

CLAIMS

What is claimed is:

1. A method of text string embedding implemented by a text string embedding device, the method comprising:

accessing, by the text string embedding device, stored history data comprising text strings and device identifiers, wherein each of the device identifiers is associated with one or more of the text strings; and

generating, by the text string embedding device, an embedding look-up table comprising a plurality of keys and a plurality of values, wherein the generating the embedding look-up table comprises:

splitting the text strings in the history data into a plurality of words;

generating a graph with the plurality of words and the device identifiers

by:

denoting each word in the plurality of words as a node of a first type;

denoting each of the device identifiers as a node of a second type;

and

for each node of the second type, building edges by linking the node of the second type denoting one of the device identifiers to ones of the nodes of the first type, the linked ones of the nodes of the first type denoting words that are comprised in one or more of the text strings associated with the one of the device identifiers;

training an embedding model using the generated graph; and

constructing the embedding look-up table based on the trained embedding model, wherein each of the plurality of keys of the embedding look-up table comprises one of the nodes of the first type and the nodes of the second type, and wherein each of the plurality of values of the embedding look-up table comprises a vector corresponding to the one of the nodes of the first type and the nodes of the second type.

2. The method according to claim 1, wherein the training of the embedding model using the generated graph comprises:

for each node in the generated graph, computing the vector based on information of the node's neighbours, wherein the node's neighbours comprise one or more of nodes linked to the node within a predetermined number of edges;

applying random walk on all the nodes in the generated graph to generate node sequences;

for every node in a node sequence of the generated node sequences, predicting the node's neighbours in the node sequence to form a predicted node distribution; and

training the embedding model based on the predicted node distribution and a true node distribution.

3. The method according to claim 1 or 2, wherein the text strings comprise addresses, and wherein the plurality of words comprise a plurality of address words.

4. The method according to claim 3, wherein when the node's neighbours comprises one or more nodes of the second type, the information of the node's neighbours includes geographic information corresponding to one or more of the device identifiers denoted by the one or more of the nodes of the second type.

5. The method according to claim 2, wherein the computing of the vector for each node in the generated graph further comprises:

if the node is a node of the first type denoting a word that has an appearance frequency lower than a threshold, computing the vector for the node as a vector for unknown words.

6. The method according to any one of claims 1 to 5, wherein the embedding model comprises a general attributed multiplex heterogeneous network embedding (GATNE) model, a hierarchical attention network (HAN) model, or a heterogeneous graph neural network (HetGNN) model.

7. A method of text string authentication implemented by a text string authentication device, the method comprising:

receiving, by the text string authentication device, a first text string from a user;

splitting, by the text string authentication device, the first text string into a first plurality of words; and

authenticating, by the text string authentication device, the first text string using one or more values in an embedding look-up table generated according to any one of claims 1 to 6 associated with keys in the embedding look-up table corresponding to each of the first plurality of words of the first text string.

8. The method according to claim 7, wherein the authenticating of the first text string comprises:

fetching one or more values in the embedding look-up table associated with keys in the embedding look-up table corresponding to each of the first plurality of words of the first text string, wherein any unknown word in the first plurality of words of the first text string is mapped to a key for unknown words having a vector for unknown words as a value; and

generating a vector for the first text string by a sum calculation or an average calculation of the fetched values of the first plurality of words.

9. The method according to claim 8, wherein the authenticating of the first text string further comprises:

receiving a second text string of the user from an official source;

splitting the second text string into a second plurality of words;

fetching one or more values in the embedding look-up table associated with keys in the embedding look-up table corresponding to each of the second plurality of words of the second text string, wherein any unknown word in the second plurality of words of the second text string is mapped to the key for unknown words having the vector for unknown words as a value; and

generating a vector for the second text string by a sum calculation or an average calculation of the fetched values of the second plurality of words.

10. The method according to claim 9, further comprising:

authenticating the first text string based on a similarity between the vector for the first text string and the vector for the second text string.

