January 02, 2019

Applying Machine Learning to Determine Documents Related to Email

Cayden Meyer

Harold Kim

Alan Green

Applying Machine Learning to Determine Documents Related to Email

Abstract:

Machine learning algorithms may determine a document or file that a user intends to attach to an email. Words in the body of the email may indicate an intent to attach a document or file to the email. Based on determining that the user intends to attach a document or file to the email, candidate documents may be found in a remote storage, such as cloud storage, based on the recipient, subject or title, and/or body of the email. The candidate documents may then be ranked and a confidence level determined. Based on the raking and/or confidence level, one or more files or documents may be suggested to the user for attachment.

An attachment determiner may determine one or more files or documents that a user intended to attach to an email. The files or documents may be stored in remote storage, such as cloud storage. The attachment determiner may use one or more machine learning algorithms to find email content and file content that are related to each other. The machine learning algorithm may determine one or more files or documents that are most relevant to the email. The files or documents may be presented to the user as suggestions for attaching to the email, allowing the user to attach the files or documents with a single click per file or document, rather than having to browse through folders and subfolders to find the file(s) or document(s).

FIG. 1 is a network diagram showing devices that may implement some of the techniques described herein. A user may log into a client 102 based on a user account associated with the user to access services such as electronic mail (email) and remote storage. The client 102 may access a remote storage 104 server, which may also be considered cloud storage, and/or an email server 106. The client 102 may access the services via a network 110 such as the Internet. The client 102 may log into the services, such as by providing a username and password, to access the services.

The remote storage 104, which may be considered cloud storage, may store files and/or documents for the user. The remote storage 104 may allow the user to interact with, such as by viewing, editing, commenting, and/or downloading the files or documents. The remote storage 104 may also allow the user to interact with files or documents that other users have shared with the user.

The email server 106 may provide email services to the user. The email server 106 may send and receive emails on behalf of the user. The email server 106 may also send data to the client 102 enabling the client 102 to present emails, including received emails, sent emails, and emails that the user is in the process of drafting, to the user. The email server 106 may allow the user to attach, to emails, files or documents stored in the remote storage 104. The email server 106 may also prompt the user to, or suggest to the user, attaching files or documents that an attachment determiner 108 determines that the user is likely to want to attach to the email.

The attachment determiner 108 may determine files or documents that a user is likely to want to attach to an email. The attachment determiner 108, as well as the remote storage 104 and email server 106, may be implemented as a separate server(s), as an application within a same server as the remote storage 104 and email server 106, or as part of a distributed system.

The attachment determiner 108 may determine files or documents that the user is likely to want to attach to the email based on the contents of the email, such as the recipient of the email, the title or subject line of the email, and the body or content of the email. The attachment determiner 108 may also determine the files or documents based on context attributes, such as files or documents that the user and/or recipient have been interacting with. The attachment determiner 108 may perform a search of files or documents within a storage area associated with the user, such as the remote storage 104, for candidate files or documents. The attachment determiner 108 may rely on and/or perform machine learning algorithms to rank and/or determine confidence scores for the candidate files or documents. The attachment determiner 108 may pass or reference one or more of the files or documents with a confidence score at or above a threshold, such as eighty percent (80%), to the email server 106, and the email server 106 may present the one or more files or documents to the user for attachment to the email.

FIG. 2 shows a pipeline of modules included in the attachment determiner 108 to determine files or documents to present to the user for attachment to an email. The arrows between the modules show data flow. The modules shown and described with respect to FIG. 2 may be included in the attachment determiner 108, or distributed between the attachment determiner 108 and the email server 106.

The attachment determiner 108 may include a tokenizer 202. The tokenizer 202 may parse the email message into tokens, such as words, terms, and/or synonyms and/or equivalent terms.

The attachment determiner 108 may include a term selector 204. The term selector 204 may select from the document the most important words terms, which may include words or phrases parsed by the tokenizer 202 and which are likely to be helpful in determining relevant files, such as terms coming after phrases likely to identify documents. Words in the title or subject, as well as the sender and recipient, may be weighted more heavily than words in the body of the email. For example, in the sentence, “I have attached the presentation,” the word, "presentation," may be relevant because it comes after the word, "attached." The term selector 204 may ignore common words such as, "a," "an," or, "the," which are unlikely to be useful in determining relevant documents.

The attachment determiner 108 may include a candidate searcher 206. The candidate searcher 206 may search for and/or find candidate documents based on the terms selected by the term selector 204. The candidate searcher 206 may search for candidate documents with terms matching and/or similar to the terms selected by the term selector 204, and/or documents that are included in a same cluster (such as a k-means cluster) as the email. The candidate searcher 206 may perform a predetermined number, such as between thirty and forty, of different searches. The candidate searcher 206 may search the remote storage 104 for candidate documents. Not all of the searches may return a candidate document. Thirty to forty searches may, for example, result in ten to fifteen separate documents.

