

[Document Name] DESCRIPTION

[Title of Invention]

DICTIONARY CREATION DEVICE FOR MONITORING TEXT
INFORMATION, DICTIONARY CREATION METHOD FOR

5 MONITORING TEXT INFORMATION, AND DICTIONARY CREATION
PROGRAM FOR MONITORING TEXT INFORMATION

[Technical Field]

[0001]

The present invention relates to dictionary generation devices for
10 monitoring text information, dictionary generation methods for monitoring
text information, and dictionary generation programs for monitoring text
information. In particular, the present invention relates to a dictionary
generation device for monitoring text information, a dictionary generation
method for monitoring text information, and a dictionary generation
15 program for monitoring text information, by which a dictionary for
monitoring text information with high precision even for unknown text is
generated.

[Background Art]

[0002]

20 For monitoring of rumors on the Internet, and the like, text
information monitoring technologies by which information contents
targeted for monitor are detected appearing from a large amount of text
have become important. Text information monitoring systems assumed in
the present invention monitor text information on the basis of dictionaries.
25 In other words, as a technique of the text information monitor technologies,
there is used a dictionary-based technique in which conditions for
detection are maintained in a dictionary for monitoring text information
and it is detected whether or not an expression in an input document
matches the conditions in the dictionary for monitoring text information.

[0003]

In the dictionary-based technique, text information can be monitored with high precision by using a high-precision dictionary. Thus, it is important that the high-precision dictionary is used.

5 [0004]

Generation of a dictionary with introspection in a text information monitoring system based on a dictionary consumes time, is prone to result in omission, and is therefore difficult. Thus, there is desired a technique of giving a positive example group, in which documents including an
10 information content targeted for monitor are gathered, and a negative example group, in which documents that do not include the information content targeted for monitor are gathered, to automatically extract an expression to be registered as a detection condition from the groups. Conventional techniques of such a method include a feature word
15 extraction technique. The feature word extraction technique is a technique that compares a positive example group and a negative example group to extract, as a feature word, a word that characteristically appears in the positive example group.

[0005]

20 An example of such a technique is PTL 1. In PTL 1, when a dictionary used in text mining is constructed, document data targeted for analysis is divided into groups, and expressions that characteristically appear in each group are used as dictionary candidates.

[Citation List]

25 [Patent Literature]

[0006]

[PTL 1]: Japanese Patent Laid-Open No. 2009-015394

[Summary of Invention]

[Technical Problem]

[0007]

However, a feature word extraction technique with a short unit at a word or modification level in the conventional art is unable to sufficiently satisfy the performance requirements of a text information monitoring system. This is because detection precision is decreased only with a short unit at a word or modification level. For example, even if one word "virus" is registered in a dictionary for monitoring text information in order to detect a description about a computer virus, a document including, e.g., "cold virus" is detected mistakenly. In this case, it is necessary to register a phrase including one or more words, such as "computer virus" or "virus email", in the dictionary for monitoring text information.

[0008]

As described above, an optimal phrase length depends on what to intend to detect, and therefore, it is impossible to decide the length as a unique value in advance. Thus, it is necessary to extract phrases with any lengths as candidates and to calculate a feature degree for each phrase in order to handle a phrase with a variable length. Further, it is impossible to appropriately handle a case in which plural phrases overlapping each other are output at an equal feature degree.

[0009]

For example, phrases as represented in Fig. 4 are extracted and "Trojan horse", "Trojan", and "horse" are extracted at an equal feature degree (= 3) by carrying out feature word extraction for phrases with various lengths when positive and negative example groups as represented in Fig. 3 are given. However, although neither "Trojan" nor "horse" appears in the negative example group, since expressions, such as "Trojan heritage" and "carousel horse", irrelevant to the virus, are conceivable, the registration of "Trojan" and "horse" in a dictionary for monitoring text information results in lower detection precision. Theoretically, the

appearance of an expression such as "Trojan heritage" or "carousel horse" in a negative example group can result in the lower feature degree of an expression such as "Trojan" or "horse" and also in lower detection precision. However, in reality, a negative example group with a sufficient amount is rarely obtained, and such a problem as described above occurs frequently.

[0010]

In PTL 1, a technique in which a word collocating with a feature word is also regarded as a dictionary registration candidate is disclosed; however, an index such as the product of TF (Term Frequency) and IDF (Inverse Document Frequency) is used in determination whether or not to carry out dictionary registration, and it is considered that there is such a problem as described above for plural phrases overlapping each other.

[0011]

As described above, conventional techniques in which a dictionary for monitoring text information is constructed with a feature degree calculated from a positive example group and a negative example group have a problem that lower detection precision is caused.

[0012] The present invention is to solve the problems described above and to provide a dictionary generation device for monitoring text information, a dictionary generation method for monitoring text information, and a dictionary generation program for monitoring text information such that it is possible to achieve high-precision detection compared to the conventional art.

[Solution to Problem]

[0013]

The present invention which is to solve the problems described above is a dictionary generation device for monitoring text information, which is used in a text information monitoring system and generates a

dictionary in which a detection condition is registered, including: a feature degree calculation unit that calculates, for a phrase as a candidate for the detection condition, a feature degree representing a degree by which the phrase matches an information content targeted for monitor; and a phrase usefulness determination unit that determines whether or not the phrase is appropriate for the detection condition based on the feature degree and a usefulness degree representing littleness of ambiguity of meaning defined by the phrase.

