] The content filter module 36 may classify publications as “relevant” or “not relevant,” or using any other suitable classification. The content filter module 36 may identify textual and non-textual data within the received publications and create a vector embedding. The content filter module

36 also may determine a relevance probability associated with each of the received publications.

[0041] In addition, for each word vector, the JS module 34 may store an entry corresponding to the word vector location in the vector embedding.

[0042] The communication interface 28 may couple the content filtering device 12, the content provider devices 14(1)-14(n), and/or the client devices 16(1)-16(n) using communication network(s) 18 and/or direct connections. Other types and/or numbers of communication networks or systems with other types and/or numbers of connections and/or configurations to other devices and/or elements may also be used as is necessary and/or desired.

[0043] In one embodiment, the communication network(s) 18 may include local area network(s) (LAN(s)) or wide area network(s) (WAN(s)), and may use TCP/IP over Ethernet and industry-standard protocols, although other types and/or numbers of protocols and/or communication networks may be used. The communication networks 18 may employ any suitable interface mechanisms and network communication technologies including, for example, teletraffic in any suitable form (e.g., voice, modem, and the like), Public Switched Telephone Network (PSTNs), Ethernet-based Packet Data Networks (PDNs), combinations thereof, and the like.

[0044] The content filtering device 12 may be a standalone device, or it may be integrated with one or more other devices or apparatuses. For example, the content filtering device 12 may include, or be hosted by, one of the content provider devices 14(1)-14(n). Other arrangements may be used as is necessary and/or desired. Moreover, one or more of the elements of the content filtering device 12 may be in a same or a different communication network, including, for example, one or more public, private, or cloud networks.

[0045] In one embodiment, the content provider devices 14(1)-14(n) may be any suitable type of computing device that is capable of storing, serving, and/or providing publications. Each of the content provider devices 14(1)-14(n) may include one or more of processors, memory, and a communication interface, which may be coupled by a bus or other communication link, although any suitable type of network devices may be used. The content provider devices 14(1)-14(n) may also include databases to store publications, etc.

[0046] The content provider devices 14(1)-14(n) may include server devices hosting news websites provided by news providing organizations publications. Exemplary news providing organizations hosting the news websites include, Bloomberg News Portal, Reuters, Investopedia, The Wall Street Journal, the Financial Times, Consumer News and Business Channel (CNBC), Cable News Network (CNN), British Broadcasting Corporation (BBC), Yahoo News, Google News, etc.

[0047] In another example, the content provider devices 14(1)-14(n) may include, for example, server devices hosting websites provided by a content providing organization. The content providing organization hosting websites may include, for example, Wikipedia, Webopedia, Google Scholar, Institute of Electrical and Electronics Engineers (IEEE), Facebook, Twitter, etc.

[0048] The content provider devices 14(1)-14(n) may receive a request for content from the content filtering device 12, process the request for content, and return relevant publications. The request for content may originate with the client devices 16(1)-16(n), for example, from applications hosted on the client devices 16(1)-16(n), although the request for content may also originate elsewhere within the network environment 10.

[0049] In one embodiment, the client devices 16(1)-16(n) may include any type of computing device that may host applications (e.g., application servers) and may transmit requests to the content filtering device 12 for processing by the news content provider devices 14(1)-14(n). Each of the client devices 16(1)-16(n) may include a processor, a memory, and a communication interface, which may be coupled by a bus or other communication link, although other numbers and/or types of network devices could be used.

[0050] Although the exemplary network environment 10 with the content filtering device 12, content provider devices 14(1)-14(n), client devices 16(1)-16(n), and communication network(s) 18 are described and illustrated herein, other types and/or numbers of systems, devices, components, and/or elements in other topologies may be used. It is to be understood that the systems of the examples described herein are for exemplary purposes, as variations of the specific hardware and software used to implement the examples are possible, as will be appreciated by those skilled in the relevant art(s).

[0051] Referring to Figure 3, an exemplary method for training a neural network for relevance probability prediction is disclosed according to one embodiment. In step 300, the content filtering device may access training data in response to, for example, an initialization request received from a client device. For example, the content filtering device may access the data storage to collect previously stored training data stored at the data storage. The previously stored training data may include historical publications that may be manually labeled as relevant publications, as not relevant publications, or using any other suitable classification. The content filtering device may use the manually-labeled publications to train a multi-resolution convolutional neural network (CNN) algorithm. An example of a suitable CNN algorithm is disclosed in Conneau et al., “Very Deep Convolutional Networks for Text Classification,”

available at <https://arxiv.org/pdf/1606.01781.pdf>. The disclosure of this document is hereby incorporated, by reference, in its entirety.