The attachment determiner 108 may include a candidate ranker 208. The candidate ranker 208 may rank and/or determine confidence scores for the files or documents found by the candidate searcher 206. The candidate ranker 208 may convert words or terms in the email and documents into vectors to compare the email to the documents. The candidate ranker 208 may rank and/or determine the confidence scores by applying machine learning to past events of users attaching or not attaching documents to emails. The machine learning algorithm may include a neural network that considers the context of the email and file. The candidate ranker 208 may consider similarity and/or overlap between the email and each document, as well as the context of the email and file. The context may include interaction by the user and the recipient with each document, such as frequency and duration of opening, viewing, editing, or commenting on the document, and references to the document with the email (e.g., “what we were talking about on Wednesday”).

The attachment determiner 108 may include a candidate selector 210. The candidate selector 210 may select the file(s) or document(s) with the highest rank or score, and/or may select a file or document with a confidence level at or above a threshold level, such as eighty percent (80%).

The attachment determiner 108 may include an attachment presenter 212. The attachment presenter 212 may present to the user, for attachment to the email, the file(s) or document(s) selected by the candidate selector 210. The user may accept or decline attaching the file(s) or document(s) with a single click.

FIG. 3 is a flowchart showing a method for adding attachments to an email. The method may be performed in response to a user clicking a send button or otherwise indicating an intent to send the email, or in response to the user completing an email message.

The method may include determining whether an attachment was intended (302). The determination of whether an attachment was intended may be based on the email referencing an attachment or a name of a file or document, using a word often associated with documents such as, “presentation,” “document,” “slides,” or, “spreadsheet,” or based on a previous email requesting a file or document. The determination may be performed in response to the user clicking a send button without including an attachment, or in response to a word or phrase being typed into the interface used to create the email. If no attachment was intended, then the method may include sending the email (314).

If an attachment was intended, then the method may include analyzing the files (304). Analyzing the files (and/or documents) may include finding candidate documents, ranking the documents, and determining confidence scores for the documents, as discussed above.

The method may include determining whether a user likely wanted to attach a particular file (306). The likelihood of attachment may be based on the ranking and/or confidence score of the document determined while analyzing the files (304). If attachment is not likely, then the method may include sending the email (314).

If attachment is likely, then the method may include presenting the attachment (308) to the user. The attachment may be presented within a graphical user interface (GUI) giving the user the option to accept or decline the presented attachment.

After presenting the attachment (308), the method may include determining whether the user accepted the attachment (310). The determining may be based on whether the user clicks a button to attach the file or document to the email or declining the proposed attachment. If the user declined the attachment, then the method may proceed to sending the email (314). If the user accepted the attachment, then the method may include adding the attachment (312) and/or linking the file to a keyword within the email, and then sending the email (314).

