(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
30 November 2017 (30.11.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/205369 A1

(51) International Patent Classification:
G06F 15/16 (2006.01) *G06K 9/18* (2006.01)
G06F 17/30 (2006.01) *H04L 12/58* (2006.01)

(21) International Application Number:
PCT/US2017/033995

(22) International Filing Date:
23 May 2017 (23.05.2017)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
15/161,746 23 May 2016 (23.05.2016) US

(71) Applicant: **GREATHORN, INC.** [US/US]; 116 Beech Street, Floor 2, Belmont, MA 02478 (US).

(72) Inventor: **WALLACE, Raymond, W. III**; c/o Greathorm, Inc., 116 Beech Street, Floor 2, Belmont, MA 02478 (US).

(74) Agent: **VALLABH, Rajesh** et al.; Foley Hoag LLP, 155 Seaport Boulevard, Boston, MA 02210-2600 (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, 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))*

(54) Title: COMPUTER-IMPLEMENTED METHODS AND SYSTEMS FOR IDENTIFYING VISUALLY SIMILAR TEXT CHARACTER STRINGS

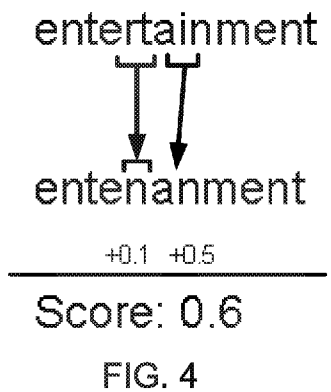


FIG. 4

(57) Abstract: Methods and systems are disclosed for selecting text character strings from a corpus of relevant strings that would commonly be considered to be visually similar to human viewer to an input string. The initial corpus may be any sufficiently broad or specific source of text, e.g., the names of users in a computer application system. The character strings in the corpus are classified such that direct, character-by-character comparisons may be limited to a small subset of likely-similar strings. The input string is then directly compared to strings that are likely to be similar to it, taking into account individual characters' similarities, combinations of characters that look similar to individual characters, transposition of characters, and simple additions and deletions.

WO 2017/205369 A1

COMPUTER-IMPLEMENTED METHODS AND SYSTEMS FOR IDENTIFYING VISUALLY SIMILAR TEXT CHARACTER STRINGS

RELATED APPLICATIONS

[0001] This application claims the benefit of priority to United States Patent Application serial number 15/161,746, filed May 23, 2016, the contents of which is hereby incorporated by reference.

BACKGROUND

[0002] The present application relates generally to computer-implemented methods and systems for analyzing an input text character string to identify one or more visually similar text character strings in a corpus of such strings.

[0003] Electronic messaging systems that exchange information between business users are essential for the communication and operation of nearly every business. More than 135 billion email messages are sent to business users daily as of 2015, and this number is expected to continue to increase. In addition, real-time communication technologies such as text and video chat, intranet sites, SMS, and voice-over-IP are also seeing increased adoption.

[0004] Over 90% of all security incidents and data breaches begin when these technologies are used fraudulently, wherein an attacker manipulates these systems to send communications that appear to be from someone who is known to the recipient. These fraudulent messages are then used to gain illegitimate access to networks, computer systems, financial resources, and protected data, often for the financial gain of the impersonating party.

[0005] This is a significant problem because communications over network-based systems are largely unprotected in terms of identity validation, and the information accessible to the accounts used within those systems is often highly valuable. For example, a recipient of a fraudulent message may be a senior executive within a business unit, with the authority to authorize the wiring of funds to third parties. If a fraudulent message convinces him or her that it is from another executive, he or she may send funds to recipients who are not who the impersonated message claims them to be.

[0006] The reason is that email, chat, voice, and other network-based communication systems contain semi-structured data fields, such as “From:” and “To:”

fields in an email, as well as unstructured data fields such as the message body or the content of the chat message being exchanged. Because both of these types of fields can be manipulated by a fraudulent party, and because no current standard exists that allows the recipient to verify that the author, sender, or creator of any individual message is who they claim to be, these attacks can be highly effective.

[0007] As a result, the use of fraudulent communication messages exchanged over a network is expected to grow as the use of these systems continues to increase. Preventing fraudulent activity without impeding the flow of legitimate communication is a particular problem for managers of information technology systems, as these managers are responsible for both determining what is and is not legitimate while simultaneously ensuring that the deliverability of communications is not delayed or prevented. Performing both of these functions simultaneously requires the information technology manager to rapidly identify extremely nuanced details about each messages structured and unstructured data fields, and to do so without slowing the rate of communications such that business processes are impacted.