[0052] For example, historical publications used for training may include press releases, news stories, websites, corporate reports and filings (e.g., financial reports associated with corporations, business filings associated with mergers and/or tax returns filings, regulatory filings, etc.), electronic books, other online publications, scanned pages of physical publications, historical websites pages etc. Other types of historical publications may be used for training as is necessary and/or desired.

[0053] In one embodiment, the textual training data may be processed in bulk, in batches, individually, or as otherwise necessary and/or desired.

[0054] In step 302, the content filtering device may use a character recognition algorithm to identify and distinguish textual data and non-textual data content in the historical publications used for training, although other manners of identifying the textual data and/or non-textual data may be used as is necessary and/or desired. Examples of non-textual include graphical images, punctuation marks (e.g., periods, commas, parentheses, etc.), numerical digits, etc. The textual data may include words and letters, etc.

[0055] In step 304, the content filtering device may identify sentences within the historical publications used for training. For example, in one embodiment, the content filtering device may identify a first period mark, and then identify the words before the first period mark as a first sentence. The content filtering device may then identify the next, or second, period mark and identify the words between the first period mark and the second period mark as a second sentence, etc.

[0056] Any other suitable method of identifying sentences may be used as is necessary and/or desired.

[0057] When the content filtering device determines that a suitable number of sentences (e.g., a threshold amount of sentences, all sentences, etc.) have been identified, the content filtering device may apply a text filtering algorithm to extract the textual data corresponding the identified sentences and filter out all non-textual data. Any other suitable filtering method or algorithm may be used as is necessary and/or desired.

[0058] Following sentence identification, in step 306, single or multiple words, n-grams, etc. may be recognized within the filtered sentences.

[0059] In step 308, the content filtering device may apply a part-of-speech (PoS) tagging algorithm to sentences to tag, or associate, each of the words in the sentences with a PoS part. The PoS parts may include, for example, nouns, pronouns, adjectives, determiners, verbs, adverbs, prepositions, conjunctions, interjections, etc. Any other suitable PoS part and/or identifiers may be used as is necessary and/or desired.

[0060] For example, if a sentence is “I went to a lot of stores but came home with only two purchases”, the PoS algorithm may identify “noun” as the PoS part for “purchases.”

[0061] In step 310, the content filtering device may transform the tagged words into corresponding word vectors. In one embodiment, the content filtering device may create a vocabulary for all the tagged words, and may store the vocabulary in the data storage. The content filtering device may then transform each of the tagged words in the vocabulary to a vector of floating point numbers, called a word vector, and may store the word vector.

[0062] In step 312, the content filtering device may determine if additional historical publications are available for review. If so, the process repeats with step 302.

[0063] If no additional historical publications are available for review, in step 314, the content filtering device may perform “back propagation,” or training, to generate a trained neural network using, for example, vector embeddings based on the word vectors or one-hot encoding using a bag-of-words technique, etc. For example, this training may be achieved by applying a gradient descent routine that iteratively attempts to minimize a loss function that quantifies the discrepancies between the manual labels and model prediction. This loss function may focus on increasing the accuracy, recall, precision, etc. of the model. An example of a suitable Gradient descent routine is described in Kingma, Diederik; Ba, Jimmy; “Adam: A Method for Stochastic Optimization,” ICLR 2015, the disclosure of which is hereby incorporated by reference in its entirety.

[0064] In one embodiment, each of the word vectors may be assigned a random initial weight (e.g., a floating point number).

[0065] In one embodiment, the neural network may include an input layer, interim convolutional and fully connected layers, and an output layer. The input layer may receive the word vectors extracted from the textual input as vector embeddings, bag-of-words, etc. The interim convolutional and fully connected layers may extract useful features from the input features. The output layer may output the relevance probability for each of the publications.

[0066] The interim convolutional layers may use various convolution sizes to extract features from local regions in the textual input. For example, a large convolution size may capture the global structure of the collection of the vector

embedding previously stored in the learning database, whereas a small convolution size may capture the local details with high resolution but it does not capture the global structure. In one embodiment, the algorithm may use a combination of convolutional layers with 2, 3, and 5-word convolution sizes in order to extract both the medium scale patterns and local details from the input features in each layer.

[0067] Upon completion of the training, the content filtering device may update the initial weights assigned to the network parameters, resulting in trained weights.

[0068] In step 316 the content filtering device may store the trained neural network. The trained neural network may be stored in the learning database as, for example, a Python object, or any other suitable object in another programming language.

[0069] The trained neural network may be updated using retraining, which may be based on manual-labeled publications.