[0014]

The present invention which is to solve the problems described above is a method for generating a dictionary used in a text information monitoring system, wherein a dictionary generation device for monitoring text information calculates, for a phrase as a candidate for a detection condition, a feature degree representing a degree by which the phrase matches an information content targeted for monitor; determines whether or not the phrase is appropriate for the detection condition based on the feature degree and a usefulness degree representing littleness of ambiguity of meaning defined by the phrase; and outputs a phrase judged to be appropriate and registers the phrase as the detection condition.

[0015]

The present invention which is to solve the problems described above is a dictionary generation program for monitoring text information, which allows a dictionary generation device for monitoring text information to execute a processing of calculating, for a phrase as a candidate for a detection condition, a feature degree representing a degree by which the phrase matches an information content targeted for monitor; a processing of determining whether or not the phrase is appropriate for the detection condition based on the feature degree and a usefulness degree representing littleness of ambiguity of meaning defined by the phrase; and

a processing of outputting a phrase judged to be appropriate and of registering the phrase as the detection condition.

[Advantageous Effects of Invention]

[0016]

5 In general, the longer length of a phrase results in the less ambiguity of meaning and in a higher matching rate for a detection condition. In the present invention, a usefulness degree is calculated based on the length of a phrase, and a phrase to be registered in a dictionary is extracted based on the usefulness degree and a feature degree.
10 In other words, priority is given to a phrase having a longer length.

[0017]

As a result, there can be generated a dictionary for monitoring text information such that it is possible to achieve high-precision detection compared to the conventional art.

15 [Brief Description of Drawings]

[0018]

[Fig. 1] Fig. 1 is a functional block diagram of a dictionary generation device.

[Fig. 2] Fig. 2 is an operation flow of a dictionary generation device.

20 [Fig. 3] Fig. 3 is an example of a positive example group and a negative example group (common to the conventional art).

[Fig. 4] Fig. 4 is an example of the frequency and feature degree of each phrase (common to the conventional art).

[Fig. 5] Fig. 5 is an example of the usefulness degree and score of each
25 phrase (Application Example 1).

[Fig. 6] Fig. 6 is an example of the usefulness degree and score of each phrase (Application Example 2).

[Fig. 7] Fig. 7 is an example of the usefulness degree and score of each phrase (Application Example 3).

[Fig. 8] Fig. 8 is an example of the usefulness degree and score of each phrase (Application Example 4).

[Fig. 9] Fig. 9 is an example of the usefulness degree and score of each phrase (Application Example 5).

5 [Description of Embodiments]

[0019] -Configuration/Operation-

The configurations and operations of exemplary embodiments of the present invention will be explained in detail below with reference to the drawings.

10 [0020]

Fig. 1 is a functional block diagram of a dictionary generation device according to the present exemplary embodiment. The dictionary generation device according to the present exemplary embodiment includes a phrase extraction unit 1, a phrase usefulness determination unit 2, a feature degree calculation unit 3, and an output unit 4. The phrase usefulness determination unit 2 includes a usefulness degree calculation unit 21 and a detection condition determination unit 22.

[0021]

The function of each configuration will be explained.

20 [0022]

It is assumed that a positive example group, in which documents including an information content targeted for monitor are collected, and a negative example group, in which documents that do not include the information content targeted for monitor are collected, are given (see Fig. 3).

25 [0023]

The phrase extraction unit 1 performs language analysis for text in the given positive example group extracts phrases having various lengths as candidates for a detection condition. The phrases are extracted by

carrying out morphological analysis to extract the phrases as specific part-of-speech tag strings, by carrying out syntactic analysis to regard subtrees of obtained syntactic trees as the phrases, or by using a combination of the analyses.

5 [0024]

The phrase usefulness determination unit 2 calculates the usefulness degree of each phrase extracted in the phrase extraction unit 1, and further determines whether the phrase is appropriate for the detection condition by combining the usefulness degree and a feature degree
10 calculated by the feature degree calculation unit 3, and.

[0025]

The usefulness degree calculation unit 21 calculates a usefulness degree by using the length of the phrase, the frequency at which the phrase appears within the positive example group, and an index pertaining to an
15 inclusion relationship between phrases for each phrase extracted by the phrase extraction unit 1. As used herein, the usefulness degree of a phrase refers to a value representing the littleness of the ambiguity of meaning defined by the phrase and to a value representing detection precision in a case in which the phrase is regarded as a detection condition.
20 As the usefulness degree, the length of the phrase or the logarithmic value thereof may be used, or the product of the length of the phrase or the logarithmic value thereof and the number of the appearances of the phrase in the positive example group or the logarithmic value thereof may be used. Alternatively, as the usefulness degree, such a C-value as proposed in NPL
25 1 may be further used based on the index pertaining to the inclusion relationship between phrases.

NPL 1: Frantzi, K. and Ananiadou, S. (1996). "Extracting Nested Collocations." In Proceedings of the 16th International Conference on Computational Linguistics (COLING 96), pp. 41-46.

[0026]

Application examples of a usefulness degree calculation will be mentioned later (Application Examples 1 to 4).

