

CLAIMS

What is claimed is:

1. A method comprising:
 - employing a computer system to receive a target string forming a part of an electronic communication;
 - in response to receiving the target string, employing the computer system to determine a string eligibility criterion according to the target string;
 - employing the computer system to pre-filter a corpus of reference strings according to the string eligibility criterion, to produce a plurality of candidate strings;
 - in response to selecting the candidate strings, employing the computer system to perform a first comparison between a character histogram of the target string and a character histogram of a candidate string of the plurality of candidate strings, and a second comparison between a timestamp of the electronic communication and a timestamp of the candidate string; and
 - employing the computer system to determine whether the electronic communication is spam or non-spam according to a result of the first comparison and the second comparison.
2. The method of claim 1, wherein the corpus of reference strings comprises a plurality of clusters, each cluster including a set of mutually-similar strings, wherein each candidate string of the plurality of candidate strings is representative of a distinct cluster, and wherein the method further comprises, in response to performing the first comparison, employing the computer system to select a cluster from the plurality of clusters and to assign the target string to the selected cluster.
3. The method of claim 2, further comprising determining whether the target communication is spam or non-spam according to a plurality of timestamps, each

timestamp of the plurality of timestamps corresponding to a member of the selected cluster.

4. The method of claim 2, further comprising:

in response to assigning the target string to the selected cluster, determining a count of cluster members of the selected cluster; and
determining whether the electronic communication is spam or non-spam according to the count of cluster members.

5. The method of claim 2, further comprising identifying the electronic communication as belonging to a selected spam wave according to the selected cluster.

6. The method of claim 1, wherein pre-filtering the corpus comprises:

determining a first string length of the target string and a second string length of a reference string of the corpus, and
when the first string length differs from the second string length by an amount smaller than a predetermined threshold, adding the reference string to the set of candidate strings.

7. The method of claim 1, wherein pre-filtering the corpus comprises:

determining a first count of distinct characters of the target string and a second count of distinct characters of a reference string of the corpus, and
when the first count differs from the second count by an amount smaller than a predetermined threshold, adding the reference string to the set of candidate strings.

8. The method of claim 1, wherein pre-filtering the corpus comprises:

determining a first string score of the target string as a function of:

$$\sum_i p_i w_i$$

wherein p_i denotes the ASCII code of the i -th character of the target string, and w_i is a character-specific weight;

determining a second string score of a reference string of the corpus; and

when the first string score differs from the second string score by an amount smaller than a predetermined threshold, adding the reference string to the set of candidate strings.

9. The method of claim 1, wherein performing the first comparison comprises determining an inter-string distance as a function of:

$$\sum_{i \in T \cap C} w_i |N^i_T - N^i_C|,$$

wherein T denotes the set of characters of the target string, C denotes the set of characters of the candidate string, N^i_T denotes a count of occurrences of character i within the target string, N^i_C denotes a count of occurrences of character i within the candidate string, and wherein w_i is a character-specific weight of character i .

10. The method of claim 9, wherein the inter-string distance is further determined as a function of:

$$\sum_{j \in T - C} w_j \cdot c,$$

wherein character j occurs within the target string, but does not occur within the candidate string, w_j is a character-specific weight of character j , and c is a number selected according to a string length of the target string.

11. The method of claim 1, wherein performing the first comparison comprises determining an inter-string distance as a function of:

$$\sum_{i \in T - C} w_i \cdot c,$$

wherein T denotes the set of characters of the target string, C denotes the set of characters of the candidate string, wherein character i occurs within the target string, but does not occur within the candidate string, w_i is a character-specific weight of character i , and c is a number selected according to a string length of the target string.

12. The method of claim 1, wherein the electronic communication comprises a blog comment.
13. The method of claim 1, wherein the electronic communication comprises a message posted on a social network site.
14. A computer system comprising at least a processor programmed to:
receive a target string forming a part of an electronic communication;
in response to receiving the target string, determine a string eligibility criterion according to the target string;
pre-filter a corpus of reference strings according to the string eligibility criterion, to produce a plurality of candidate strings;
in response to selecting the candidate strings, perform a first comparison between a character histogram of the target string and a character histogram of a candidate string of the plurality of candidate strings, and a second comparison between a timestamp of the electronic communication and a timestamp of the candidate string; and
determine whether the electronic communication is spam or non-spam according to a result of the first comparison and the second comparison.
15. The system of claim 14, wherein the corpus of reference strings comprises a plurality of clusters, each cluster including a set of similar strings, wherein each candidate string of the plurality of candidate strings is representative of a distinct cluster, and wherein the processor is further programmed, in response to performing the first comparison, to select a cluster from the plurality of clusters and to assign the target string to the selected cluster.
16. The system of claim 15, further comprising determining whether the target communication is spam or non-spam according to a plurality of timestamps, each

timestamp of the plurality of timestamps corresponding to a member of the selected cluster.

