

WHAT IS CLAIMED IS:

1. A method for adding hyperlink for text, comprising:

creating a hyperlink word list in advance, the hyperlink word list comprising a plurality of hyperlink words;

collecting a variety of texts, and generating a characteristic word list by implementing word segmentation processing for each of the texts, the characteristic word list comprising a plurality of characteristic words;

respectively determining an IDF (inverse document frequency) value for each characteristic word after generating a characteristic word list by implementing word segmentation processing for each of the texts, wherein the IDF value is calculated by following processes: obtaining a quotient by a quantity of the variety of texts collected divided by a quantity of texts appearing the characteristic word, and calculating a logarithm of the quotient;

for each of the characteristic words, computing a co-occurrence frequency between each of the characteristic words and each of the hyperlink words;

considering each text to be added a hyperlink as a text X, and processing the text X by following steps:

carrying out the word segmentation processing to the text X, and obtaining a segmentation result;

extracting the hyperlink words occurred in the hyperlink word list and the characteristic words occurred in the characteristic word list from the segmentation result;

computing a weight of each of the hyperlink words that are occurred in the hyperlink word list, and computing a weight of each of the characteristic words that are occurred in the characteristic word list;

determining a final weight of each of the hyperlink words according to each co-occurrence frequency and the weights of the hyperlink words;

descendingly sorting the hyperlink words occurred in the hyperlink word list according to the final weights of the hyperlink words, and obtaining K numbers of hyperlink words that are arranged in first; and

adding hyperlinks for the K numbers of hyperlink words, wherein K is a positive integer.

2. The method according to claim 1, the step of generating a characteristic word list by implementing word segmentation processing for each of the texts, comprising:

extracting all unrepeated words obtained by the word segmentation processing as the characteristic words; or

removing high-frequency words, stop words and low-frequency words from all the unrepeated words obtained by the word segmentation processing, and considering remaining words as the characteristic words.

3. The method according to claim 1, the step of for each of the characteristic words, computing a co-occurrence frequency between each of the characteristic words and each of the hyperlink words, comprising:

for each characteristic word y and each hyperlink word x, calculating the co-occurrence frequency $P(x/y)$ using a formula of:

$P(x/y) = \text{a quantity of } xy \text{ co-occurrences} / \text{a quantity of } y \text{ occurrences};$

wherein, the quantity of xy co-occurrences represents the number of texts appearing the characteristic words y and the hyperlink words x in the same time; the quantity of y occurrences represents the number of texts appearing the characteristic words y;

or,

for each characteristic word y and each hyperlink word x, calculating the co-occurrence frequency $P(x/y)$ using a formula of:

$P(x/y) = H(x,y) / I(x,y) = H(x,y) / (H(x) + H(y) - H(x,y));$

wherein, H represents information entropy, I represents mutual information.

4. The method according to claim 3, ~~after the step of generating a characteristic word list by implementing word segmentation processing for each of the texts,~~ further comprising:

~~respectively determining an IDF (inverse document frequency) value for each characteristic word, wherein the IDF value is calculated by following processes: obtaining a quotient by a quantity of the variety of texts collected divided by a quantity of texts appearing the characteristic word, and calculating a logarithm of the quotient;~~

the step of computing a weight of each of the hyperlink words that are occurred in the hyperlink word list and computing a weight of each of the characteristic words that are occurred in the characteristic word list, comprising:

for each hyperlink word H, calculating the weight W_H of the hyperlink word H:

$$W_H = TF_H * IDF_H;$$

wherein, TF_H represents TF (term frequency) value of the hyperlink word H, that is, TF_H refers to the quantity of the hyperlink word H appearing in the text X, and IDF_H represents IDF value of the hyperlink word H;

for each characteristic word F, calculating the weight W_F of the characteristic word F:

$$W_F = TFF * IDFF;$$

wherein, TFF represents TF value of the characteristic word F, and $IDFF$ represents IDF value of the characteristic word F.

5. The method according to claim 4, the step of determining a final weight of each of the hyperlink words according to each co-occurrence frequency and the weights of the hyperlink words, comprising:

for each hyperlink word H, calculating the final weight W_H' of the hyperlink word H:

$$W_H' = W_H * \sum_{i=1}^n P(H/F_i) * W_{F_i};$$

wherein, n represents a quantity of the characteristic words.

6. A device for adding hyperlink to text, comprising:

a preprocessing module, configured to create a hyperlink word list in advance, collect a variety of texts, generate a characteristic word list by implementing word

segmentation processing for each of the texts, and for each of the characteristic words computer a co-occurrence frequency between each characteristic word list and each hyperlink word, wherein the hyperlink word list comprising a plurality of hyperlink words and the characteristic word list comprising a plurality of characteristic words;

the preprocessing module, further configured to respectively determine an IDF (inverse document frequency) value for each characteristic word, wherein the IDF value is calculated by following processes: obtaining a quotient by a quantity of the variety of texts collected divided by a quantity of texts appearing the characteristic words, and calculating a logarithm of the quotient;

an adding module, configured to consider each text to be added a hyperlink as a text X, and process the text by following steps:

carrying out the word segmentation processing to the text X, and obtaining a segmentation result;

extracting the hyperlink words occurred in the hyperlink word list and the characteristic words occurred in the characteristic word list from the segmentation result;

computing a weight of each of the hyperlink words that are occurred in the hyperlink word list, and computing a weight of each of the characteristic words that are occurred in the characteristic word list;

determining a final weight of each of the hyperlink words according to each co-occurrence frequency and the weights of the hyperlink words;

descendingly sorting the hyperlink words occurred in the hyperlink word list according to the final weights of the hyperlink words, and obtaining K numbers of hyperlink words that are arranged in first; and

adding hyperlinks for the K numbers of hyperlink words, wherein, K is a positive integer.