[0027]

5 For each phrase, the detection condition determination unit 22 determines whether or not the phrase is appropriate for the detection condition by using the usefulness degree calculated by the usefulness degree calculation unit 21 and the feature degree calculated by the feature degree calculation unit 3. For example, the detection condition
10 determination unit 22 evaluates the appropriateness as the detection condition with the product of the usefulness degree and the feature degree, and determines that the phrase is appropriate for the detection condition in a case in which the value of the product is greater than a threshold value. The detection condition determination unit 22 can also exclude phrases, of
15 which the usefulness degrees are less than the threshold value, to reduce phrases of which the feature degrees are calculated and to reduce a calculation amount (Application Example 5).

[0028]

 The feature degree calculation unit 3 compares the statistics of the
20 positive example group and the negative example group, and calculates the degree by which a phrase of interest appears in the positive example group as the feature degree. The feature degree is calculated by using an existing measure, such as a chi-square value, a mutual information content, or ESC (Extended Stochastic Complexity), used in text mining. The
25 calculation of the feature degree in this case may be carried out for all the phrases extracted in the phrase extraction unit 1 or only for phrases necessary for determination in the phrase usefulness determination unit 2.

[0029]

The output unit 4 outputs, as a phrase to be registered in a

dictionary, a phrase judged to be appropriate for the detection condition by the phrase usefulness determination unit 2. The output unit 4 not only outputs only a phrase to be registered in the dictionary but also outputs the phrase together with a usefulness degree, a feature degree, a score
5 representing appropriateness as the detection condition, and the like, whereby phrases to be registered in the dictionary using manpower with reference to the score and the like can be sorted to lighten an operation for constructing a dictionary for monitoring text information.

[0030]

10 Fig. 2 is an operation flow of a dictionary generation device. A dictionary generation program allows the dictionary generation device to execute each processing of the operation flow. When the program is executed, the phrase extraction unit 1, the phrase usefulness determination unit 2, the feature degree calculation unit 3, and the output unit 4 are
15 operated.

[0031]

First, the phrase extraction unit 1 subjects text in a given positive example group to language analysis to extract phrases having various lengths as candidates for a detection condition (step S1).

20 [0032]

Then, the usefulness degree calculation unit 21 calculates the usefulness degree of each phrase extracted by the phrase extraction unit 1 (step S2).

[0033]

25 On the other hand, the feature degree calculation unit 3 calculates the feature degree of a phrase of interest (step S3).

[0034]

Then, for each phrase, the detection condition determination unit 22 determines whether or not the phrase is appropriate for the detection

condition by using the usefulness degree calculated by the usefulness degree calculation unit 21 and the feature degree calculated by the feature degree calculation unit 3 (step S4). For example, the detection condition determination unit 22 calculates a score based on the usefulness degree and the feature degree, and carries out the determination based on the score.

[0035]

Finally, the output unit 4 outputs a phrase to be registered in a dictionary (step S5), and the processing is finished.

[0036]

Either of step S2 and step 3 may be carried out earlier, or the steps may be carried out simultaneously.

[0037]

In step S3 and step S4, the feature degree of only a phrase of which the usefulness degree is not less than the threshold value may also be calculated to determine whether or not the phrase is appropriate for the detection condition.

[0038]

-Specific Example of Conventional Art-

The dictionary creation device according to the conventional art includes a phrase extraction unit 1, a feature degree calculation unit 3, and an output unit 4 (illustration is omitted). In other words, the dictionary generation device according to the conventional art is common to the present exemplary embodiment except the presence or absence of the phrase usefulness determination unit 2.

[0039]

The text information monitoring system according to the present invention matches a character string with the dictionary for monitoring text information to thereby monitor text information and to register the character string as a detection condition in the dictionary for monitoring

text information. However, the text information monitoring system according to the present invention is not limited to the above-described system, and the present invention is also effective in a system that monitors text information with a part-of-speech tag or a syntactic structure as a condition.

[0040]

The dictionary generation device generates a dictionary used in the dictionary for monitoring text information.

[0041]

Fig. 3 is an example of the positive example group and the negative example group. It is assumed that such positive and negative example groups are given.

[0042]

First, the phrase extraction unit 1 extracts a candidate for a detection condition from the positive example group. For example, when all phrases having three or less chunks are extracted from the positive example group of Fig. 3, the phrases such as "Trojan horse", "Trojan", "horse", "Trojan horse infection", "horse infection", "infection", and "email" are extracted as candidates for detection conditions.

[0043]

Then, the feature degree calculation unit 3 calculates the feature degree of each candidate for a detection condition. Fig. 4 is an example of the frequency and feature degree of each phrase. For example, assuming that a feature degree is calculated by

$$\text{feature degree} = (\text{frequency in positive example group}) - (\text{frequency in negative example group})$$

it is calculated that the feature degree of "Trojan horse" is 3, the feature degree of "Trojan" is 3, the feature degree of "horse" is 3, the feature degree of "Trojan horse infection" is 2, the feature degree of "horse

"infection" is 2, the feature degree of "infection" is 1, and the feature degree of "email" is 1.

[0044]

The output unit 4 outputs, for example, the phrases "Trojan horse",
5 "Trojan", and "horse", having the high feature degrees, and registers the phrases in the dictionary.