[0070] Referring to FIG. 4, an exemplary method for content filtering using a trained neural network is illustrated according to one embodiment. In step 400, the content filtering device may receive one or more publications for evaluation from a content provider device. This may be, for example, in response to a request for content sent by the content filtering device. The received publications may include, for example, website pages, corporate reports and filings, other online publications, etc.

[0071] In one embodiment, publications within a certain time period may also be requested and obtained.

[0072] In one embodiment, the content filtering device may request and obtain some or all information from one or more particular data source, within a certain trailing time period, and/or within a general category of information, such as business news. This may obtain an initial dataset that is larger than keyword searching, as the request may not be based on keywords.

[0073] In step 402, the content filtering device may apply a character recognition algorithm to identify and distinguish textual data and non-textual data in content in the one of the publications. This may be similar to step 302, above.

[0074] In step 404, the content filtering device may identify sentences in the publication. This may be similar to step 304, above.

[0075] In step 406, single or multiple words, n-grams, etc. may be recognized within the filtered sentences. This may be similar to step 306, above.

[0076] In step 408, the content filtering device may apply a part-of-speech (PoS) tagging algorithm to sentences to tag, or associate, each of the words in the sentences with a PoS part. This may be similar to step 308, above.

[0077] In step 410, the content filtering device may transform the tagged words into corresponding word vectors, or may apply a bag-of-words technique. This may be similar to step 310, above.

[0078] In step 412, the content filtering device may retrieve the stored trained neural network (e.g., the neural network stored in step 316, above).

[0079] In step 414, the content filtering device may predict a relevance probability for each of the publications using the previously stored trained neural network, including the trained weights assigned to the word vectors.

[0080] As discussed above, the trained neural network may include an input layer, interim convolutional and fully connected layers, and an output layer. The input layer may receive the word vectors extracted from the textual input as vector embeddings. The interim convolutional and fully connected layers may extract useful features from the input features. The output layer may output the relevance probability for each of the publications.

[0081] For example, the output layer may determine the relevance probability that a publication is relevant $P(\text{Relevant})=0.7$ and a probability that the publication is not relevant $P(\text{Not relevant}) = 0.3$. The determined probabilities quantify the chances of the publications being relevant or not relevant publications.

[0082] In step 416, the content filtering device may transmit the relevance probability associated with each of the publications to a client device for validation.

[0083] In step 418, if a manual overwrite indicator is received from one of the client devices (e.g., a change to the labeling of the publication), in step 422, the content filtering device may override the existing relevance probability and label the publication as relevant or non-relevant based on the override. In step 424, the content filtering device may update the training data by adding the manually-labeled publications with their manual classifications to the data storage. The manually-labeled publications and their manual classifications may be used by the content filtering device to re-train the stored neural network.

[0084] In one embodiment, the re-training may be performed periodically, on demand, or as necessary and/or desired.

[0085] If, in step 418, no manual overwrite indicator is received, in step 420, the content filtering device may classify the publications based on the existing relevance probability.

[0086] In one embodiment, the content filtering device may use a pre-defined threshold to classify the publications. For example, for a pre-defined threshold of $P(\text{Relevant}) = 0.65$, the content filtering device may classify a publication as a relevant publication when the relevance probability of the publication is equal to or exceeds 0.65, and as not relevant when the publication has a relevance probability of $P(\text{Relevant})$ of less than 0.65. This threshold is exemplary only, and may vary as is necessary and/or desired.

[0087] The publication may be stored in the data storage along with the assigned classification.

[0088] As additional publications are received, the process may be repeated.

[0089] Having thus described the basic concept of the invention, it will be rather apparent to those skilled in the art that the foregoing detailed disclosure is intended to be presented by way of example only, and is not limiting. Various alterations, improvements, and modifications will occur and are intended to those skilled in the art, though not expressly stated herein. These alterations, improvements, and modifications are intended to be suggested hereby, and are within the spirit and scope of the invention. Additionally, the recited order of processing elements or sequences, or the use of numbers, letters, or other designations therefore, is not intended to limit the claimed processes to any order except as may be specified in the claims. Accordingly, the invention is limited only by the following claims and equivalents thereto.

CLAIMS

What is claimed is:

1. A method for content filtering of publications, comprising:
in an information processing apparatus comprising at least one computer processor:
receiving a trained neural network, the trained neural network being trained to predict a relevance probability of a publication and comprising a plurality of word vectors, each word vector having a set of trained weights;
receiving a publication for evaluation;
identifying textual data in the publication;
identifying a plurality of sentences in the textual data;
identifying a plurality of words in the plurality of sentences;
transforming each of the words into word vectors;
applying the trained neural network to predict the relevance probability for the publication using the word vectors; and
outputting the relevance probability for the publication.
2. The method of claim 1, further comprising:
tagging each of the words with a part of speech information;
wherein the tagged words are transformed into word vectors.
3. The method of claim 1, wherein the publication for evaluation comprises at least one of a press release, a news publication, a website, a corporate report, a corporate filing, and an electronic book.