7. The device according to claim 6, wherein the preprocessing module, comprises:

a first processing unit, configured to create a hyperlink word list;

a second processing unit, configured to collect a variety of texts, generate a characteristic word list by implementing word segmentation processing for each of the texts, and compute the co-occurrence frequency between each characteristic word and each hyperlink word.

8. The device according to claim 7, wherein the second processing unit comprises:

a first processing sub-unit, configured to collect a variety of texts; and

a second processing sub-unit, configured to carry out the word segmentation processing for each of the texts, and extract all unrepeated words obtained by the word segmentation processing as the characteristic words to form a characteristic word list; or, remove high-frequency words, stop words and low-frequency words from all the unrepeated words obtained by the word segmentation processing, and consider remaining words as the characteristic words to form a characteristic word list;

for each characteristic word y and each hyperlink word x , calculating the co-occurrence frequency $P(x/y)$ using a formula of: $P(x/y) = \text{a quantity of } xy \text{ co-occurrences} / \text{a quantity of } y \text{ occurrences}$; wherein, the quantity of xy co-occurrences represents the number of texts appearing the characteristic words y and the hyperlink words x in the same time; the quantity of y occurrences represents the number of texts appearing the characteristic words y ;

or,

for each characteristic word y and each hyperlink word x , calculating the co-occurrence frequency $P(x/y)$ using a formula of: $P(x/y) = H(x,y)/I(x,y) = H(x,y)/(H(x)+H(y)-H(x,y))$, wherein, H represents information entropy; I represents mutual information.

9. The device according to claim 8, wherein the adding module comprises:

a third processing unit, configured to carry out the word segmentation processing to the text X and obtain a segmentation result; and

a fourth processing unit, configured to extract the hyperlink words occurred in the hyperlink word list and the characteristic words occurred in the characteristic

word list from the segmentation result, compute a weight of each of the hyperlink words that are occurred in the hyperlink word list, and compute a weight of each of the characteristic words that are occurred in the characteristic word list, and determine the final weight of each of the hyperlink word according to each co-occurrence frequency and the weights of the hyperlink words; and

a fifth processing unit, configured to descendingly sort the hyperlink words occurred in the hyperlink word list according to the final weights of the hyperlink words, and obtain K numbers of hyperlink words that are arranged in first, and add hyperlinks for the K numbers of hyperlink words, wherein, K is a positive integer.

10. The device according to claim 9, wherein ~~the second processing sub-unit is further configured to respectively determine an IDF (inverse document frequency) value for each characteristic word, the IDF value is calculated by following processes: obtaining a quotient by a quantity of the variety of texts collected divided by a quantity of texts appearing the characteristic words, and calculating a logarithm of the quotient;~~

the fourth processing unit comprises:

a third processing sub-unit, configured to extract the hyperlink words occurred in the hyperlink word list and the characteristic words occurred in the characteristic word list from the segmentation result; for each hyperlink word H, calculate the weight W_H of the hyperlink word H: $W_H = TF_H * IDF_H$, wherein TF_H represents TF (term frequency) value of the hyperlink word H, that is, TF_H refers to the quantity of the hyperlink word H appearing in the text X, and IDF_H represents IDF value of the hyperlink word H; for each characteristic word F, calculate the weight W_F of the extracted characteristic word F: $W_F = TFF * IDFF$; wherein TFF represents TF value of the characteristic word F, and $IDFF$ represents IDF value of the characteristic word F; and

a fourth processing sub-unit, configured to, for each hyperlink word H, calculate

the final weight W_H' of the hyperlink word H: $W_H' = W_H * \sum_{i=1}^n P(H/F_i) * W_{F_i}$; wherein, n represents a quantity of the characteristic words.

11. A non-transitory computer readable storage medium, storing one or more programs for execution by one or more processors of a computer having a display, the one or more programs comprising instructions for:

creating a hyperlink word list in advance, the hyperlink word list comprising a plurality of hyperlink words;

collecting a variety of texts, and generating a characteristic word list by implementing word segmentation processing for each of the texts, the characteristic word list comprising a plurality of characteristic words;

respectively determining an IDF (inverse document frequency) value for each characteristic word after generating a characteristic word list by implementing word segmentation processing for each of the texts, wherein the IDF value is calculated by following processes: obtaining a quotient by a quantity of the variety of texts collected divided by a quantity of texts appearing the characteristic word, and calculating a logarithm of the quotient;

for each of the characteristic words, computing a co-occurrence frequency between each of the characteristic words and each of the hyperlink words;

considering each text to be added a hyperlink as a text X, and processing the text X by following steps:

carrying out the word segmentation processing to the text X, and obtaining a segmentation result;

extracting the hyperlink words occurred in the hyperlink word list and the characteristic words occurred in the characteristic word list from the segmentation result;

computing a weight of each of the hyperlink words that are occurred in the hyperlink word list, and computing a weight of each of the characteristic words that are occurred in the characteristic word list;

determining a final weight of each of the hyperlink words according to each co-occurrence frequency and the weights of the hyperlink words;

descendingly sorting the hyperlink words occurred in the hyperlink word list according to the final weights of the hyperlink words, and obtaining K numbers of hyperlink words that are arranged in first; and

adding hyperlinks for the K numbers of hyperlink words, wherein K is a positive integer.