[0045]

-Specific Application Example 1-

The operations of the phrase extraction unit 1 and the feature
10 degree calculation unit 3 are similar to those of the conventional art. In other words, candidates for detection conditions are extracted from a positive example group, and the feature degree of each candidate for a detection condition is calculated.

[0046]

15 Further, the usefulness degree calculation unit 21 calculates the usefulness degree of each candidate for a detection condition. Fig. 5 is an example of the usefulness degree and score of each phrase (mentioned later). For example, the usefulness degree is calculated based on the product of the length of a phrase and a frequency in the positive example
20 group. In other words, when the usefulness degree is calculated by

usefulness degree = (length of the phrase) × (frequency in positive example group)

it is calculated that the usefulness degree of "Trojan horse" is 6, the usefulness degree of "Trojan" is 3, the usefulness degree of "horse" is 3,
25 the usefulness degree of "Trojan horse infection" is 6, the usefulness degree of "horse infection" is 4, the usefulness degree of "infection" is 2, and the usefulness degree of "email" is 2. In this case, the length of each phrase is calculated based on the number of chunks. In addition to the number of chunks, however, the length may also be calculated based on the

number of morphemes, the number of characters, a byte length, and the like.

[0047]

Then, the detection condition determination unit 22 evaluates each candidate for a detection condition (see Fig. 5). For example, the detection condition determination unit 22 calculates a score representing appropriateness for the detection condition based on the product of a usefulness degree and a feature degree. In other words, when the score is calculated by

$$\text{score} = \text{feature degree} \times \text{usefulness degree}$$

the detection condition determination unit 22 calculates that the score of "Trojan horse" is 18, the score of "Trojan" is 9, the score of "horse" is 9, the score of "Trojan horse infection" is 12, the score of "horse infection" is 8, the score of "infection" is 2, and the score of "email" is 2. For example, when phrases having a score of 10 or more are adopted as the detection conditions, the detection condition determination unit 22 determines that two of "Trojan horse" and "Trojan horse infection" are appropriate for the detection conditions.

[0048]

The output unit 4 outputs the phrases "Trojan horse" and "Trojan horse infection" based on the determination results from the detection condition determination unit 22, and registers the phrases in the dictionary.

[0049]

-Effect-

The effect of the present exemplary embodiment will be explained in comparison with the conventional art.

[0050]

In the conventional art in which a detection condition is determined based only on a feature degree, "Trojan horse", "Trojan", and "horse" have

feature degree = 3, which is maximum, and are detection conditions. However, expressions, such as "Trojan heritage" for "Trojan" and "carousel horse" for "horse", which are not intrinsically intended to be detected are detected, so that detection precision are decreased.

5 [0051]

In contrast, in the present exemplary embodiment, the phrase usefulness determination unit 2 uses the length of a phrase as a candidate to calculate a usefulness degree representing goodness for a detection condition in a case in which the phrase is the detection condition. The phrase usefulness determination unit 2 determines a phrase to be registered in the dictionary by using the obtained usefulness degree and a feature degree that is separately calculated.

[0052]

In general, the longer length of a phrase results in the less ambiguity of meaning and in a higher matching rate for a detection condition. Thus, higher-precision detection than that in the case of using only a feature degree is enabled by selecting a phrase having a long length in a case in which phrases that overlap each other have an equal feature degree.

20 [0053]

In addition to the length of a phrase, the frequency at which the phrase appears within a document group is further used to calculate a usefulness degree. The longer length of the phrase results in a higher matching rate but is considered to result in a lower recall rate since the appearance probability of the phrase is decreased. Thus, the consideration of the frequency as well as the length of the phrase allows a usefulness degree, at which a matching rate and a recall rate are balanced, to be calculated, and enables higher-precision detection.

[0054]

In the present exemplary embodiment, "Trojan horse" and "Trojan horse infection" are detection conditions, and neither "Trojan" nor "horse" is registered in the dictionary. As a result, higher-precision detection than that in the conventional art can be achieved.

5 [0055]

-Specific Application Example 2-

In Application Example 1 as described above, the usefulness degree calculation unit 21 calculates usefulness degrees based on the products of the lengths of phrases and frequencies in a positive example group;

10 however, when the differences between the usefulness degrees are intended to be more significant, a correction value may be subtracted from the lengths of the phrases.

[0056]

Fig. 6 is another example of the usefulness degree and score of each phrase. For example, the usefulness degree calculation unit 21 calculates a usefulness degree based on the product of a value, obtained by subtracting a correction value from the length of a phrase, and a frequency in a positive example group. The correction value may be determined experientially. In this example, the correction value is assumed to be
20 "-0.5". In other words, in the case of calculation by

usefulness degree = (length of phrase - 0.5) × (frequency in positive example group)

it is calculated that the usefulness degree of "Trojan horse" is 4.5, the usefulness degree of "Trojan" is 1.5, the usefulness degree of "horse" is 1.5,
25 the usefulness degree of "Trojan horse infection" is 5, the usefulness degree of "horse infection" is 3, the usefulness degree of "infection" is 1, and the usefulness degree of "email" is 1.

[0057]

As described above, the length of a phrase is corrected to be more

emphasized.

