

METHOD, DEVICE, AND COMPUTER STORAGE MEDIA FOR ADDING HYPERLINK TO TEXT

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the priority benefit of Chinese Patent Application No. 201210087642.0 filed on March 29, 2012, the content of which is incorporated by reference herein in its entirety for all purposes.

FIELD

[0002] The present disclosure relates to text processing technology field, and specifically to a method and a device for adding hyperlink to text.

BACKGROUND

[0003] A hyperlink is a reference to a link in the Internet that points to a specific target, which will automatically jump to the specific target by clicking the hyperlink.

[0004] Different texts are linked together by adding the hyperlinks in the texts. As shown in FIG. 1, FIG. 1 is a schematic diagram of an existing text adding the hyperlinks. In the FIG. 1, certain words such as “Hall of Supreme Harmony”, “Hall of Complete Harmony” and “Hall of Preserving Harmony” are added the hyperlinks. In practical applications, the words added the hyperlinks are usually called hyperlink words which are often represented by blue fonts with underlines.

[0005] In the prior art, the following methods are generally used to add the hyperlink to the text:

[0006] 1) to each text to be added the hyperlink, manually determining which words need to be added the hyperlinks;

[0007] 2) creating a hyperlink word list, and matching texts to be added the hyperlinks with the hyperlink word list, and adding the hyperlinks to the words appearing in the hyperlink word list according the matching results.

[0008] However, in practice, the above two methods exist some problems:

[0009] to method 1) due to the need for manual operations, this method is not convenient to achieve, especially adding the hyperlinks to a large number of texts;

[0010] to method 2) although this method is relatively easy to realize, but it just only adds the hyperlinks to the matched words, which may result in a lower correlation between the texts and the hyperlinks added to the texts.

SUMMARY

[0011] In view of this, a main purpose of present disclosure is providing a method for adding hyperlink to text and a device for adding hyperlink to text, which can improve the correlation between the texts and the hyperlinks added to the texts, and be easy to implement.

[0012] In order to achieve the above object, the technology solution of the present disclosure is realized as:

[0013] a method for adding hyperlink to text, including:

[0014] creating a hyperlink word list in advance, the hyperlink word list comprising a plurality of hyperlink words;

[0015] 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;

[0016] for each of the characteristic words, computing a co-occurrence frequency between each of the characteristic words and each of the hyperlink words;

[0017] considering each text to be added a hyperlink as a text X, and processing the text X by following steps:

[0018] carrying out the word segmentation processing to the text X, and obtaining a segmentation result;

[0019] extracting the hyperlink words occurred in the hyperlink word list and the characteristic words occurred in the characteristic word list from the segmentation result;

[0020] 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;

[0021] determining a final weight of each of the hyperlink words according to each co-occurrence frequency and the weights of the hyperlink words;

[0022] 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

[0023] adding hyperlinks for the K numbers of hyperlink words, wherein K is a positive integer.

[0024] A device for adding hyperlink to text, includes:

[0025] 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;

[0026] an adding module, configured to consider each text to be added a hyperlink as a text X, and process the text by following steps:

[0027] carrying out the word segmentation processing to the text X, and obtaining a segmentation result;

[0028] extracting the hyperlink words occurred in the hyperlink word list and the characteristic words occurred in the characteristic word list from the segmentation result;

[0029] 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;

[0030] determining a final weight of each of the hyperlink words according to each co-occurrence frequency and the weights of the hyperlink words;

[0031] 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

[0032] adding hyperlinks for the K numbers of hyperlink words, wherein, K is a positive integer.

[0033] According to the above technical solutions of the present disclosure, by using the embodiments of the present invention, the correlation between the words by

counting the co-occurrence relationship between the words of the collected text, and the final weights of each extracted hyperlink word are obtained according to the co-occurrence frequency of each extracted characteristic word and each extracted hyperlink word from the texts to be added the hyperlinks, and the hyperlink words with larger final weight values are added the hyperlinks, thereby increasing the correlation between the texts and the hyperlinks to be added. Moreover, by use of the solution of the present disclosure, it can automatically determine to add the hyperlinks to which words without manual operations, which are easy to implement.

BRIEF DESCRIPTION OF THE DRAWINGS

[0034] FIG. 1 is a schematic diagram of an existing text adding the hyperlinks.

[0035] FIG. 2 is a flowchart of a method for adding hyperlink to text provided in one embodiment of the present invention.

[0036] FIG.3 is a schematic diagram of a device for adding hyperlink to text provided in one embodiment of the present invention.