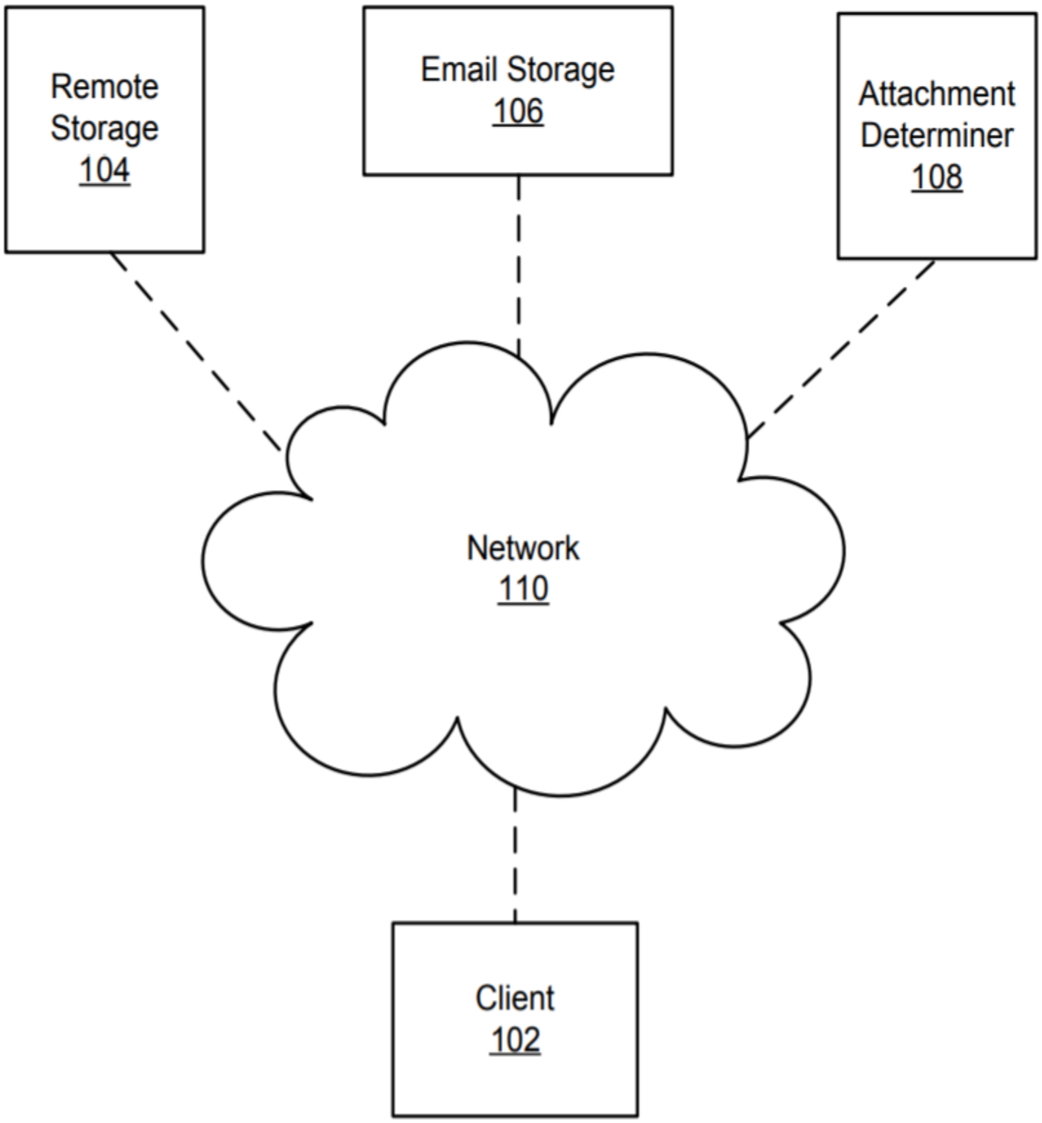
Remote Storage 104
Email Storage 106
Attachment Determiner 108
Network 110
Client 102

FIG. 1

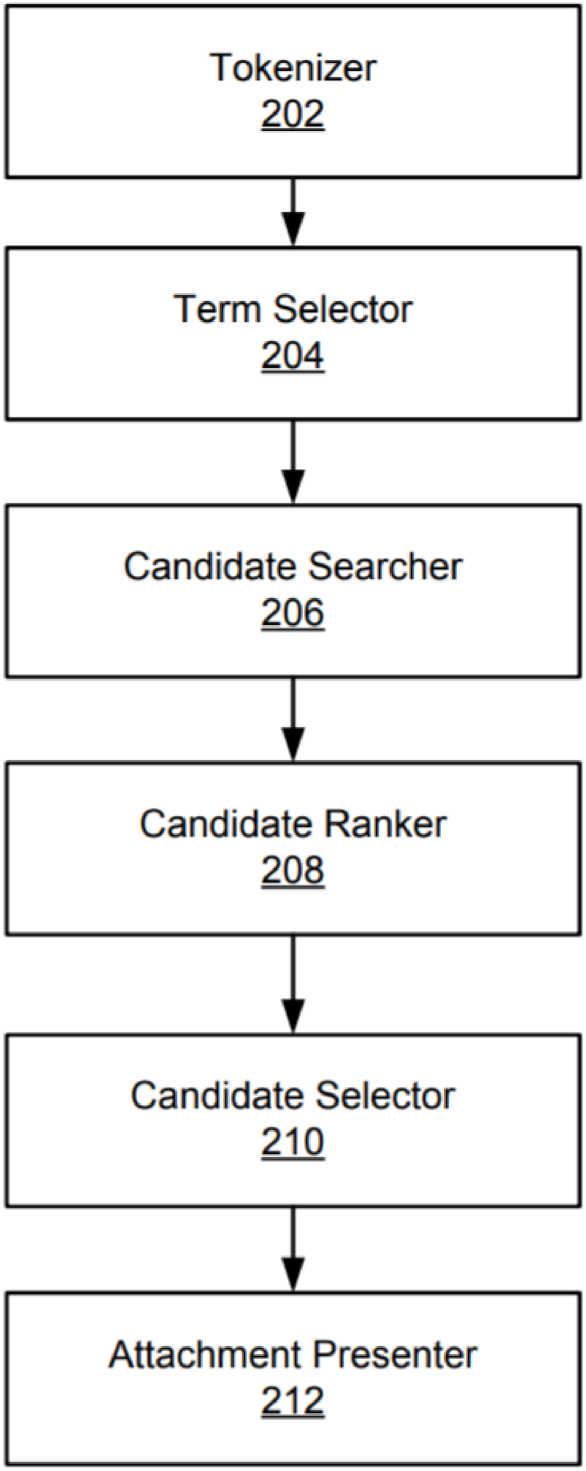
Tokenizer
202
Term Selector
204
Candidate Searcher
206
Candidate Ranker
208
Candidate Selector
210
Attachment Presenter
212