11. A text string embedding device comprising:

a first storage device for storing history data comprising text strings and device identifiers, wherein each of the device identifiers is associated with one or more of the text strings;

a training device coupled to the first storage device for generating an embedding look-up table comprising a plurality of keys and a plurality of values, the training device being configured to:

access the stored history data;

generate the embedding look-up table, wherein the generating the embedding look-up table comprises:

splitting the text strings in the stored history data into a plurality of words;

generating a graph with the plurality of words and the device identifiers by:

denoting each word in the plurality of words as a node of a first type;

denoting each of the device identifiers as a node of a second type; and

for each node of the second type, building edges by linking the node of the second type denoting one of the device identifiers to ones of the nodes of the first type, the linked ones of the nodes of the first type denoting words that are comprised in one or more of the text strings associated with the one of the device identifiers;

training an embedding model using the generated graph; and

constructing the embedding look-up table based on the trained embedding model, wherein each of the plurality of keys of the embedding look-up table comprises one of the nodes of the first type and the nodes of the second type, and wherein each of the plurality of values of the embedding look-up table comprises a vector corresponding to the one of the nodes of the first type and the nodes of the second type; and

a second storage device for storing the embedding look-up table.

12. The device according to claim 11, wherein when training the embedding model using the generated graph, the training device is configured to:

for each node in the generated graph, compute the vector based on information of the node's neighbours, wherein the node's neighbours comprise one or more of nodes linked to the node within a predetermined number of edges;

apply random walk on all the nodes in the generated graph to generate node sequences;

for every node in a node sequence of the generated node sequences, predict the node's neighbours in the node sequence to form a predicted node distribution; and

train the embedding model based on the predicted node distribution and a true node distribution.

13. The device according to claim 11 or 12, wherein the text strings comprise addresses, and wherein the plurality of words comprise a plurality of address words.

14. The device according to claim 13, wherein when the node's neighbours comprises one or more nodes of the second type, the information of the node's neighbours includes geographic information corresponding to one or more of the device identifiers denoted by the one or more of the nodes of the second type.

15. The device according to claim 12, wherein the training device when computing the vector for each node in the generated graph is further configured to:

if the node is a node of the first type denoting a word that has an appearance frequency lower than a threshold, compute the vector for the node as a vector for unknown words.

16. The device according to any one of claims 11 to 15, wherein the embedding model comprises a general attributed multiplex heterogeneous network embedding (GATNE) model, a hierarchical attention network (HAN) model, or a heterogeneous graph neural network (HetGNN) model.

17. A text string authentication device coupled to a text string embedding device according to any one of claims 11 to 16, the text string embedding device including an embedding look-up table generated according to any one of claims 1 to 6, wherein the text string authentication device is configured to:

receive a first text string;

split the first text string into a first plurality of words; and

authenticate the first text string using one or more values in the embedding look-up table associated with keys in the embedding look-up table corresponding to each of the first plurality of words of the first text string.

18. The device according to claim 17, wherein the text string authentication device is configured to:

fetch one or more values in the embedding look-up table associated with keys in the embedding look-up table corresponding to each of the first plurality of words of the first text string, wherein any unknown word in the first plurality of words of the first text string is mapped to a key for unknown words having a vector for unknown words as a value; and

generate a vector for the first text string by a sum calculation or an average calculation of the fetched values of the first plurality of words.

19. The device according to claim 18, wherein the text string authentication device is further configured to:

- receive a second text string of the user from an official source;

- split the second text string into a second plurality of words;

- fetch one or more values in the embedding look-up table associated with keys in the embedding look-up table corresponding to each of the second plurality of words of the second text string, wherein any unknown word in the second plurality of words of the second text string is mapped to the key for unknown words having the vector for unknown words as a value; and

- generate a vector for the second text string by a sum calculation or an average calculation of the fetched values of the second plurality of words.

20. The device according to claim 19, wherein the text string authentication device is further configured to:

- authenticate the first text string based on a similarity between the vector for the first text string and the vector for the second text string.