CLAIMS

1. A method for determining feature words, comprising:

obtaining text data;

automatically extracting a first feature word from the text data;

updating a word segmentation library by including the first feature word as a determined feature word to obtain an updated word segmentation library, the word segmentation library including a first plurality of determined feature words for representing a plurality of attribute types before the updating, the first feature word being different from any one of the first plurality of determined feature words before the updating;

automatically extracting a second feature word from the same text data based on the updated word segmentation library that includes a second plurality of determined feature words including the first feature word and the first plurality of determined feature words; and

classifying the text data based on the second feature word,

wherein the extracting the first feature word from the text data includes:

selecting, from the text data, one or more character strings each having a character length less than a length threshold as one or more candidate character strings;

for each candidate character string of the one or more character strings,

calculating a first probability of the candidate character string occurring in the text data;

determining one or more split combinations of the candidate character string;

calculating a second probability of a split combination of the one or more split combinations occurring in the text data;

determining a coagulation degree of the candidate character string based on a ratio between the first probability and the second probability;

calculating an index parameter of each of the candidate character string, the index parameter including the coagulation degree, a higher value of the coagulation degree indicating a higher probability of the candidate character string being the first feature word; and

determining, based on the index parameter, a candidate character string meeting a first threshold from the one or more candidate character strings as the first feature word.

2. The method according to claim 1, wherein the text data includes one or more of a transaction postscript or a text label.

3. The method according to claim 1, wherein the index parameter includes a frequency.

4. The method according to claim 1, further comprising:

before the selecting, from the text data, the one or more character strings each having a character length less than the length threshold as one or more candidate character strings, filtering out invalid character strings from the text data.

5. The method according to claim 1, wherein the extracting the second feature word from the text data based on the updated word segmentation library includes:

performing word segmentation processing on the text data based on the updated word segmentation library to obtain a plurality of word segmentation units;

performing word vectorization processing on each of the plurality of word segmentation units to obtain word vectors of word segment units of the plurality of word segmentation units; and

determining, based on the word vectors of the word segmentation units and word vectors of the plurality of determined feature words, a word segmentation unit meeting a second threshold among the plurality of word segmentation units as the second feature word.

6. The method according to claim 5, wherein the text data includes a Chinese character, and the performing word vectorization processing on each of the plurality of word segmentation units to obtain the word vectors of word segmentation units of the plurality of word segmentation units includes:

splitting the Chinese character in a word segmentation unit into a plurality of strokes;

obtaining a stroke vector of the word segmentation unit based on the plurality of strokes of the word segmentation unit;

obtaining a word connected to the word segmentation unit from the text data as a context word;

obtaining a word vector of the context word; and

determining a word vector of the word segmentation unit based on the stroke vector of the word segmentation unit and the word vector of the context word.

7. The method according to claim 5, wherein the determining, based on the word vectors of the word segmentation units and word vectors of the plurality of determined feature words, the word segmentation unit meeting the second threshold among the plurality of word segmentation units as the second feature word includes:

extracting a determined number of determined feature words from the plurality of determined feature words as test words;

establishing a labeled sample set based on word vectors of determined feature words of the plurality of predetermined feature words except for the test words;

establishing an unlabeled sample set based on the word vectors of the test words and the word vectors of the word segmentation units;

determining a fitting score threshold through iterative fitting based on the labeled sample set and the unlabeled sample set; and

determining, based on the fitting score threshold, a word segmentation unit meeting the second threshold among the word segmentation units as the second feature word.

8. The method according to claim 5, further comprising:

before the determining, based on the word vectors of the word segmentation units and word vectors of the plurality of determined feature words, the word segmentation unit meeting the second threshold among the plurality of word segmentation units as the second feature word, filtering out word vectors of stop words from the word vectors of the word segmentation units.

9. A server, comprising a processor and a memory for storing instructions executable by the processor, wherein when executing the instructions the processor implements acts including:

obtaining text data;

automatically extracting a first feature word from the text data;

updating a word segmentation library by including the first feature word as a determined feature word to obtain an updated word segmentation library, the word segmentation library including a first plurality of determined feature words for representing a plurality of attribute types before the updating, the first feature being different from any one of the first plurality of determined feature work before the updating;

automatically extracting a second feature word from the same text data based on the updated word segmentation library that includes a second plurality of determined feature words including the first feature word and the first plurality of determined feature words; and

classifying the text data based on the second feature word,

wherein the extracting the first feature word from the text data includes:

selecting, from the text data, one or more character strings each having a character length less than a length threshold as one or more candidate character strings;

for each candidate character string of the one or more character strings,

calculating a first probability of the candidate character string occurring in the text data;

determining one or more split combinations of the candidate character string;

calculating a second probability of a split combination of the one or more split combinations occurring in the text data;

determining a coagulation degree of the candidate character string based on a ratio between the first probability and the second probability;

calculating an index parameter of each of the candidate character string, the index parameter including the coagulation degree, a higher value of the coagulation degree indicating a higher probability of the candidate character string being the first feature word; and

determining, based on the index parameter, a candidate character string meeting a first threshold from the one or more candidate character strings as the first feature word.