[0008] Despite this contradictory set of requirements, regulatory and audit frameworks such as the Health Insurance Portability and Accountability Act of 1996 and the Sarbanes-Oxley Act may require information technology managers to track and protect against the exposure of data that is linked to these kinds of communication system accounts.

[0009] Various embodiments disclosed herein relate to improved techniques for separating legitimate and illegitimate messages within these systems, without introducing any delay to the use of these systems.

BRIEF SUMMARY OF THE DISCLOSURE

[0010] A computer-implemented method in accordance with one or more embodiments analyzes an input text character string to identify one or more visually similar text character strings in a corpus of text character strings. The method comprises the steps, each performed by a computer system, of: (a) receiving the input text character string; (b) determining a scalar estimate for the input text character string based on a model of visually similar text characters stored in a computer storage system, said model comprising a plurality of look-alike sets of text characters, each set having an assigned numerical value;

(c) identifying a set of candidate text character strings from the corpus of text character strings stored in the computer storage system by comparing the scalar estimate of the input text character string with pre-calculated scalar estimates of each of said text character strings in the corpus, and selecting text character strings having scalar estimates within a given range of the scalar estimate of the input text character string; (d) analyzing the input text character string with each of the candidate text character strings to determine a visual similarity score between the input text character string and each of the candidate text character strings; and (e) identifying one or more candidate text character strings deemed to be most visually similar to the input text character string based on the similarity scores.

[0011] A computer system in accordance with one or more further embodiments comprises at least one processor, memory associated with the at least one processor, and a program supported in the memory for analyzing an input text character string to identify one or more visually similar text character strings in a corpus of text character strings, the program containing a plurality of instructions which, when executed by the at least one processor, cause the at least one processor to: (a) receive the input text character string; (b) determine a scalar estimate for the input text character string based on a model of visually similar text characters stored in a computer storage system, said model comprising a plurality of look-alike sets of text characters, each set having an assigned numerical value; (c) identify a set of candidate text character strings from the corpus of text character strings stored in the computer storage system by comparing the scalar estimate of the input text character string with precalculated scalar estimates of each of said text character strings in the corpus, and selecting text character strings having scalar estimates within a given range of the scalar estimate of the input text character string; (d) analyze the input text character string with each of the candidate text character strings to determine a visual similarity score between the input text character string and each of the candidate text character strings; and (e) identify one or more candidate text character strings deemed to be most visually similar to the input text character string based on the similarity scores.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] FIG. 1A shows a list of example text strings, all variations of the word "entertainment." Each variation differs from the original in a manner that makes it likely to

be mistaken for the original word, especially when viewed in isolation. FIG. 1B shows the example text strings of FIG. 1A, with the particular differences with the original underlined.

[0013] FIG. 2 shows an example of a partial look-alike set table, which is used to identify and score visually similar individual characters in accordance with one or more embodiments.

[0014] FIG. 3 shows an exemplary calculation for determining the Scalar Estimate of the word "pail" using the look-alike sets shown in FIG. 2.

[0015] FIG. 4 is a visual depiction of an exemplary Comparison Score calculation in accordance with one or more embodiments, comparing the word "entertainment" to another character string bearing two differences from the word.

[0016] FIG. 5 is a simplified block diagram illustrating an exemplary computer system for identifying visually similar text character strings in accordance with one or more embodiments.

DETAILED DESCRIPTION

[0017] Spelling mistakes in natural writing are common because the pronunciation of one or more characters can, at times, differ very little from one or several other combinations of characters. People are generally good, however, at noticing these different arrangements of characters by sight. Though common misspellings may sound similar to the proper spelling of a word, they are generally visually discernable. When they know the correct spelling of a word, a human being is capable of noticing such transliterations.

[0018] Another class of misspellings, though may sound rather different when correctly pronounced, can be very difficult for a person to notice when reading from a page or computer screen. This kind of misspelling, when done with intention, can be used to fool even a very attentive person. This is the basis of the what is commonly known as a "homograph attack" – an attacker utilizes an otherwise legitimately named resource that is visually similar to a resource their victim is familiar with in order to fool the victim into taking some desired action, e.g., responding to financial request or providing information known to the victim.

[0019] Various embodiments disclosed herein are directed to methods and systems for determining the degree of similarity between two potentially unique but visually similar

text character strings (representing, e.g., domain names used in email message exchanges), offering improved efficiency over existing string comparison techniques.