[0058]

Then, the detection condition determination unit 22 calculates, from score = feature degree $\times$ usefulness degree, that the score of "Trojan horse" is 13.5, the score of "Trojan" is 4.5, the score of "horse" is 4.5, the score of "Trojan horse infection" is 10, the score of "horse infection" is 6, the score of "infection" is 1, and the score of "email" is 1. For example, when phrases having a score of 10 or more are adopted for detection conditions, the detection condition determination unit 22 determines that two of "Trojan horse" and "Trojan horse infection" are appropriate for the detection conditions.

[0059]

In comparison with Application Example 1, the rate of the score of "Trojan" or "horse" is decreased with respect to the score of "Trojan horse". In other words, "Trojan horse" is more reliably registered in a dictionary, and "Trojan" and "horse" are more reliably excluded from dictionary registration. As a result, precision is improved.

[0060]

-Specific Application Example 3-

In Application Example 1 and Application Example 2 as described above, the detection condition determination unit 22 is set to adopt a phrase having a score of 10 or more as a detection condition, and therefore, "horse infection" is not registered in a dictionary but can be registered depending on settings. "Horse infection" is included in "Trojan horse infection", and is used as the expression of "Trojan horse infection", so-called a set phrase, in most cases. Thus, there is no point in registering both "horse infection" and "Trojan horse infection" in the dictionary.

[0061] Thus, the usefulness degree calculation unit 21 calculates a

usefulness degree based on an index representing an inclusion relationship between phrases as well as the length of a phrase and a frequency in a positive example group. For example, a C-value may be assumed to be the usefulness degree. The C-value is a value calculated from the following expression. Fig. 7 is another example of the usefulness degree (C-value) and score of each phrase.

Definition of C-value

C-value = (phrase length) $\times$ (frequency in positive example group - T/C) (in case of C>0)

10 C-value = (phrase length) $\times$ (frequency in positive example group) (in case of C=0)

T: Total of the frequency of appearance of a phrase that includes a phrase of interest and is longer than the phrase of interest

C: Cardinality of phrases that include a phrase of interest and are longer than the phrase of interest (i.e., the number of such phrases)
[0062]

T and C will be specifically explained below (see Fig. 7).
[0063]

Phrase of interest: "Trojan horse"

20 Phrase that includes the phrase of interest and is longer than the phrase of interest: "Trojan horse infection"

T=2: Frequency of appearance of "Trojan horse infection": 2

C=1: Phrase that includes the phrase of interest and is longer than the phrase of interest: 1

25 [0064]

Phrase of interest: "Trojan"

Phrases that include the phrase of interest and are longer than the phrase of interest: "Trojan horse" and "Trojan horse infection"

T=3+2=5: Frequency of appearance of "Trojan horse": 3, and

frequency of appearance of "Trojan horse infection": 2

C=2: Phrases that include the phrase of interest and are longer than the phrase of interest: 2

[0065]

5 Phrase of interest: "horse"

Phrases that include the phrase of interest and are longer than the phrase of interest: "Trojan horse", "Trojan horse infection", and "horse infection"

T=3+2+2=7: Frequency of appearance of "Trojan horse": 3, frequency of appearance of "Trojan horse infection": 2, and frequency of appearance of

10 "horse infection": 2

C=3: phrases that include the phrase of interest and are longer than the phrase of interest: 3

[0066]

Phrase of interest: "Trojan horse infection"

15 Phrase that includes the phrase of interest and is longer than the phrase of interest: none

T=0

C=0

[0067]

20 Phrase of interest: "horse infection"

Phrase that includes the phrase of interest and is longer than the phrase of interest: "Trojan horse infection"

T=2: Frequency of appearance of "Trojan horse infection": 2

C=1: Phrase that includes the phrase of interest and is longer than the phrase of interest: 1

25

[0068]

Phrase of interest: "Infection"

Phrases that include the phrase of interest and are longer than the phrase of interest: "Trojan horse infection" and "horse infection"

$T=2+2=4$: Frequency of appearance of "Trojan horse infection": 2, and frequency of appearance of "horse infection": 2

$C=2$: phrases that include the phrase of interest and are longer than the phrase of interest: 2

5 [0069]

Phrase of interest: "email"

Phrase that includes the phrase of interest and is longer than the phrase of interest: none

$T=0$

10 $C=0$

[0070]

Due to the correction with T and C , it is calculated that the usefulness degree of "Trojan horse" is 2, the usefulness degree of "Trojan" is 0.5, the usefulness degree of "horse" is 0.67, the usefulness degree of "Trojan horse infection" is 6, the usefulness degree of "horse infection" is 0, the usefulness degree of "infection" is 0, and the usefulness degree of "email" is 2.

[0071]

The usefulness degree of "Trojan horse infection" is 6 whereas the usefulness degree of "horse infection" is 0. The results show that since "horse infection" is a set phrase that is necessarily used as the expression of "Trojan horse infection" in a positive example document group, the term property of "horse infection" is low, and there is no point in adding "horse infection" as a condition if "Trojan horse infection" exists as a detection condition.

[0072]

On the other hand, the usefulness degree of "Trojan horse" is 2. Since "Trojan horse" has usage examples other than "Trojan horse infection", the term property and C -value of "Trojan horse" are higher than

those of "horse infection".

[0073]

A term property is an index representing the easiness of use as a group of phrases. A high term property means easy use as a group of phrases.