DETAILED DESCRIPTION OF ILLUSTRATED EMBODIMENTS

[0037] Embodiment of the present disclosure provides a solution for adding hyperlink to text, which can improve the correlation between texts and hyperlinks added to the texts, and be easy to implement.

[0038] Further description of present disclosure will be illustrated, which is combined with embodiments in drawings and detailed embodiments to make the purpose, the technical solutions and the advantages of the present invention more clearly.

[0039] Referring to FIG. 2, FIG.2 is a flowchart of a method for adding hyperlink to text provided in one embodiment of the present invention. As shown in FIG. 2, the method may include steps as below.

[0040] In step 21: creating a hyperlink word list in advance and collecting a variety of texts, and generating a characteristic word list by implementing word segmentation processing for each of the texts, and for each of the characteristic words, respectively

computing a co-occurrence frequency between each of the characteristic word lists and each of the hyperlink words.

[0041] In this step, the hyperlink word list is first generated, and specifically includes hyperlink words which can be determined based on actual need. The hyperlink word list can be generated by the way of manual editing, or also automatically generated by machine; this does not make a specific limit.

[0042] The texts can be collected from the Internet to generating the characteristic word list and counting co-occurrence relationship between words. How to collect the texts is an existing technology, in theory, the number of the collected texts is the more the better, and the specific number of the texts can be determined according to the actual needs.

[0043] The specific processes of generating the characteristic word list and computing co-occurrence relationship between words may include:

[0044] 1) carrying out the word segmentation processing for each collected text;

[0045] 2) extracting all unrepeated words obtained by the word segmentation processing as the characteristic words to form a characteristic word list; or, in order to reduce subsequent processing workload, 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 to form a characteristic word list; how to segment words and how to distinguish the high-frequency words, the stop words and the low-frequency words from all the unrepeated words are existing technology;

[0046] Moreover, after obtaining the characteristic word list, respectively determining IDF (inverse document frequency) value of each characteristic word, the IDF value can be calculated by the 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;

[0047] 3) respectively computing the co-occurrence frequency between each characteristic word and each hyperlink word, specifically including processes:

[0048] for each characteristic word y and each hyperlink word x , calculating the co-occurrence frequency $P(x/y)$ using a formula of:

[0049] $P(x/y)$ = a quantity of xy co-occurrences / a quantity y occurrences; (1)

[0050] 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 in all the selected texts; the quantity of y occurrences represents the number of texts appearing the characteristic words y ;

[0051] or,

[0052] for each characteristic word y and each hyperlink word x , calculating the co-occurrence frequency $P(x/y)$ using a formula of:

[0053] $P(x/y) = H(x,y) / I(x,y) = H(x,y) / (H(x) + H(y) - H(x,y))$; (2)

[0054] wherein, H represents information entropy; I represents mutual information; the information entropy H is a measure of the uncertainty associated with random variables of the hyperlink word x and the characteristic word y , whose unit is bit, and the mutual information I is a measure of the mutual dependence of two random variables of the hyperlink word x and the characteristic word y , whose unit is bit. The specific calculations of the information entropy H and the mutual information I are known in the art;

[0055] in practical applications, any one of the above two ways can be selected according to actual needs.

[0056] In step 22: considering each text to be added a hyperlink as a text X , and processing the text X to be added a hyperlink respectively according to the following processes as shown in steps 23-26.

[0057] In order to facilitate presentation, the text X is used here to represent any one text which needs to be added the hyperlink.

[0058] In step 23: carrying out the word segmentation processing to the text X and obtaining a segmentation result.

[0059] In step 24: extracting the hyperlink words occurred in the hyperlink word list and the characteristic words occurred in the characteristic word list from the segmentation results and computing a weight of each of the extracted hyperlink words and each of the extracted characteristic words.

[0060] Wherein, the segmentation result of word segmentation is matched with the hyperlink word list and the characteristic word list generated in the step 21, to extract

the hyperlink words occurred in the hyperlink word list and the characteristic words occurred in the characteristic word list.

[0061] Furthermore, in this step, for each hyperlink word H, the weight WH of the extracted hyperlink word H can be calculated by the following formula (3):

$$[0062] \text{WH} = \text{TFH} * \text{IDFH}; \quad (3)$$

[0063] wherein, TFH represents TF (term frequency) value of the hyperlink word H, that is, TFH refers to the quantity of the hyperlink word H appearing in the text X, and IDFH represents IDF value of the hyperlink word H.