[0020] Techniques disclosed herein analyze an input text character string and select text character strings from a corpus of relevant strings that would commonly be considered to be visually similar to human viewer. The corpus may be any sufficiently broad or specific source of text, e.g., the names of users in a computer application system. The character strings in the corpus are classified such that direct, character-by-character comparisons may be limited to a small subset of likely-similar strings. The input string is then directly compared to strings that are likely to be similar to it, taking into account individual characters' similarities, combinations of characters that look similar to individual characters, transposition of characters, and simple additions and deletions.

[0021] A model of similar characters is used to determine whether two character strings are visually similar. The model used can depend greatly on the context the strings will be seen in, particularly common characteristics of the font faces used to display or print the strings. As such, specific applications of the techniques disclosed herein may use models tuned to the particular context of the application.

[0022] In separating the set strings from a given corpus of strings that are similar to an input or "suspect" string, it is preferable to know the exact Similarity Score of each pair of corpus and suspect strings. For large corpuses this is quite time consuming. As such, a Scalar Estimate of which "similarity class" each character string in the corpus is pre-calculated. A precise definition of a "similarity class" need not be made here, as its existence in the abstract is sufficient to calculate an approximation of what class a character string would belong to that is appropriate to compare to other Scalar Estimates.

[0023] To the calculate the Scalar Estimate, the available set of characters is first portioned into approximate "look-alike sets," where each set contains characters that are visually similar (within the context of interest) enough. Each look-alike set should be similar in size, though they needn't be equal, and each character in the available character set should be represented in exactly one alike-set. Each look-alike set is given a numerical value arbitrarily. FIG. 2 shows an example of a partial look-alike set table used to identify and score visually similar individual characters in accordance with one or more embodiments.

[0024] For a given alike-set partitioning, an individual character string's Scalar Estimate is the sum of the numerical value of the string's characters, modified by their position in the string, such that characters nearest the start of the string are valued more closely to their alike-set value than those at the end, weighting the Estimate toward the front of the string – where a person's focus often lies when evaluating small strings. FIG. 3 shows an exemplary calculation for determining the Scalar Estimate of the word "pail" using the look-alike sets shown in FIG. 2.

[0025] The Scalar Estimate can be used to filter a large string corpus to a "candidate set" of only those strings whose Scalar Estimates are within a particular delta of the "suspect" string's Estimate. This delta should be chosen as appropriate to the corpus size and desired tolerance of the filter such that the candidate set is of a size small enough to be directly compared to the suspect string within the desired time constraints.

[0026] With the candidate set selected, the suspect string is compared to each string to find the string or strings that it is most visually similar to. This is a done by calculating a Similarity Score for each pairing of the suspect string with a character strings in the candidate set.

[0027] Prior to comparing the suspect string to the candidate set, a table of look alike-sets and "ligature" table is determined.

[0028] The table of alike-sets used to calculate a Similarity Score is similar to the alike-set partitioning used to calculate the Scalar Estimate, however not every character in the available character set need be represented in any alike-set, and a given alike-set should include only particularly visually similar characters, given the context the suspect string may be found in.

[0029] The ligature table is constructed as a mapping of input character strings to output character strings such that a given input string of one or two characters may be provided and, if the input string is present in the mapping, produce an output string that visually similar to the output. For example, in many common contexts, the input character string 'rn' is visually similar to the output character string 'm' as shown in FIG. 4.

[0030] For a given candidate string and suspect string, a Similarity Score is calculated from 0.0 when the two character strings are identical, to a value equal to the length of the longer of the two strings when similarities between the two character strings

exist. Beginning at 0.0, the Score is increased for each instance where the character strings differ visually. The Score is increased by different amounts depending on the type of visual difference encountered:

- Corresponding characters from the same look-alike set increases the Score by 0.1
- A character in one character string corresponding to a "ligature" in the opposite character string increases the Score by 0.1
- A transposition of corresponding pairs of characters increases the Score by 0.25
- A character missing from one character string at a corresponding position in the other string increases the Score by 0.5
- Corresponding characters that are strictly different increases the Score by 1.0

[0031] Where two types of visual difference might be represented by the same, or intersecting, corresponding characters, the choice of difference type(s) representing a lower overall Similarity Score is preferred.

[0032] As such, the overall Similarity Score is the sum of the scores of all visual differences between a given pair of character strings.