10. The sever according to claim 9, wherein the text data includes one or more of a transaction postscript or a text label.

11. The sever according to claim 9, wherein the index parameter includes a frequency.

12. The sever according to claim 9, wherein the acts include:

before the selecting, from the text data, the one or more character strings each having a character length less than the length threshold as one or more candidate character strings, filtering out invalid character strings from the text data.

13. The sever according to claim 9, wherein the extracting the second feature word from the text data based on the updated word segmentation library includes:

performing word segmentation processing on the text data based on the updated word segmentation library to obtain a plurality of word segmentation units;

performing word vectorization processing on each of the plurality of word segmentation units to obtain word vectors of word segment units of the plurality of word segmentation units; and

determining, based on the word vectors of the word segmentation units and word vectors of the plurality of determined feature words, a word segmentation unit meeting a second threshold among the plurality of word segmentation units as the second feature word.

14. The sever according to claim 13, wherein the text data includes a Chinese character, and the performing word vectorization processing on each of the plurality of word segmentation units to obtain the word vectors of word segmentation units of the plurality of word segmentation units includes:

splitting the Chinese character in a word segmentation unit into a plurality of strokes;

obtaining a stroke vector of the word segmentation unit based on the plurality of strokes of the word segmentation unit;

obtaining a word connected to the word segmentation unit from the text data as a context word;

obtaining a word vector of the context word; and

determining a word vector of the word segmentation unit based on the stroke vector of the word segmentation unit and the word vector of the context word.

15. The sever according to claim 13, wherein the determining, based on the word vectors of the word segmentation units and word vectors of the plurality of determined feature words, the word segmentation unit meeting the second threshold among the plurality of word segmentation units as the second feature word includes:

extracting a determined number of determined feature words from the plurality of determined feature words as test words;

establishing a labeled sample set based on word vectors of determined feature words of the plurality of predetermined feature words except for the test words;

establishing an unlabeled sample set based on the word vectors of the test words and the word vectors of the word segmentation units;

determining a fitting score threshold through iterative fitting based on the labeled sample set and the unlabeled sample set; and

determining, based on the fitting score threshold, a word segmentation unit meeting the second threshold among the word segmentation units as the second feature word.

16. The sever according to claim 13, wherein the acts include:

before the determining, based on the word vectors of the word segmentation units and word vectors of the plurality of determined feature words, the word segmentation unit meeting the second threshold among the plurality of word segmentation units as the second feature word, filtering out word vectors of stop words from the word vectors of the word segmentation units.

17. A non-transitory computer-readable storage medium storing computer instructions, which when executed by a processor, configures the processor to implement acts comprising:

obtaining text data;

automatically extracting a first feature word from the text data;

updating a word segmentation library by including the first feature word as a determined feature word to obtain an updated word segmentation library, the word segmentation library including a first plurality of determined feature words for representing a plurality of attribute types before the updating, the first feature word being different from any one of the first plurality of determined feature words before the updating;

automatically extracting a second feature word from the same text data based on the updated word segmentation library that includes a second plurality of determined feature words including the first feature word and the first plurality of determined feature words; and

classifying the text data based on the second feature word,

wherein the extracting the first feature word from the text data includes:

selecting, from the text data, one or more character strings each having a character length less than a length threshold as one or more candidate character strings;

for each candidate character string of the one or more character strings,

calculating a first probability of the candidate character string occurring in the text data;

determining one or more split combinations of the candidate character string;

calculating a second probability of a split combination of the one or more split combinations occurring in the text data;

determining a coagulation degree of the candidate character string based on a ratio between the first probability and the second probability;

calculating an index parameter of each of the candidate character string, the index parameter including the coagulation degree, a higher value of the coagulation degree indicating a higher probability of the candidate character string being the first feature word; and

determining, based on the index parameter, a candidate character string meeting a first threshold from the one or more candidate character strings as the first feature word.

18. The non-transitory computer-readable storage medium according to claim 17, wherein the extracting the second feature word from the text data based on the updated word segmentation library includes:

performing word segmentation processing on the text data based on the updated word segmentation library to obtain a plurality of word segmentation units;

performing word vectorization processing on each of the plurality of word segmentation units to obtain word vectors of word segment units of the plurality of word segmentation units; and

determining, based on the word vectors of the word segmentation units and word vectors of the plurality of determined feature words, a word segmentation unit meeting a second threshold among the plurality of word segmentation units as the second feature word.

19. The non-transitory computer-readable storage medium according to claim 17, wherein the index parameter includes a frequency.

20. The non-transitory computer-readable storage medium according to claim 17, wherein the acts include:

before the selecting, from the text data, the one or more character strings each having a character length less than the length threshold as one or more candidate character strings, filtering out invalid character strings from the text data.