FIG. 2

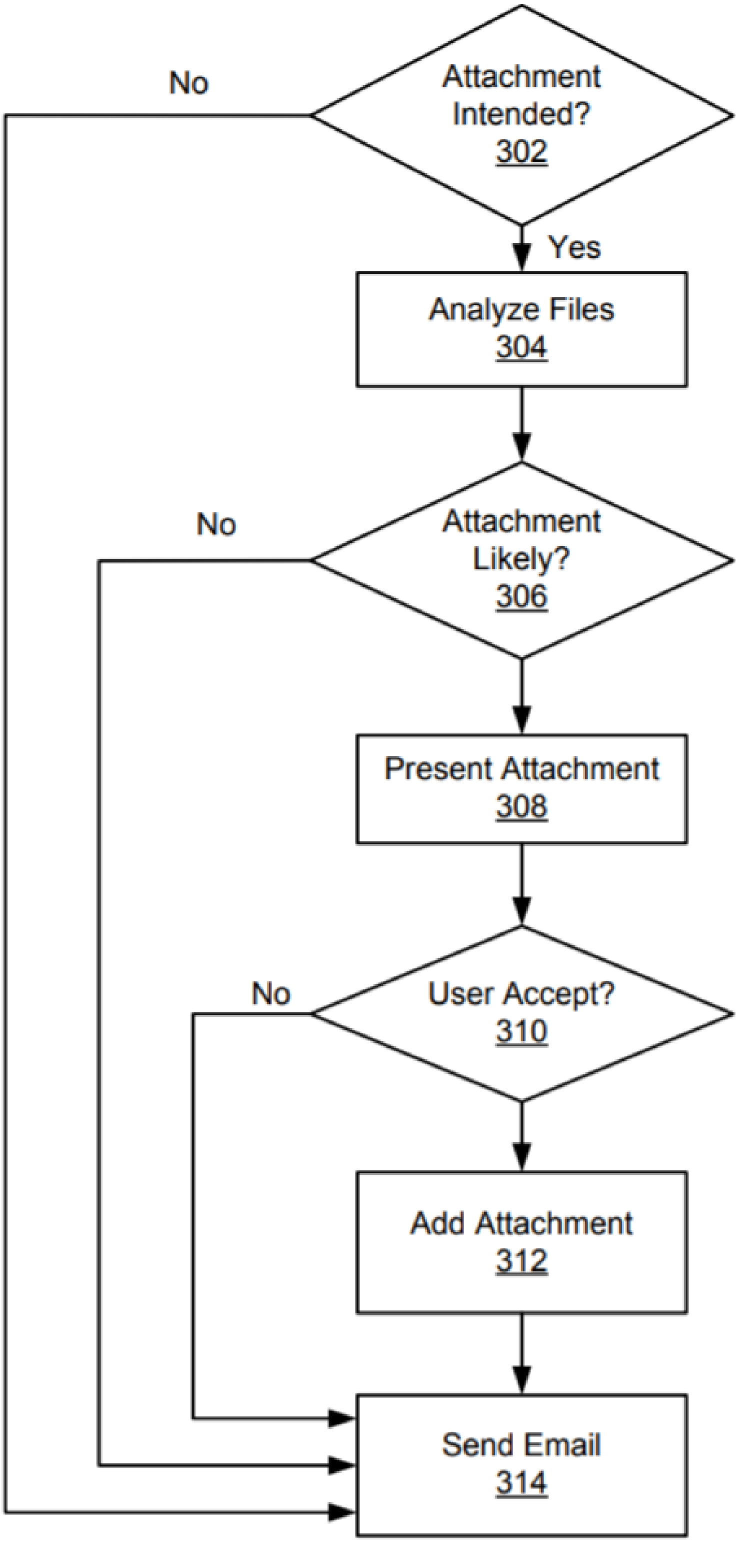
Attachment Intended?
302
No
Yes
Analyze Files
304
Attachment Likely?
306
No
Present Attachment
308
User Accept?
310
No
Add Attachment
312
Send Email
314

FIG. 3