4. The method of claim 1, wherein the textual data in the publication is identified using a character recognition algorithm.

5. The method of claim 1, wherein the plurality of sentences are identified using punctuation in the textual data.

6. The method of claim 2, wherein the plurality of words are tagged with part of speech information using a part-of-speech tagging algorithm.

7. The method of claim 1, wherein the tagged words are transformed into word vectors by creating a vocabulary of tagged words, and assigning each of the tagged words a floating point number.

8. The method of claim 1, further comprising:
receiving an override of the relevance probability for the publication;
labeling the relevance probability based on the override; and
re-training the trained neural network using the publication and the override.

9. The method of claim 1, wherein the trained neural network is trained by:

receiving a plurality of historical publications;
identifying textual data in each of the historical publications;
identifying a plurality of sentences in the textual data;
identifying a plurality of words in the plurality of sentences;
transforming each of the words into word vectors;
generating the trained neural network using the word vectors; and
storing the trained neural network.

10. The method of claim 9, wherein the trained neural network is further trained by

tagging each of the words with a part of speech information;
wherein the tagged words are transformed into word vectors.

11. A system for content filtering of publications, comprising:
a content provider device providing a publication to be evaluated for a relevance probability;
a client device that receives the publication and the relevance probability;
and

a content filtering device in communication with the content provider device and the client device and comprising at least one computer processor, the content filtering device performing the following:

receive a trained neural network, the trained neural network being trained to predict a relevance probability of a publication and comprising a plurality of word vectors, each word vector having a set of trained weights;

receive a publication for evaluation;

identify textual data in the publication;

identify a plurality of sentences in the textual data;

identify a plurality of words in the plurality of sentences;

transform each of the words into word vectors;

apply the trained neural network to predict the relevance probability for the publication using the word vectors; and

output the relevance probability for the publication.

12. The system of claim 11, wherein the content filtering device further tags each of the words with a part of speech information, and the tagged words are transformed into word vectors.

13. The system of claim 11, wherein the publication for evaluation comprises at least one of a press release, a news publication, a website, a corporate report, a corporate filing, and an electronic book.

14. The system of claim 11, wherein the textual data in the publication is identified using a character recognition algorithm.

15. The system of claim 11, wherein the plurality of sentences are identified using punctuation in the textual data.

16. The system of claim 12, wherein the plurality of words are tagged with a part of speech using a part-of-speech tagging algorithm.

17. The system of claim 11, wherein the tagged words are transformed into word vectors by creating a vocabulary of tagged words, and assigning each of the tagged words a vector of floating point numbers.

18. The system of claim 11, wherein the content filtering device further:

receives an override of the relevance probability for the publication;
labels the relevance probability based on the override; and

re-trains the trained neural network using the publication and the override.