EXAMPLES

[0033] An exemplary computer-implemented method for identifying visually similar character strings in accordance with one or more embodiments is illustrated below. First, an input text string is analyzed to identify a set of possible strings that look visually similar to the input string.

[0034] (a) The available text characters are partitioned into rough look-alike sets, each given a different numerical value.

[0035] (b) An input text character string is processed, calculating a Scalar Estimate such that:

[0036] (i) the Estimate begins at 0.0

[0037] (ii) repeated characters - characters identical the character that precedes them - are discarded

[0038] (iii) each remaining character is assigned the numerical value of the look-alike set it belongs to

[0039] (iv) the numerical value of each character is modified by multiplying the numerical value by the ratio of the distance of the character's position from the end of the "input" string and the full length of the string

[0040] (v) each modified numerical value is added to the Estimate

[0041] (c) The final Scalar Estimate places the string in an approximate class of character strings which encompasses at least the possible strings that look visually similar to the input string.

[0042] A method In accordance with one or more further embodiments includes the following steps:

[0043] (a) The available text characters are partitioned into look-alike sets - sets of characters that look similar.

[0044] (b) The available text characters are partitioned into a ligature table - a look up table mapping individual key characters to specific ligature character strings that look similar to the key character.

[0045] (c) Two character strings, the "suspect" (i.e., input) and "candidate" strings, are compared by a backtracking process, wherein characters of corresponding string positions are compared and a Similarity Score is calculated:

[0046] (i) the Score begins at 0.0

[0047] (ii) the position in the "suspect" and "candidate" strings begins at their first character

[0048] (iii) for each iteration, one of the following options is chosen such that the Score is minimized:

[0049] 1. if the two current characters are identical, the Score is not changed and both string positions are advanced by one

[0050] 2. if the two current characters belong to the same look-alike set, the Score is increased by 0.1 and both string positions are advanced by one

[0051] 3. if the current character in the "suspect" string is a key character in the ligature table and the current and next characters in the "candidate" string are the corresponding ligature character string, the Score is increased by 0.1 and the "suspect" string is advanced by one and the "candidate" string is advanced by two

[0052] 4. if the current character in the “candidate" string is a key character in the ligature table and the current and next characters in the “suspect” string are the corresponding ligature character string, the Score is increased by 0.1 and the “candidate” string is advanced by one and the “suspect” string is advanced by two

[0053] 5. if the current character in the “candidate” string is identical to the next character in the “suspect” string and the current character in the “suspect” string is identical to the current character in the “candidate” string, the Score is increased by 0.25 and both string positions are advanced by two

[0054] 6. if the current character in the “suspect” string is identical to the next character in the “candidate” string, the Score is increased by 0.5 and the “suspect” string position is advanced by one and the “candidate” string position is advanced by two

[0055] 7. if the current character in the “candidate” string is identical to the next character in the “suspect” string, the Score is increased by 0.5 and the “candidate” string position is advanced by one and the “suspect” string position is advanced by two

[0056] 8. otherwise, the Score is increased by 1.0 and both string positions are advanced by one

[0057] (iv) when one of the character strings is exhausted, the total number of characters remaining in the other string as added to the Score

[0058] (v) the final Similarity Score represents how visually similar the two character strings would appear on a computer screen, where two completely identical character strings would result in a Score of 0.0 and the more visually different the two character strings, the higher the Score.

[0059] A method in accordance with one or more further embodiments includes the following steps:

[0060] (a) A “suspect” text character string and a set of “corpus” text character strings are considered.

[0061] (b) The Scalar Estimate, as described above, is calculated for the “suspect” character string.

[0062] (c) The Scalar Estimate, as described above, is calculated for each of the “corpus” character strings.

[0063] (d) Of the set of "corpus" strings, a sub-set of "candidate" strings is defined such that the absolute difference between the Scalar Estimate of a "candidate" and the Scalar Estimate of the "suspect" string is less than a chosen Delta Value.

[0064] (e) For each "candidate" string, a Similarity Score, as above, is calculated for the "candidate" string and the "suspect" string.

[0065] (f) The set of "candidate" character strings is ordered by the calculated Similarity Score.

[0066] (g) Of the initial set of "corpus" character strings, the "candidate" string or strings with the lowest Similarity Scores are the most visually similar to the "suspect" string.