[0064] In his step, for each characteristic word F, the weight WH of the characteristic word F can be calculated by the following formula (4):

$$[0065] \text{WF} = \text{TFF} * \text{IDFF}; \quad (4)$$

[0066] wherein, TFF represents TF value of the characteristic word F, and IDFF represents IDF value of the characteristic word F.

[0067] In the embodiment of the present invention, each IDF value has been figured out in the step 21.

[0068] In step 25: determining a final weight of each of the hyperlink words according to each co-occurrence frequency and the weights of the hyperlink words.

[0069] In this step, for each hyperlink word H, the final weight WH' of the extracted hyperlink word H can be calculated by the following formula (5):

$$[0070] W_H' = W_H * \sum_{i=1}^n P(H/F_i) * W_{Fi}; \quad (5)$$

[0071] wherein, n represents the quantity of the characteristic words.

[0072] The value of P(H/F_i) has been calculated in the step 21.

[0073] In step 26: 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 arranged in first, and adding hyperlinks for the K numbers of hyperlink words, wherein, K is a positive integer.

[0074] In summary, the description for the method provided in the embodiments of the present invention is completed.

[0075] Based on the above description for the method, referring to FIG. 3, a device for adding hyperlink to text is provided. FIG.3 is a schematic diagram of a device for

adding hyperlink to text provided in one embodiment of the present invention. The device may include: a preprocessing module and an adding module.

[0076] The preprocessing module, is capable of creating a hyperlink word list in advance and collecting a variety of texts, and generating a characteristic word list by implementing word segmentation processing for each of the texts, and for each of the characteristic words computing a co-occurrence frequency between each of the characteristic words and each of the hyperlink words.

[0077] The adding module, is capable of considering each text to be added a hyperlink as a text X and processing the text X respectively according to the following steps in detail:

[0078] carrying out the word segmentation processing to the text X and obtaining a segmentation result;

[0079] extracting the hyperlink words occurred in the hyperlink word list and the characteristic words occurred in the characteristic word list from the segmentation result and 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;

[0080] determining a final weight of each hyperlink word according to each co-occurrence frequency and the weights of the hyperlink words; and

[0081] 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 the hyperlinks for the K numbers of hyperlink words, wherein, K is a positive integer.

[0082] In the embodiment of the present invention, the preprocessing module may include: a first processing unit and a second processing unit.

[0083] The first processing unit, is capable of creating a hyperlink word list.

[0084] The second processing unit, is capable of collecting a variety of texts, and generating a characteristic word list by implementing word segmentation processing for each text, and respectively computing the co-occurrence frequency between each characteristic word and each hyperlink word.

[0085] The second processing unit can further include: a second processing sub-unit and a second processing sub-unit, which are not shown in FIG. 3 to simplify the drawing.

[0086] The first processing sub-unit, is capable of collecting a variety of texts.

[0087] The second processing sub-unit, is capable of carrying out the word segmentation processing for each collected text, and extracting all unrepeated words obtained by the word segmentation processing as the characteristic words to form a characteristic word list; 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 to form a characteristic word list;

[0088] moreover, for each characteristic word y and each hyperlink word x , calculating the co-occurrence frequency $P(x/y)$ by a formula (1):

[0089] $P(x/y) = \text{a quantity of } xy \text{ co-occurrences} / \text{a quantity of } y \text{ occurrences}$ (1)

[0090] 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 in all the selected texts; the quantity of y occurrences represents the number of texts appearing the characteristic words y ;

[0091] or, for each characteristic word y and each hyperlink word x , calculating the co-occurrence frequency $P(x/y)$ by a formula (2):

[0092] $P(x/y) = H(x,y) / I(x,y) = H(x,y) / (H(x) + H(y) - H(x,y))$ (2)

[0093] wherein, H represents information entropy; I represents mutual information; the information entropy H is a measure of the uncertainty associated with random variables of the hyperlink word x and the characteristic word y , whose unit is bit, and the mutual information I is a measure of the mutual dependence of two random variables of the hyperlink word x and the characteristic word y , whose unit is bit.

[0094] The adding module may include: a third processing unit, a fourth processing unit and a fifth processing unit.

[0095] The third processing unit, is capable of carrying out the word segmentation processing to the text X and obtaining a segmentation result.

[0096] The fourth processing unit, is capable of extracting the hyperlink words occurred in the hyperlink word list and the characteristic words occurred in the characteristic word list from the segmentation result and computing a weight of each of the extracted hyperlink word that are occurred in the hyperlink word list, and computing a weight of each of the extracted characteristic words that are occurred in the characteristic word list, and determining a final weight of each of the hyperlink words according to each co-occurrence frequency and the weights of the hyperlink words.