[0074]

Use of a C-value as a usefulness degree as described above results in the lower value of a phrase included in another longer phrase, eliminates addition of a redundant detection condition, and enables improvement of dictionary precision.

[0075]

Then, the detection condition determination unit 22 calculates, from score = feature degree $\times$ usefulness degree, that the score of "Trojan horse" is 6, the score of "Trojan" is 1.5, the score of "horse" is 2, the score of "Trojan horse infection" is 12, the score of "horse infection" is 0, the score of "infection" is 0, and the score of "email" is 2. For example, when phrases having a score of 5 or more are adopted as detection conditions, the detection condition determination unit 22 determines that two of "Trojan horse" and "Trojan horse infection" are appropriate for the detection conditions.

[0076]

-Specific Application Example 4-

In Application Example 3, a correction value as explained in Application Example 2 may be used. In this example, the correction value is assumed to be "-1". Fig. 8 is another example of the usefulness degree (C-value) and score of each phrase.

Definition of C-value

C-value = (phrase length - 1) $\times$ (frequency in positive example group - T/C) (in case of C>0)

C-value = (phrase length - 1) × (frequency in positive example group) (in case of C=0)

T: Total of the frequency of appearance of a phrase that includes a phrase of interest and is longer than the phrase of interest

5 C: Cardinality of phrases that include a phrase of interest and are longer than the phrase of interest (i.e., the number of such phrases)

The value "-1" in the terms of the phrase lengths is similar to the correction value "-0.5" described in Application Example 2. In other words, the value "-1" is a correction value for more emphasizing the length
10 of a phrase.

[0077]

As a result, the differences between usefulness degrees become more significant.

[0078]

15 -Application Example 5-

Only for phrases of which the usefulness degrees are not less than a threshold value, the feature degree calculation unit 3 calculates the feature degrees of the phrases, and the detection condition determination unit 22 determines whether or not the phrases are appropriate for detection
20 conditions.

[0079]

Specific explanation will be given in comparison with Application Example 2. Fig. 8 is another example of the usefulness degree and score of each phrase.

25 [0080]

Similarly to Application Example 2, the usefulness degree calculation unit 21 calculate that the usefulness degree of "Trojan horse" is 4.5, the usefulness degree of "Trojan" is 1.5, the usefulness degree of "horse" is 1.5, the usefulness degree of "Trojan horse infection" is 5, the

usefulness degree of "horse infection" is 3, the usefulness degree of "infection" is 1, and the usefulness degree of "email" is 1.

[0081]

The feature degree calculation unit 3 calculates the feature degrees of, for example, only the phrases having a usefulness degree of 3 or more: "Trojan horse", "Trojan horse infection", and "horse infection". Then, the detection condition determination unit 22 calculates, from score = feature degree $\times$ usefulness degree, that the score of "Trojan horse" is 13.5, the score of "Trojan horse infection" is 10, and the score of "horse infection" is 6. For example, when phrases having a score of 10 or more are adopted as detection conditions, the detection condition determination unit 22 determines that two of "Trojan horse" and "Trojan horse infection" are appropriate for the detection conditions.

[0082]

All the phrases (seven phrases) are subjected to feature degree calculation and determination in Application Example 2, whereas only the three phrases "Trojan horse", "Trojan horse infection", and "horse infection" are subjected to feature degree calculation and determination in Application Example 5. However, Application Example 2 and Application Example 5 have the same determination results and equal precision.

[0083]

As a result, a calculation amount can be reduced while maintaining precision.

[0084]

-Supplement-

Application Example 1 mainly explains the details of claim 4 and claim 7. Application Example 2 mainly explains claim 3 except claim 4. Application Examples 3 and 4 mainly explain claim 5 and claim 6.

Application Example 5 mainly explains claim 8.

[0085]

The present invention is a device for generating a dictionary used in a text information monitoring system, and can also be applied to a rumor
5 monitoring system or a reputation extraction system, targeted for the Internet, or the like.

[0086]

-Supplemental Notes-

In the exemplary embodiments described above, each unit may be
10 constructed with hardware, or may be achieved by a computer program. In this case, functions and operations similar to those mentioned above are achieved by a processor operated by the program stored in a program memory. Only some functions may also be achieved by the computer program.

15 [0087]

Some or all of the exemplary embodiments described above can also be described as in the following supplemental notes, but are not limited to the following.

[0088]

20 The present invention is a dictionary generation device for monitoring text information, which is used in a text information monitoring system and generates a dictionary in which a detection condition is registered, the dictionary generation device for monitoring text information, including:

25 a feature degree calculation unit that calculates, for a phrase as a candidate for the detection condition, a feature degree representing a degree by which the phrase matches an information content targeted for monitor; and

a phrase usefulness determination unit that determines whether or

not the phrase is appropriate for the detection condition based on the feature degree and a usefulness degree representing littleness of ambiguity of meaning defined by the phrase.

[0089]

5 In the dictionary generation device for monitoring text information of the present invention, the phrase usefulness determination unit preferably includes:

 a usefulness degree calculation unit which calculates the usefulness degree based on the length of a phrase; and

10 a detection condition determination unit which determines whether or not the phrase is appropriate for a detection condition based on the usefulness degree calculated by the usefulness degree calculation unit and on the feature degree.