[0067] The techniques for identify visually similar text character strings described herein may be implemented in one or more computer programs executing on a programmable computer system. FIG. 5 is a simplified block diagram illustrating an exemplary computer system 10, on which the computer programs may operate as a set of computer instructions. The computer system 10 includes at least one computer processor 12 and a system memory 14 (including a random access memory and a read-only memory) readable by the processor 12. The computer system also includes a mass storage device 15 (e.g., a hard disk drive, a solid-state storage device, an optical disk device, etc.). The computer processor 12 is capable of processing instructions stored in the system memory or mass storage device. The computer system additionally includes input/output devices 16, 18 (e.g., a display, keyboard, or pointer device), a graphics module 20 for generating graphical objects, and a communication module or network interface 22, which manages communication with other computer systems 26 via the Internet or other communications network or combination of networks 24. Each computer program can be a set of instructions or program code in a code module resident in the random access memory of the computer system. Until required by the computer system, the set of instructions may be stored in the mass storage device or on another computer system and downloaded via the Internet or other network.

[0068] Having thus described several illustrative embodiments, it is to be appreciated that various alterations, modifications, and improvements will readily occur to those skilled in the art. Such alterations, modifications, and improvements are intended to form a part of this disclosure, and are intended to be within the spirit and scope of this

disclosure. While some examples presented herein involve specific combinations of functions or structural elements, it should be understood that those functions and elements may be combined in other ways according to the present disclosure to accomplish the same or different objectives. In particular, acts, elements, and features discussed in connection with one embodiment are not intended to be excluded from similar or other roles in other embodiments.

[0069] Additionally, elements and components described herein may be further divided into additional components or joined together to form fewer components for performing the same functions. For example, the computer system may comprise one or more physical machines, or virtual machines running on one or more physical machines. In addition, the computer system may comprise a cluster of computers or numerous distributed computers that are connected by the Internet or another network.

[0070] Accordingly, the foregoing description and attached drawings are by way of example only, and are not intended to be limiting.

[0071] What is claimed is:

CLAIMS

1. A computer-implemented method for analyzing an input text character string to identify one or more visually similar text character strings in a corpus of text character strings, the method comprising the steps, each performed by a computer system, of:

(a) receiving the input text character string;

(b) determining a scalar estimate for the input text character string based on a model of visually similar text characters stored in a computer storage system, said model comprising a plurality of look-alike sets of text characters, each set having an assigned numerical value;

(c) identifying a set of candidate text character strings from the corpus of text character strings stored in the computer storage system by comparing the scalar estimate of the input text character string with precalculated scalar estimates of each of said text character strings in the corpus, and selecting text character strings having scalar estimates within a given range of the scalar estimate of the input text character string;

(d) analyzing the input text character string with each of the candidate text character strings to determine a visual similarity score between the input text character string and each of the candidate text character strings; and

(e) identifying one or more candidate text character strings deemed to be most visually similar to the input text character string based on the similarity scores.

2. The method of claim 1, wherein the input text character string comprises a domain name in an e-mail message, and the method is used to identify fraudulent e-mail messages.

3. The method of claim 2, further comprising flagging the e-mail message when the similarity score of the one or more text character strings deemed to be most visually similar to the input text character string is within a given similarity threshold.

4. The method of claim 1, the scalar estimate of the input text character string is determined in step (b) based on the sum of the numerical value of each character in the input string, modified by its position in the string such that the numerical values of characters nearer to the start of the string are given greater weight than the numerical values of the characters nearer to the end of the string.

5. The method of claim 1, wherein the visual similarity scores determined in step (d) are based on individual character similarities, combinations of characters that look similar to individual characters, transposition of characters, and/or simple additions and deletions of characters.

6. The method of claim 1, wherein step (d) includes comparing characters of the input text character string with characters in each of the candidate text character strings using a table of look-alike sets, each set comprising visually similar characters.

7. The method of claim 1, wherein step (d) includes comparing characters of the input text character string with characters in each of the candidate text character strings using a ligature table mapping individual characters with a ligature character string having a similar visual appearance.

8. The method of claim 1, wherein determining a scalar estimate for the input text character string based on the model of visually similar text characters in step (b) comprises:

(i) discarding one or more characters in the input text character string that are identical to a preceding character;

(ii) assigning each remaining character the numerical value of the look-alike set it belongs to;

(iii) modifying the numerical value assigned to each character by multiplying the numerical value by the ratio of the distance of the character's position from the end of the input text character string and the full length of the input text character string; and

(iv) adding the modified numerical values assigned to each character to determine the scalar estimate.