17. The system of claim 15, wherein the processor is further programmed to:

in response to assigning the target string to the selected cluster, determine a count of cluster members of the selected cluster; and
determine whether the electronic communication is spam or non-spam according to the count of cluster members.

18. The system of claim 15, wherein the processor is further programmed to identify the electronic communication as belonging to a selected spam wave according to the selected cluster.

19. The system of claim 14, wherein pre-filtering the corpus comprises:

determining a first string length of the target string and a second string length of a reference string of the corpus, and
when the first string length differs from the second string length by an amount smaller than a predetermined threshold, adding the reference string to the set of candidate strings.

20. The system of claim 14, wherein pre-filtering the corpus comprises:

determining a first count of distinct characters of the target string and a second count of distinct characters of a reference string of the corpus, and
when the first count differs from the second count by an amount smaller than a predetermined threshold, adding the reference string to the set of candidate strings.

21. The system of claim 14, wherein pre-filtering the corpus comprises:

determining a first string score of the target string as a function of:

$$\sum_i p_i w_i$$

wherein p_i denotes the ASCII code of the i -th character of the target string, and w_i is a character-specific weight;

determining a second string score of a reference string of the corpus; and

when the first string score differs from the second string score by an amount smaller than a predetermined threshold, adding the reference string to the set of candidate strings.

22. The system of claim 14, wherein performing the first comparison comprises determining an inter-string distance as a function of:

$$\sum_{i \in T \cap C} w_i |N^i_T - N^i_C|,$$

wherein T denotes the set of characters of the target string, C denotes the set of characters of the candidate string, N^i_T denotes a count of occurrences of character i within the target string, N^i_C denotes a count of occurrences of character i within the candidate string, and wherein w_i is a character-specific weight of character i .

23. The system of claim 22, wherein the inter-string distance is further determined as a function of:

$$\sum_{j \in T - C} w_j \cdot c,$$

wherein character j occurs within the target string, but does not occur within the candidate string, w_j is a character-specific weight of character j , and c is a number selected according to a string length of the target string.

24. The system of claim 14, wherein performing the first comparison comprises determining an inter-string distance as a function of:

$$\sum_{i \in T - C} w_i \cdot c,$$

wherein T denotes the set of characters of the target string, C denotes the set of characters of the candidate string, wherein character i occurs within the target string, but does not occur within the candidate string, w_i is a character-specific weight of character i , and c is a number selected according to a string length of the target string.

25. The system of claim 14, wherein the electronic communication comprises a blog comment.
26. The system of claim 14, wherein the electronic communication comprises a message posted on a social network site.
27. A method comprising:
- employing a computer system to receive an electronic communication;
 - in response to receiving the electronic communication, employing the computer system to extract a target string from the electronic communication;
 - employing the computer system to transmit the target string to an anti-spam server; and
 - in response to transmitting the target string, receiving a target label indicative of whether the electronic communication is spam or non-spam, wherein the target label is determined at the anti-spam server and wherein determining the target label comprises:
 - employing the anti-spam server to determine an eligibility criterion according to the target string;
 - employing the anti-spam server to pre-filter a corpus of reference strings according to the criterion condition, to produce a plurality of candidate strings;
 - in response to selecting the candidate strings, employing the anti-spam server to perform a first comparison between a character histogram of the target string and a character histogram of a candidate string of the plurality of candidate strings, and a second comparison between a timestamp of the electronic communication and a timestamp of the candidate string; and
 - employing the anti-spam server to determine the target label according to a result of the first comparison and the second comparison.

28. A method comprising:

employing a computer system to receive a target string forming a part of an electronic communication;

in response to receiving the target string, employing the computer system to determine a string eligibility criterion according to the target string;

employing the computer system to pre-filter a corpus of reference strings according to the string eligibility criterion, to produce a plurality of candidate strings;

in response to selecting the candidate strings, employing the computer system to determine an inter-string distance separating the target string from a candidate string of the plurality of candidate strings, the inter-string distance determined according to a count of occurrences of a selected character within the target string and a count of occurrences of the selected character within the candidate string;
and

employing the computer system to determine whether the electronic communication is spam or non-spam according to the inter-string distance.

CLAIMS

1. A method comprising:

in response to receiving a target string forming a part of an electronic communication, employing at least one processor of a computer system to select a plurality of candidate strings from a corpus of reference strings, wherein selecting the plurality of candidate strings comprises:

comparing a string length of the target string to a string length of a reference string of the corpus, and

in response, selecting the reference string into the plurality of candidate strings according to a result of the comparison of string lengths;

in response to selecting the plurality of candidate strings, employing the at least one processor to perform a first comparison between the target string and a candidate string of the plurality of candidate strings, and a second comparison between the target string and the candidate string; and

employing the at least one processor to determine whether the electronic communication is spam or non-spam according to a result of the first comparison and the second comparison,

wherein the first comparison comprises comparing, for each character of a plurality of distinct alphanumeric characters, a count of occurrences of the each character within the target string to a count of occurrences of the each character within the reference string, wherein the count of occurrences of the each character within the target string is determined without regard to a position of the each character relative to other characters within the target string, and

wherein the second comparison comprises comparing a timestamp of the electronic communication to a timestamp of another electronic communication, the another electronic communication containing the candidate string.