[0090]

15 In the dictionary generation device for monitoring text information of the present invention, the usefulness degree calculation unit more preferably calculates a usefulness degree based on the length of the phrase and on a frequency in a document group.

[0091]

20 In general, the longer length of a phrase results in the less ambiguity of meaning and in a higher matching rate for a detection condition. In the present invention, priority is given to a phrase having a longer length by the configuration described above. As a result, it is possible to achieve high-precision detection compared to the conventional
25 art.

[0092]

For example,

 the usefulness degree calculation unit calculates a usefulness degree based on the product of the length of a phrase or the logarithmic

value thereof and a frequency in a document group or the logarithmic value thereof.

[0093]

In the dictionary generation device for monitoring text information
5 of the present invention, the usefulness degree calculation unit preferably
calculates a usefulness degree based on the length of the phrase, a
frequency in a document group, and an index representing an inclusion
relationship between phrases.

[0094]

10 More preferably,
when another phrase that is longer than a phrase of interest includes
the phrase of interest,

the index representing an inclusion relationship between phrases is
a ratio between the total of a frequency at which the other phrase appears
15 and the number of the other phrase.

[0095]

Consideration of the inclusion relationship results in the lower
value of a phrase included in the other longer phrase, eliminates addition
of a redundant detection condition, and enables improvement of dictionary
20 precision.

[0096]

In the dictionary generation device for monitoring text information
of the present invention, preferably,

the detection condition determination unit determines whether or
25 not a phrase is appropriate for a detection condition based on the product
of the usefulness degree or the logarithmic value thereof and the feature
degree or the logarithmic value thereof.

[0097]

As a result, it is possible to carry out detection in consideration of a

usefulness degree.

[0098]

In the dictionary generation device for monitoring text information of the present invention, more preferably,

5 for a phrase of which the usefulness degree calculated by the usefulness degree calculation unit is not less than a threshold value, the feature degree calculation unit calculates a feature degree, and the detection condition determination unit determines whether or not the phrase is appropriate for a detection condition.

10 [0099]

As a result, a calculation amount can be reduced while maintaining precision.

[0100]

15 The present invention is a dictionary generation method for monitoring text information, which is a method for generating a dictionary used in a text information monitoring system,

wherein a dictionary generation device for monitoring text information

20 calculates, for a phrase as a candidate for a detection condition, a feature degree representing a degree by which the phrase matches an information content targeted for monitor;

determines whether or not the phrase is appropriate for the detection condition based on the feature degree and a usefulness degree representing littleness of ambiguity of meaning defined by the phrase; and

25 outputs a phrase judged to be appropriate and registers the phrase as the detection condition.

[0101]

In the dictionary generation method for monitoring text information of the present invention, preferably,

the usefulness degree is calculated based on the length of a phrase;
and

whether or not the phrase is appropriate for a detection condition is
determined based on the usefulness degree and the feature degree.

5 [0102]

More preferably, a usefulness degree is calculated based on the
length of the phrase and on a frequency in a document group.

[0103]

For example,

10 a usefulness degree is calculated based on the product of the length
of a phrase or the logarithmic value thereof and a frequency in a document
group or the logarithmic value thereof.

[0104] In the dictionary generation method for monitoring text information
of the present invention, preferably,

15 a usefulness degree is calculated based on the length of the phrase,
a frequency in a document group, and an index representing an inclusion
relationship between phrases.

[0105]

More preferably,

20 when another phrase that is longer than a phrase of interest includes
the phrase of interest,

the index representing an inclusion relationship between phrases is
a ratio between the total of a frequency at which the other phrase appears
and the number of the other phrase.

25 [0106]

In the dictionary generation method for monitoring text information
of the present invention, preferably,

whether or not a phrase is appropriate for a detection condition is
determined based on the product of the usefulness degree or the

logarithmic value thereof and the feature degree or the logarithmic value thereof.

[0107]

In the dictionary generation method for monitoring text information
5 of the present invention, more preferably,

for a phrase of which the usefulness degree calculated by the
usefulness degree calculation unit is not less than a threshold value,
a feature degree is calculated, and

whether or not the phrase is appropriate for a detection condition is
10 determined.

[0108]

The present invention is a dictionary generation program for
monitoring text information, which causes a dictionary generation device
for monitoring text information to execute

15 a processing of calculating, for a phrase as a candidate for a
detection condition, a feature degree representing a degree by which the
phrase matches an information content targeted for monitor;

a processing of determining whether or not the phrase is appropriate
for the detection condition based on the feature degree and a usefulness
20 degree representing littleness of ambiguity of meaning defined by the
phrase; and

a processing of outputting a phrase judged to be appropriate and of
registering the phrase as the detection condition.

[0109]

25 The dictionary generation program for monitoring text information
of the present invention preferably causes

a processing of calculating the usefulness degree based on the
length of a phrase; and

a processing of determining whether or not the phrase is appropriate

for a detection condition based on the usefulness degree and the feature degree

to be executed.

[0110]

5 In the dictionary generation program for monitoring text information of the present invention, more preferably,

a usefulness degree is calculated based on the length of the phrase and on a frequency in a document group in the usefulness degree calculation processing.

10 [0111]

For example,

a usefulness degree is calculated based on the product of the length of a phrase or the logarithmic value thereof and a frequency in a document group or the logarithmic value thereof in the usefulness degree calculation processing.