9. The method of claim 1, wherein the text characters are partitioned into look-alike sets of characters that look similar, and the text characters are also partitioned into a ligature table mapping individual key characters to particular ligature character strings that look similar to the key characters,

wherein analyzing the input text character string with each of the candidate text character strings to determine a visual similarity score in step (d) comprises assigning a starting similarity score of 0.0, and starting at the first character in the input text character string and a first character in a candidate text character string, choosing for each

comparison iteration, one of the following options such that the visual similarity score is minimized:

(i) if the corresponding characters in the input text character string and the candidate text character string are identical, the similarity score is not changed and both string positions are advanced by one;

(ii) if the corresponding characters in the input text character string and the candidate text character string belong to the same look-alike set, the similarity score is increased by 0.1 and both string positions are advanced by one;

(iii) if the current character in the input text character string is a key character in the ligature table and the current and next characters in the candidate text character string are the corresponding ligature character string in the ligature table, the similarity score is increased by 0.1, the input text character string is advanced by one, and the candidate text character string is advanced by two;

(iv) if the current character in the candidate text character string is a key character in the ligature table and the current and next characters in the input text character string are the corresponding ligature character string in the ligature table, the similarity score is increased by 0.1, the candidate text character string is advanced by one, and the input text character string is advanced by two;

(v) if the current character in the candidate text character string is identical to the next character in the input text character string and the current character in the input text character string is identical to the current character in the candidate text character string, the similarity score is increased by 0.25 and both string positions are advanced by two;

(vi) if the current character in the input text character string is identical to the next character in the candidate text character string, the similarity score is increased by 0.5 and the input text character string position is advanced by one and the candidate text character string position is advanced by two;

(vii) if the current character in the candidate text character string is identical to the next character in the input text character string, the similarity score is increased by 0.5 and the candidate text character string position is advanced by one and input text character string position is advanced by two; and

(viii) otherwise, the similarity score is increased by 1.0 and both string positions are advanced by one;

wherein when one of the character strings is exhausted, the total number of characters remaining in the other string as added to the similarity score; and

wherein the final similarity score represents how visually similar the two character strings would appear on a computer screen, where two completely identical character strings would result in a similarity score of 0.0 and the more visually different the two character strings, the higher the similarity score.

10. A computer system, comprising:

at least one processor;

memory associated with the at least one processor; and

a program supported in the memory for analyzing an input text character string to identify one or more visually similar text character strings in a corpus of text character strings, the program containing a plurality of instructions which, when executed by the at least one processor, cause the at least one processor to:

(a) receive the input text character string;

(b) determine a scalar estimate for the input text character string based on a model of visually similar text characters stored in a computer storage system, said model comprising a plurality of look-alike sets of text characters, each set having an assigned numerical value;

(c) identify a set of candidate text character strings from the corpus of text character strings stored in the computer storage system by comparing the scalar estimate of the input text character string with precalculated scalar estimates of each of said text character strings in the corpus, and selecting text character strings having scalar estimates within a given range of the scalar estimate of the input text character string;

(d) analyze the input text character string with each of the candidate text character strings to determine a visual similarity score between the input text character string and each of the candidate text character strings; and

(e) identify one or more candidate text character strings deemed to be most visually similar to the input text character string based on the similarity scores.

11. The system of claim 10, wherein the input text character string comprises a domain name in an e-mail message, and the system is used to identify fraudulent e-mail messages.

12. The system of claim 11, wherein the program further comprises instructions for flagging the e-mail message when the similarity score of the one or more text character strings deemed to be most visually similar to the input text character string is within a given similarity threshold.

13. The system of claim 10, the scalar estimate of the input text character string is determined in (b) based on the sum of the numerical value of each character in the input string, modified by its position in the string such that the numerical values of characters nearer to the start of the string are given greater weight than the numerical values of the characters nearer to the end of the string.

14. The system of claim 10, wherein the visual similarity scores determined in (d) are based on individual character similarities, combinations of characters that look similar to individual characters, transposition of characters, and/or simple additions and deletions of characters.

15. The system of claim 10, wherein (d) includes comparing characters of the input text character string with characters in each of the candidate text character strings using a table of look-alike sets, each set comprising visually similar characters.

16. The system of claim 10, wherein (d) includes comparing characters of the input text character string with characters in each of the candidate text character strings using a ligature table mapping individual characters with a ligature character string having a similar visual appearance.