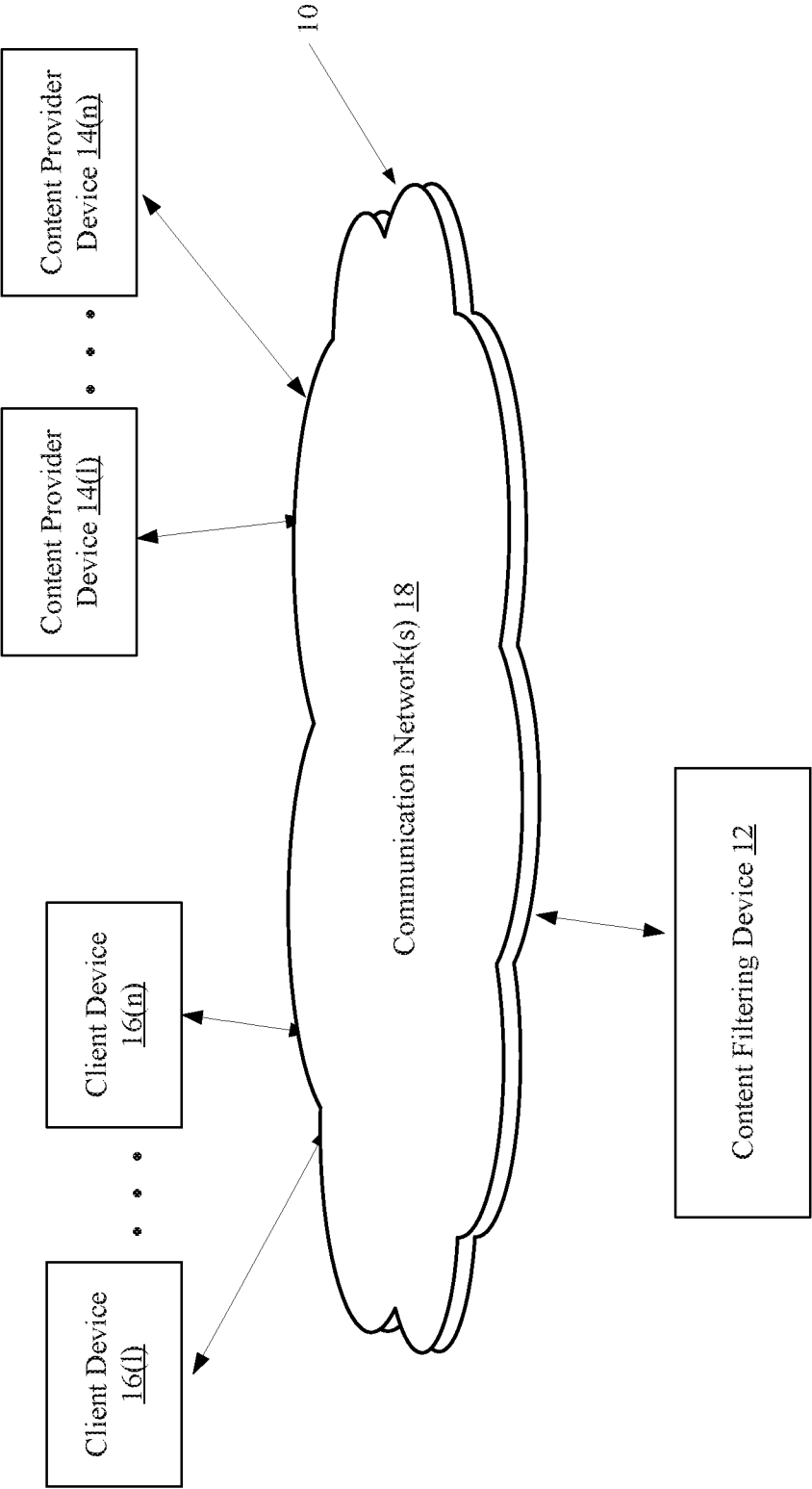


FIGURE 1

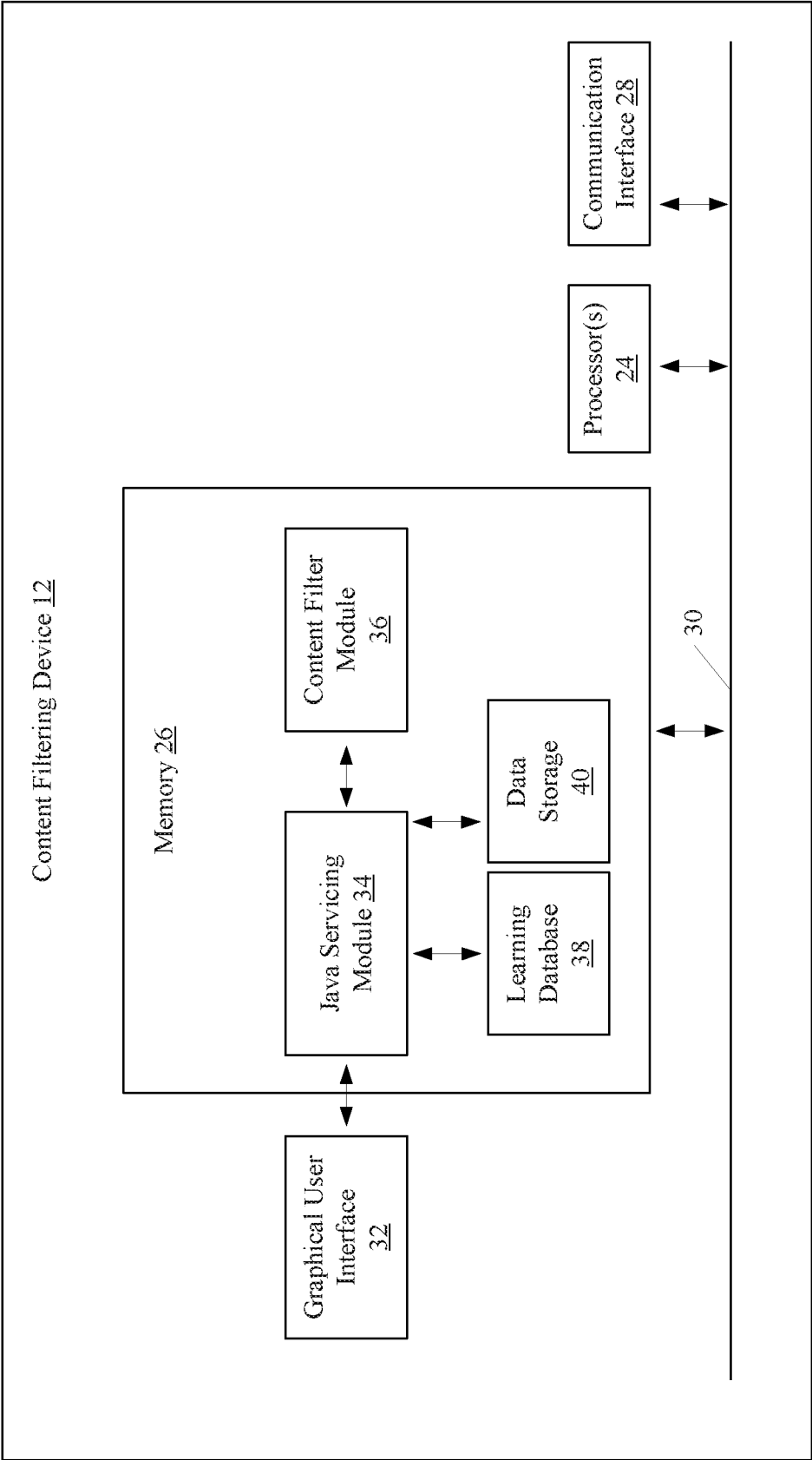


FIGURE 2

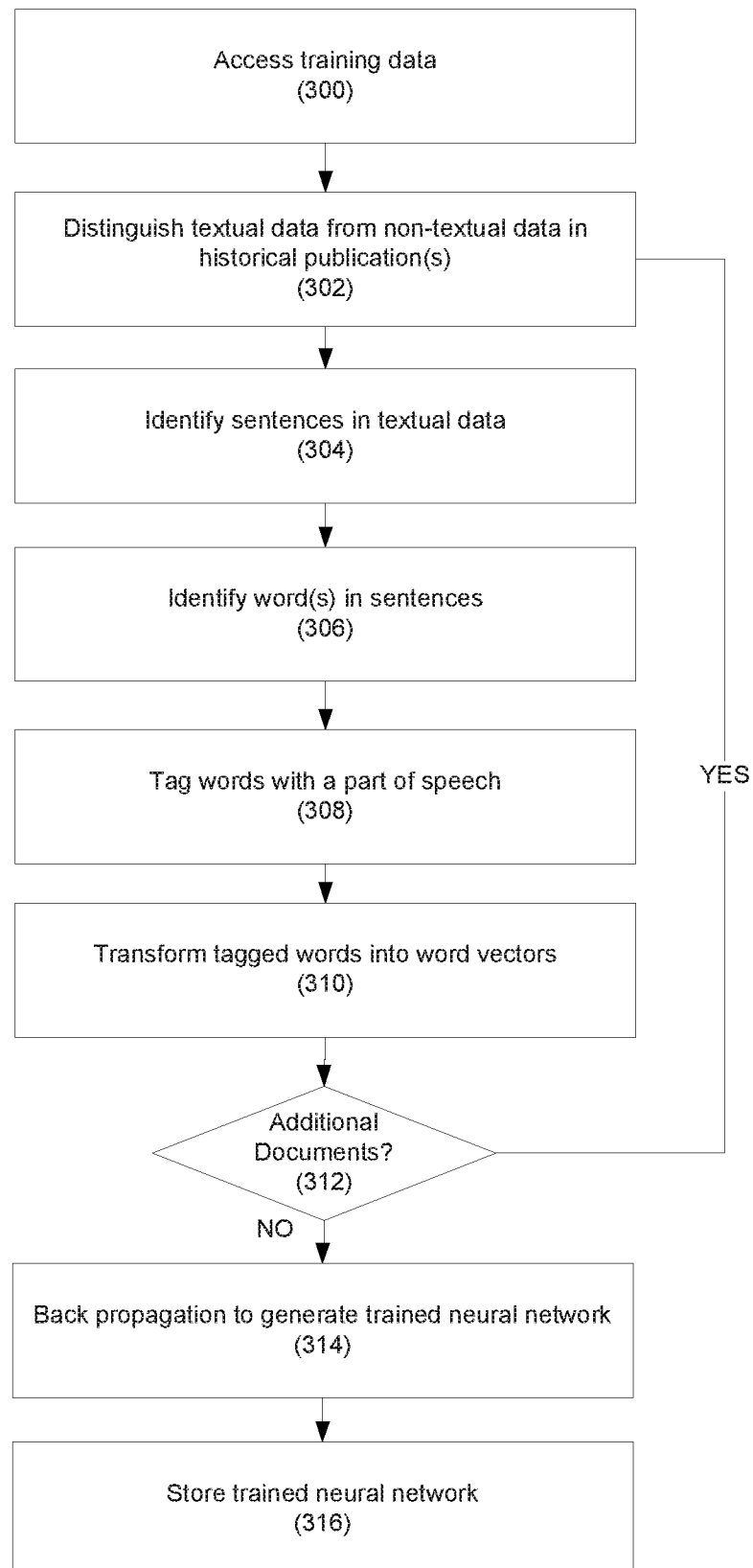


FIGURE 3

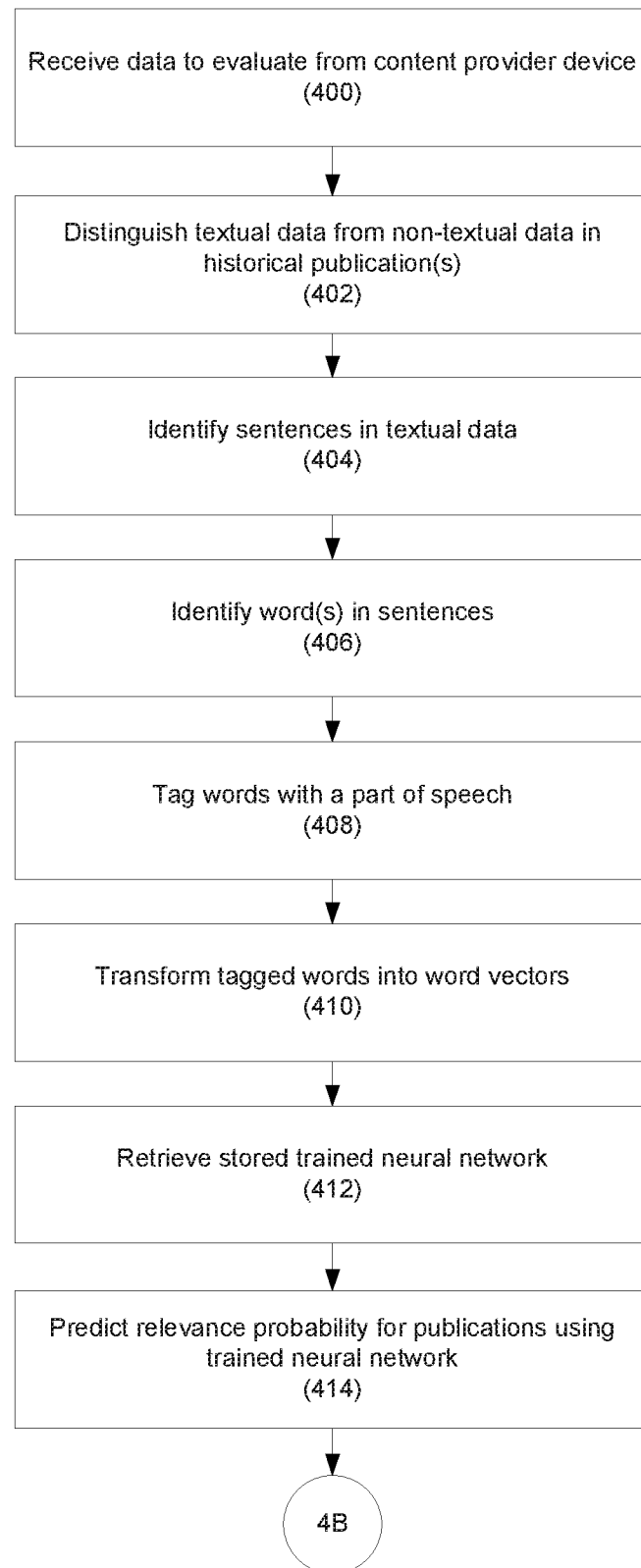


FIGURE 4A

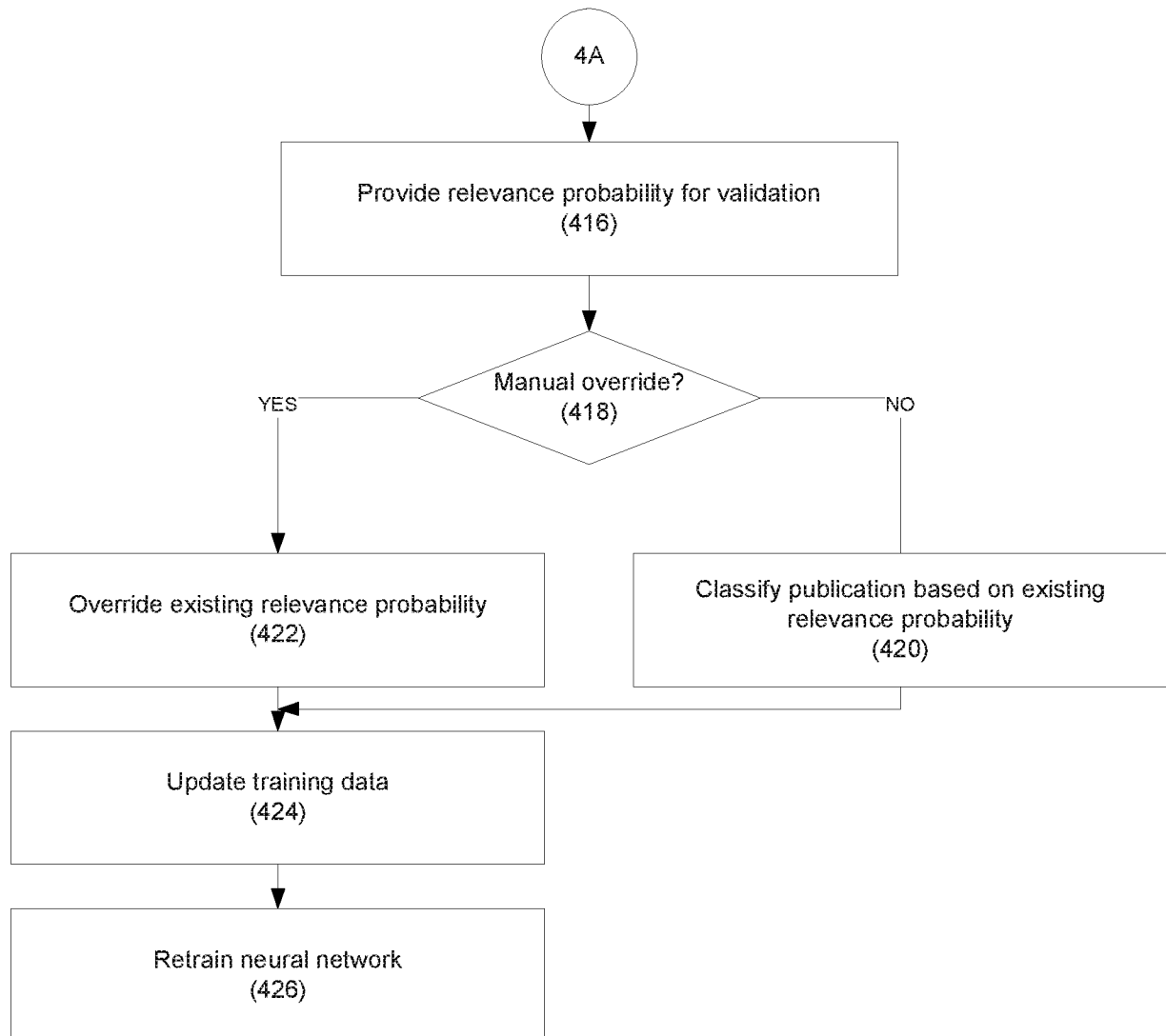


FIGURE 4B

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2019/060262

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F40/30 G06F16/35
