

Claims

1. A computer-implemented method for removing a duplicated web page, the method comprising:

acquiring (S101) a plurality of web pages of a predetermined type;
extracting (S102) a feature code of a current web page and a number of text characters contained in the current web page, extracting the feature code of the current web page comprising:

acquiring paragraphs contained in the text of the current web page;
for each paragraph, selecting a second pre-set number of text characters on the left and right sides of a central location of the paragraph; and
generating the feature code by a calculation on a character string that combines all the selected text characters of the paragraphs,

wherein when a number of characters in a paragraph is smaller than a pre-set first number, the feature code is supplemented with a corresponding number of pre-defined characters, and
wherein the second pre-set number is half of the first pre-set number;

looking up a data table to determine (S102) whether the feature code is contained in the data table; and
after determining (S102) that the feature code is contained in the data table:

reading a number of text characters of the web page referred to in the data table corresponding to the feature code, and

discarding the current web page when a difference between the read number of text characters and the extracted number of text characters is within a range.

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

extracting a feature code of a second current web page and a number of text characters contained in second current web page;

looking up a data table to determine whether the feature code of the second current web page is contained in the data table; and

after determining that the feature code of the second current web page is not contained in the data table after looking up the data table, writing correspondence between the extracted feature code and the extracted number of text characters of the current web page into the data table.

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

extracting a feature code of a second current web page and a number of text characters contained in second current web page;

looking up a data table to determine whether the feature code of the second current web page is contained in the data table; and

after determining that the feature code of the second current web page is contained in the data table:

reading a number of text characters of the web page referred to in the data table corresponding to the feature code of the second current web page, and

5 writing the correspondence between the extracted feature code and the extracted number of text characters of the second current web page into the data table when the difference between the read number of text characters and the extracted number of text characters of the second current web page is not within the range.

10 4. The method according to any of claims 1-3, wherein the second number is 3 to 8.

5. The method according to claim 4, wherein the second number is 5.

15 6. The method according to any of the preceding claims, wherein acquiring the plurality of web pages of the predetermined type comprises:

acquiring web pages relevant to a keyword, and
selecting, from the acquired web pages, web pages containing texts to determine the plurality of webpages.

20 7. An apparatus for removing a duplicated web page, the apparatus comprising:

an acquisition module (100) configured to acquire a plurality of web pages of a predetermined type; and
a first processing module (200) configured to:

25 extract (S101) a feature code of a current web page and a number of text characters contained in the current web page for each web page,
look up a data table to determine (S102) whether the feature code is contained in the data table,
in response to the feature code being contained in the data table, read a number of text characters of the web page referred to in the data table corresponding to the feature code, and
30 discard the current web page when a difference between the read number of text characters and the extracted number of text characters is within a range,
wherein to extract the feature code of the current web page, the first processing module is configured to:

35 acquire paragraphs contained in the text of the current web page,
for each paragraph, select a second pre-set number of text characters on the left and right sides of a central location of the paragraph, and
generate the feature code by a calculation on a character string that combines all the selected text characters of the paragraphs,

40 wherein when a number of characters in a paragraph is smaller than a pre-set first number, the feature code is supplemented with a corresponding number of pre-defined characters, and

wherein the second pre-set number is half of the first pre-set number.

45 8. The apparatus according to claim 7, further comprising:

a second processing module (300) configured to write the correspondence between the extracted feature code and the extracted number of text characters of the current web page into the data table if the feature code is not contained in the data table after looking up the data table to determine whether the feature code is contained in the data table.

50

9. The apparatus according to claim 7, further comprising:

55 a third processing module (400) configured to write the correspondence between the extracted feature code and the extracted number of text characters of the current web page into the data table when the difference between the read number of text characters and the extracted number of text characters is not within the range.

10. The apparatus according to claim of any of claims 7 to 9, wherein the second number is 3 to 8.

11. The apparatus according to claim 10, wherein the second number is 5.

12. The apparatus according to any of claims 7 to 11, wherein to acquire the plurality of web pages of the predetermined type, the first processing module (200) is configured to:

5

acquire web pages relevant to a keyword, and
select, from the acquired web pages, web pages containing texts to determine the plurality of webpages.

13. A computer-readable storage medium storing instructions which, when executed by a processor of an apparatus,
cause the apparatus to perform the method of any one of claims 1 to 6.

10