2. The method of claim 1, wherein the corpus of reference strings comprises a plurality of clusters, each cluster including a set of mutually-similar strings, wherein each candidate string of the plurality of candidate strings is representative of a distinct cluster, and wherein the method further comprises, in response to performing the first comparison, employing the computer system to select a cluster from the plurality of clusters and to assign the target string to the selected cluster.

3. The method of claim 2, further comprising determining whether the target communication is spam or non-spam according to a plurality of timestamps, each timestamp of the plurality of timestamps corresponding to a member of the selected cluster.

4. The method of claim 2, further comprising:

in response to assigning the target string to the selected cluster, determining a count of cluster members of the selected cluster; and

determining whether the electronic communication is spam or non-spam according to the count of cluster members.

5. The method of claim 2, further comprising identifying the electronic communication as belonging to a selected spam wave according to the selected cluster.

6. The method of claim 1, wherein selecting the plurality of candidate strings further comprises:

determining a first count of distinct characters of the target string and a second count of distinct characters of the reference string, and

when the first count differs from the second count by an amount smaller than a predetermined threshold, selecting the reference string into the plurality of candidate strings.

7. The method of claim 1, wherein selecting the plurality of candidate strings further comprises:

determining a first string score of the target string as a function of:

$$\sum_i p_i w_i$$

wherein p_i denotes an ASCII code of the i -th character of the target string, and w_i is a character-specific weight;

determining a second string score of the reference string; and

when the first string score differs from the second string score by an amount smaller than a predetermined threshold, selecting the reference string into the plurality of candidate strings.

8. The method of claim 1, wherein performing the first comparison comprises determining an inter-string distance as a function of:

$$\sum_{i \in T \cap C} w_i |N_T^i - N_C^i|,$$

wherein T denotes the set of characters of the target string, C denotes the set of characters of the candidate string, N_T^i denotes a count of occurrences of character i within the target string, N_C^i denotes a count of occurrences of character i within the candidate string, and wherein w_i is a character-specific weight of character i .

9. The method of claim 8, wherein the inter-string distance is further determined as a function of:

$$\sum_{j \in T-C} w_j \cdot c,$$

wherein character j occurs within the target string, but does not occur within the candidate string, w_j is a character-specific weight of character j , and c is a number selected according to the string length of the target string.

10. The method of claim 1, wherein performing the first comparison comprises determining an inter-string distance as a function of:

$$\sum_{i \in T-C} w_i \cdot c,$$

wherein T denotes the set of characters of the target string, C denotes the set of characters of the candidate string, wherein character i occurs within the target string, but does not occur within the candidate string, w_i is a character-specific weight of character i , and c is a number selected according to the string length of the target string.

11. The method of claim 1, wherein the electronic communication comprises a blog comment.

12. The method of claim 1, wherein the electronic communication comprises a message posted on a social network site.

13. A computer system comprising at least one processor programmed to:

in response to receiving a target string forming part of an electronic communication, select a plurality of candidate strings from a corpus of reference strings, wherein selecting the plurality of candidate strings comprises:

comparing a string length of the target string to a string length of a reference string of the corpus, and

in response, selecting the reference string into the plurality of candidate strings according to a result of the comparison of string lengths;

in response to selecting the candidate strings, perform a first comparison between the target string and a candidate string of the plurality of candidate strings, and a second comparison between the target string and the candidate string; and

determine whether the electronic communication is spam or non-spam according to a result of the first comparison and the second comparison,

wherein the first comparison comprises comparing, for each character of a plurality of distinct alphanumeric characters, a count of occurrences of the each character within the target string to a count of occurrences of the each character within the reference string, wherein the count of occurrences of the each character within the target string is determined without regard to a position of the each character relative to other characters within the target string, and

wherein the second comparison comprises comparing a timestamp of the electronic communication to a timestamp of a another electronic communication containing the candidate string.

14. The computer system of claim 13, wherein the corpus of reference strings comprises a plurality of clusters, each cluster including a set of similar strings, wherein each candidate string of the plurality of candidate strings is representative of a distinct cluster, and wherein the processor is further programmed, in response to performing the first comparison, to select a cluster from the plurality of clusters and to assign the target string to the selected cluster.