17. The system of claim 10, wherein determining a scalar estimate for the input text character string based on the model of visually similar text characters in (b) comprises:

(i) discarding one or more characters in the input text character string that are identical to a preceding character;

(ii) assigning each remaining character the numerical value of the look-alike set it belongs to;

(iii) modifying the numerical value assigned to each character by multiplying the numerical value by the ratio of the distance of the character's position from the end of the input text character string and the full length of the input text character string; and

(iv) adding the modified numerical values assigned to each character to determine the scalar estimate.

18. The system of claim 10, wherein the text characters are partitioned into look-alike sets of characters that look similar, and the text characters are also partitioned into a ligature table mapping individual key characters to particular ligature character strings that look similar to the key characters,

wherein analyzing the input text character string with each of the candidate text character strings to determine a visual similarity score in (d) comprises assigning a starting similarity score of 0.0, and starting at the first character in the input text character string and a first character in a candidate text character string, choosing for each comparison iteration, one of the following options such that the visual similarity score is minimized:

(i) if the corresponding characters in the input text character string and the candidate text character string are identical, the similarity score is not changed and both string positions are advanced by one;

(ii) if the corresponding characters in the input text character string and the candidate text character string belong to the same look-alike set, the similarity score is increased by 0.1 and both string positions are advanced by one;

(iii) if the current character in the input text character string is a key character in the ligature table and the current and next characters in the candidate text character string are the corresponding ligature character string in the ligature table, the similarity score is increased by 0.1, the input text character string is advanced by one, and the candidate text character string is advanced by two;

(iv) if the current character in the candidate text character string is a key character in the ligature table and the current and next characters in the input text character string are the corresponding ligature character string in the ligature table, the similarity score is increased by 0.1, the candidate text character string is advanced by one, and the input text character string is advanced by two;

(v) if the current character in the candidate text character string is identical to the next character in the input text character string and the current character in the input text character string is identical to the current character in the candidate text character string, the similarity score is increased by 0.25 and both string positions are advanced by two;

(vi) if the current character in the input text character string is identical to the next character in the candidate text character string, the similarity score is increased by 0.5 and the input text character string position is advanced by one and the candidate text character string position is advanced by two;

(vii) if the current character in the candidate text character string is identical to the next character in the input text character string, the similarity score is increased by 0.5 and the candidate text character string position is advanced by one and input text character string position is advanced by two; and

(viii) otherwise, the similarity score is increased by 1.0 and both string positions are advanced by one;

wherein when one of the character strings is exhausted, the total number of characters remaining in the other string as added to the similarity score; and

wherein the final similarity score represents how visually similar the two character strings would appear on a computer screen, where two completely identical character strings would result in a similarity score of 0.0 and the more visually different the two character strings, the higher the similarity score.

entertainment
entenainment
entertainmcnt
entertalnment

FIG. 1A

entertainment
entenainment
entertainmcnt
entertalnment

FIG. 1B

0.0: _ - . ,
1.0: 1 l i
2.0: a o u O c
3.0: g j y p q
. . .