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, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2018/196873 A1 (YEREBAKAN HALID ZIYA [US] ET AL) 12 July 2018 (2018-07-12) claim 1	1-18
X	----- ALEXIS CONNEAU ET AL: "Very Deep Convolutional Networks for Text Classification", PROCEEDINGS OF THE 15TH CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: VOLUME 1, LONG PAPERS, 1 January 2017 (2017-01-01), pages 1107-1116, XP055682353, Stroudsburg, PA, USA DOI: 10.18653/v1/E17-1104 cited in the application sections 2, 4.1 and 5 ----- -/-	1-18



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

3 April 2020

Date of mailing of the international search report

16/04/2020

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

Abram, Robert

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2019/060262

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2017/169475 A1 (KORPUSIK MANDY BARRETT [US] ET AL) 15 June 2017 (2017-06-15) paragraph [0054] -----	1-18
A	US 2016/357855 A1 (FAN JAMES J [US] ET AL) 8 December 2016 (2016-12-08) the whole document -----	1-18

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2019/060262

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2018196873 A1	12-07-2018	NONE	
US 2017169475 A1	15-06-2017	JP 2017107569 A	15-06-2017
		US 2017169475 A1	15-06-2017
US 2016357855 A1	08-12-2016	US 2016357855 A1	08-12-2016
		US 2016358094 A1	08-12-2016
		US 2019377747 A1	12-12-2019