15. The computer system of claim 14, wherein the at least one processor is further programmed to determine whether the target communication is spam or non-spam according to a plurality of timestamps, each timestamp of the plurality of timestamps corresponding to a member of the selected cluster.

16. The computer system of claim 14, wherein the at least one processor is further programmed to:

in response to assigning the target string to the selected cluster, determine a count of cluster members of the selected cluster; and

determine whether the electronic communication is spam or non-spam according to the count of cluster members.

17. The computer system of claim 14, wherein the at least one processor is further programmed to identify the electronic communication as belonging to a selected spam wave according to the selected cluster.

18. The computer system of claim 13, wherein selecting the plurality of candidate strings further comprises:

determining a first count of distinct characters of the target string and a second count of distinct characters of the reference string, and

when the first count differs from the second count by an amount smaller than a predetermined threshold, selecting the reference string into the plurality of candidate strings.

19. The computer system of claim 13, wherein selecting the plurality of candidate strings comprises: determining a first string score of the target string as a function of:

$$\sum_i p_i w_i,$$

wherein p_i denotes an ASCII code of the i -th character of the target string, and w_i is a character-specific weight;

determining a second string score of the reference string; and

when the first string score differs from the second string score by an amount smaller than a predetermined threshold, selecting the reference string into the plurality of candidate strings.

20. The computer system of claim 13, wherein performing the first comparison comprises determining an inter-string distance as a function of:

$$\sum_{i \in T \cap C} w_i |N_T^i - N_C^i|,$$

wherein T denotes the set of characters of the target string, C denotes the set of characters of the candidate string, N_T^i denotes a count of occurrences of character i within the target string, N_C^i denotes a count of occurrences of character i within the candidate string, and wherein w_i is a character-specific weight of character i .

21. The computer system of claim 20, wherein the inter-string distance is further determined as a function of.

$$\sum_{j \in T - C} w_j \cdot c,$$

wherein character j occurs within the target string, but does not occur within the candidate string, w_j is a character-specific weight of character j , and c is a number selected according to the string length of the target string.

22. The computer system of claim 13, wherein performing the first comparison comprises determining an inter-string distance as a function of:

$$\sum_{i \in T-C} w_i \cdot c,$$

wherein T denotes the set of characters of the target string, C denotes the set of characters of the candidate string, wherein character i occurs within the target string, but does not occur within the candidate string, w_i is a character-specific weight of character i , and c is a number selected according to the string length of the target string.

23. The computer system of claim 13, wherein the electronic communication comprises a blog comment.

24. The computer system of claim 13, wherein the electronic communication comprises a message posted on a social network site.

25. A method comprising:

employing at least one processor of a computer system to receive an electronic communication;

in response to receiving the electronic communication, employing the at least one processor to extract a target string from the electronic communication;

employing the at least one processor to transmit the target string to an anti-spam server; and

in response to transmitting the target string, employing the at least one processor to receive a target label indicative of whether the electronic communication is spam or non-spam, wherein the target label is determined at the anti-spam server and wherein determining the target label comprises:

employing the anti-spam server to select a plurality of candidate strings from a corpus of reference strings, wherein selecting the plurality of candidate strings comprises:

comparing a string length of the target string to a string length of a reference string of the corpus, and

in response, selecting the reference string into the plurality of candidate strings according to a result of the comparison of string lengths;

in response to selecting the candidate strings, employing the anti-spam server to perform a first comparison between the target string and a candidate string of the plurality of candidate strings, and a second comparison between the target string and the candidate string; and

employing the anti-spam server to determine the target label according to a result of the first comparison and the second comparison,

wherein the first comparison comprises comparing, for each character of a plurality of distinct alphanumeric characters, a count of occurrences of the each character within the target string to a count of occurrences of the each character within the reference string, wherein the count of occurrences of the each character within the target string is determined without regard to a position of the each character relative to other characters within the target string, and

wherein the second comparison comprises comparing a timestamp of the electronic communication to a timestamp of another electronic communication containing the candidate string.

26. A method comprising:

in response to receiving a target string forming part of an electronic communication, employing at least one processor of a computer system to select a plurality of candidate strings from a corpus of reference strings, wherein selecting the plurality of candidate strings comprises:

comparing a string length of the target string to a string length of a reference string of the corpus, and

in response, select the reference string into the plurality of candidate strings according to a result of the comparison of string lengths;

in response to selecting the candidate strings, employing the at least one processor to determine an inter-string distance separating the target string from a candidate string of the plurality of candidate strings, the inter-string distance determined according to a count of occurrences within the target string of each character of a plurality of distinct alphanumeric characters, and according to a count of occurrences of the each character within the candidate string, wherein the count of occurrences of the each character within the target string is determined without regard to a position of the each character relative to other characters within the target string; and

employing the at least one processor to determine whether the electronic communication is spam or non-spam according to the inter-string distance.