FIG. 2

	p	a	i	l
unmodified:	3.0	2.0	1.0	1.0
modified:	3.0	1.5	0.5	0.25

Scalar Estimate: 5.25

FIG. 3

FIG. 4

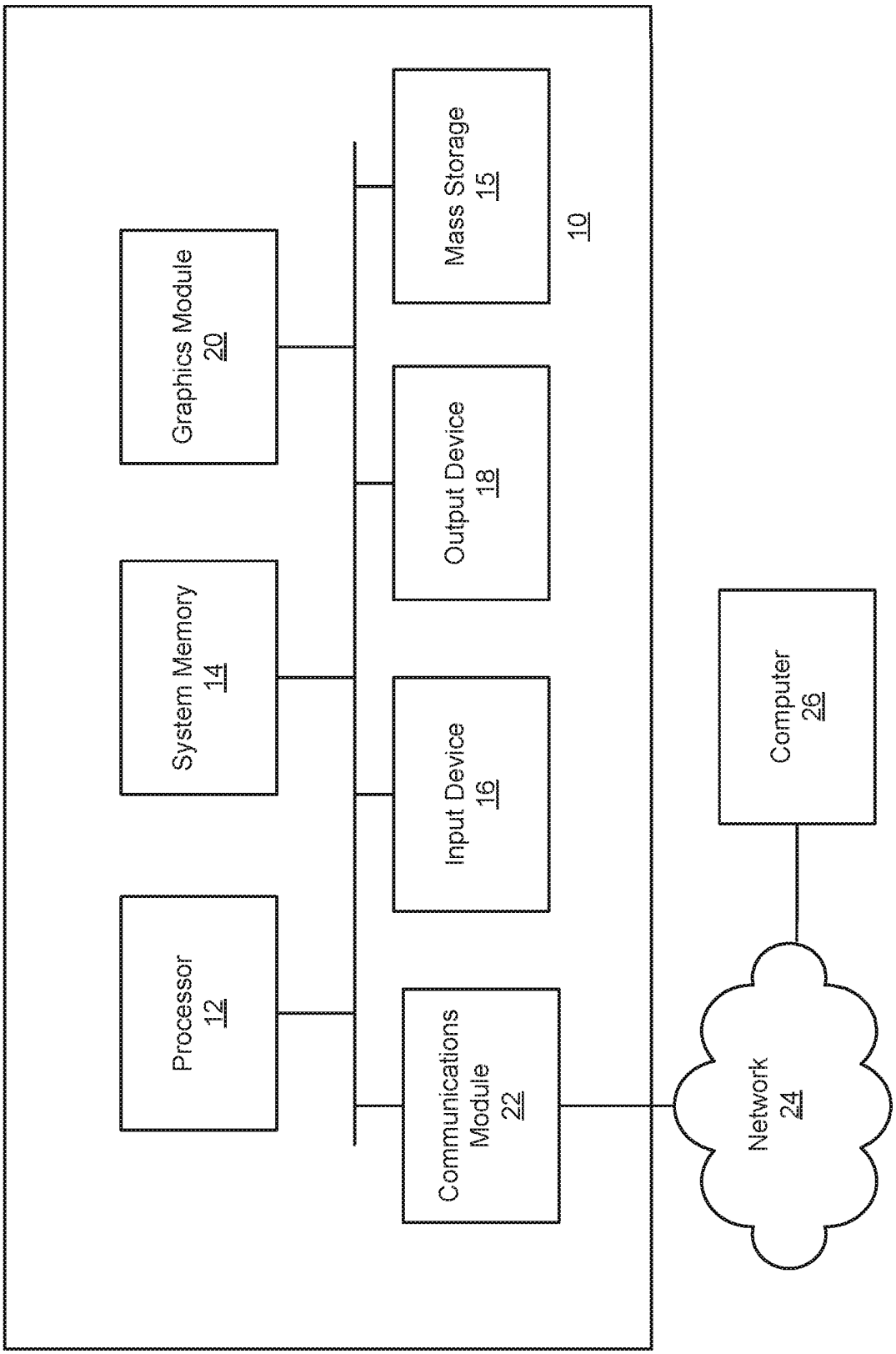

Processor
12
System Memory
14
Graphics Module
20
Communications
Module
22
Input Device
16
Output Device
18
Mass Storage
15
10
Network
24
Computer
26

FIG. 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2017/033995

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - G06F 15/16; G06F 17/30; G06K 9/18; H04L 12/58 (2017.01)

CPC - G06F 17/211; G06F 17/273; G06F 17/30303; G06F 17/30554; G06F 17/30563; G06F 17/30675; G06F 17/30705; G06F 17/30958; G06Q 30/02; G09B 17/02; H04L 51/12 (2017.02)

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)

See Search History document

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

USPC - 382/182; 707/723; 709/206 (keyword delimited)

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

See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2014/0115704 A1 (HEWLETT-PACKARD DEVELOPMENT COMPAN) 24 April 2014 (24.04.2014), entire document	1-3, 5-7, 10-12, 14-16
---		---
Y		4, 8, 9, 13, 17, 18
Y	US 2016/0004937 A1 (GENERAL ELECTRIC COMPANY) 07 January 2016 (07.01.2016), entire document	4, 8, 9, 13, 17, 18
A	US 2012/0141031 A1 (BOEGELUND) 07 June 2012 (07.06.2012), entire document	1-18
A	US 6,421,709 B1 (MCCORMICK et al) 16 July 2002 (16.07.2002), entire document	1-18
A	US 2016/0092557 A1 (ORACLE INTERNATIONAL CORPORATION) 31 March 2016 (31.03.2016), entire document	1-18
A	US 2005/0262210 A1 (YU) 24 November 2005 (24.11.2005), entire document	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	Date of mailing of the international search report
08 July 2017	04 AUG 2017
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, VA 22313-1450 Facsimile No. 571-273-8300	Authorized officer Blaine R. Copenheaver PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774