[0097] The fifth processing unit, is capable of 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.

[0098] Moreover, the second processing sub-unit is further capable of respectively determining an IDF value of each characteristic word, the IDF value can be calculated by the 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.

[0099] The fourth processing unit can further include: a third processing sub-unit and a fourth processing sub-unit, which are not shown in FIG. 3 to simplify the drawing.

[00100] The third processing sub-unit, is capable of extract the hyperlink words occurred in the hyperlink word list and the characteristic words occurred in the characteristic word list from the segmentation result.

[00101] Furthermore, for each hyperlink word H, the weight W_H of the hyperlink word H can be calculated by a following formula (3):

$$[00102] W_H = TF_H * IDF_H; \quad (3)$$

[00103] 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.

[00104] For each characteristic word F, the weight W_F of the characteristic word F can be calculated by a following formula (4):

[00105] $WF = TFF * IDFF;$ (4)

[00106] wherein, TFF represents TF value of the characteristic word F, and IDFF represents IDF value of the characteristic word F.

[00107] The fourth processing sub-unit, is capable of, for each hyperlink word H, calculating the final weight W_H' of the hyperlink word H by a following formula (5):

[00108] $W_H' = W_H * \sum_{i=1}^n P(H/F_i) * W_{Fi};$ (5)

[00109] wherein, n represents the quantity of the extracted characteristic words.

[00110] The specific workflow of the device provided in the embodiment of FIG. 3 can be found in the method provided in the embodiment of FIG. 2, which is not repeated here.

[00111] The present disclosure further provides a computer readable storage media, 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 executing the method for adding hyperlink to text by a machine. Specifically, a system or a device configured with the storage media can be provided, and the storage media stores software program codes which can execute and realize the function of any embodiment of the above embodiments, and computers (such as CPUs or MPUs) of the system or the device read and execute the program codes stored in the storage media.

[00112] In this case, the program codes stored in the storage media themselves can execute and realize the function of any embodiment of the above embodiments, therefore, the program codes and the storage media including the program codes constitute one part of the present disclosure.

[00113] Embodiments of the storage media for supplying the program codes include: a floppy disk, a hard disk, a CD (compact disc) (such as, CD-ROM, CD-R, CD-RW, DVD-ROM, DVD-RAM, DVD-RW, DVD+RW), a magnetic tape, a non-volatile memory card and ROM. Alternatively, the program codes can be downloaded from server computers by a communication network.

[00114] Furthermore, it should be noted that the operating system of the computer can execute and complete some or all the actual operations by executing the program

codes read by the computer or by executing the instructions based on the program codes, therefore realizing the functions of any embodiment of the above embodiments.

[00115] Moreover, it should be understood that the program codes read out from the storage media are written in memories of expansion boards inserted in the computer, or are written in memories provided in expansion units connected to the computer, and then the instructions based on the program codes can enable the CPUs mounted on the expansion boards or the expansion units to execute and perform some or all the actual operations, therefore realizing the functions of any embodiment of the above embodiments.

[00116] While the present disclosure has been described with reference to particular embodiments, it will be understood that the embodiments are illustrative and that the invention scope is not so limited. Alternative embodiments of the present invention will become apparent to those having ordinary skill in the art to which the present invention pertains. Such alternate embodiments are considered to be encompassed within the spirit and scope of the present disclosure. Accordingly, the scope of the present invention is described by the appended claims and is supported by the foregoing description.

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, further comprising:

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 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.

METHOD, DEVICE, AND COMPUTER STORAGE MEDIA FOR ADDING HYPERLINK TO TEXT

ABSTRACT

Methods and devices for adding hyperlink to text are disclosed: generating hyperlink word list and characteristic word list in advance, determining co-occurrence frequency with each hyperlink word; to each text X which to be added the hyperlink, word segmentation processing them respectively, extracting the hyperlink word occurred in the hyperlink word list and the characteristic word occurred in the characteristic word list from results of word segmentation, determining weights of each extracted hyperlink word and extracted characteristic word, getting final weights of each extracted hypertext link word according to the co-occurrence frequency of each extracted characteristic word and each extracted hyperlink word and the weights; descendingly sorting each extracted hyperlink word according to the final weights, adding hyperlink to first k hyperlink words, and K is positive integer. Applying the solution, it can improve the relativity of the added hyperlink and the text, and it is easy to implement.

FIG. 2