15 [0112]

In the dictionary generation program for monitoring text information of the present invention, preferably,

a usefulness degree is calculated based on the length of the phrase, 20 a frequency in a document group, and an index representing an inclusion relationship between phrases in the usefulness degree calculation processing.

[0113]

More preferably,

25 when another phrase that is longer than a phrase of interest includes the phrase of interest,

the index representing an inclusion relationship between phrases is a ratio between the total of a frequency at which the other phrase appears and the number of the other phrase.

[0114]

In the dictionary generation program for monitoring text information of the present invention, preferably,

whether or not a phrase is appropriate for a detection condition is
 5 determined based on the product of the usefulness degree or the logarithmic value thereof and the feature degree or the logarithmic value thereof in the detection condition determination processing.

[0115]

In the dictionary generation program for monitoring text
 10 information of the present invention, more preferably,

for a phrase of which the usefulness degree calculated by the usefulness degree calculation processing is not less than a threshold value,
 a feature degree is calculated in the usefulness degree calculation processing, and

15 whether or not the phrase is appropriate for a detection condition is determined in the detection condition determination processing.

[0116]

This application is based upon and claims the benefit of priority
 20 from Japanese patent application No. 2012-213536, filed on September 27, 2012, the disclosure of which is incorporated herein in its entirety by reference.

[Reference signs List]

[0117]

25 1 Phrase extraction unit
 2 Phrase usefulness determination unit
 3 Feature degree calculation unit
 4 Output unit
 21 Usefulness degree calculation unit

22 Detection condition determination unit

[Document Name] CLAIMS

[Claim 1]

A dictionary generation device for monitoring text information, that
5 is used in a text information monitoring system and generates a dictionary
in which a detection condition is registered, the dictionary generation
device comprising:

a feature degree calculation unit that calculates, for a phrase to be a
candidate for the detection condition, a feature degree representing a
10 degree by which the phrase matches an information content targeted for
monitor; and

a phrase usefulness determination unit that determines whether or
not the phrase is appropriate for the detection condition based on the
feature degree and a usefulness degree representing littleness of ambiguity
15 of meaning defined by the phrase.

[Claim 2]

The dictionary generation device for monitoring text information
according to claim 1, wherein the phrase usefulness determination unit
includes:

20 a usefulness degree calculation unit which calculates the usefulness
degree based on the length of the phrase, and

a detection condition determination unit which determines whether
or not the phrase is appropriate for a detection condition based on the
usefulness degree calculated by the usefulness degree calculation unit and
25 on the feature degree.

[Claim 3]

The dictionary generation device for monitoring text information
according to claim 2, wherein

the usefulness degree calculation unit calculates a usefulness

degree based on the length of the phrase and on a frequency in a document group.

[Claim 4]

The dictionary generation device for monitoring text information
5 according to claim 3, wherein

the usefulness degree calculation unit calculates a usefulness degree based on the product of the length of the phrase or the logarithmic value thereof and a frequency in a document group or the logarithmic value thereof.

10 [Claim 5]

The dictionary generation device for monitoring text information according to claim 2, wherein

the usefulness degree calculation unit calculates a usefulness degree based on the length of the phrase, a frequency in a document group,
15 and an index representing an inclusion relationship between phrases.

[Claim 6]

The dictionary generation device for monitoring text information according to claim 5, wherein

when another phrase that is longer than a phrase of interest includes
20 the phrase of interest,

the index representing an inclusion relationship between phrases is a ratio between the total of a frequency at which the other phrase appears and the number of the other phrase.

[Claim 7]

25 The dictionary generation device for monitoring text information according to claims 2 to 6, wherein

the detection condition determination unit determines whether or not the phrase is appropriate for a detection condition based on the product of the usefulness degree or the logarithmic value thereof and the feature

degree or the logarithmic value thereof.

[Claim 8]

The dictionary generation device for monitoring text information according to claims 2 to 7, wherein

5 for the phrase of which the usefulness degree calculated by the usefulness degree calculation unit is not less than a threshold value, the feature degree calculation unit calculates a feature degree, and the detection condition determination unit determines whether or not the phrase is appropriate for a detection condition.

10 [Claim 9]

A dictionary generation method for monitoring text information, that is a method for generating a dictionary used in a text information monitoring system,

15 wherein a dictionary generation device for monitoring text information

calculates, for a phrase as a candidate for a detection condition, a feature degree representing a degree by which the phrase matches an information content targeted for monitor;

20 determines whether or not the phrase is appropriate for the detection condition based on the feature degree and a usefulness degree representing littleness of ambiguity of meaning defined by the phrase; and outputs the phrase judged to be appropriate and registers the phrase as the detection condition.

[Claim 10]

25 A dictionary generation program for monitoring text information, that causes a dictionary generation device for monitoring text information to execute

a processing of calculating, for a phrase as a candidate for a detection condition, a feature degree representing a degree by which the

phrase matches an information content targeted for monitor;

a processing of determining whether or not the phrase is appropriate for the detection condition based on the feature degree and a usefulness degree representing littleness of ambiguity of meaning defined by the

5 phrase; and

a processing of outputting the phrase judged to be appropriate and of registering the phrase as the detection condition.