30 OCT 2002

166166

G00002

INFORMATION PROCESSING SYSTEM

BACKGROUND OF THE INVENTION

The present invention relates to an information processing system which is employed in language discrimination to discriminate a language in a document and keyword search in a full text search to search/register an input text while observing all texts (character string) contained in an input document as an object.

As the method conducted in the prior art to discriminate the language of the text (character strings) described in the document in the field of information searching process, there are the method which discriminates the language based on discrimination of words by providing a word dictionary of the language, as disclosed in Unexamined Japanese Patent Publication Hei 8-137886; the method which discriminates the language based on particular bits (two bits herein) of the character codes of the language, as disclosed in Unexamined Japanese Patent Publication Hei 8-160929; and the method which discriminates the language based on particular information (e.g., the domain name, etc.) peculiar to the language, as disclosed in Unexamined Japanese Patent Publication Hei 10-171810.

Also, as the method conducted in the prior art in the keyword search to register/search the text described in the document in the field of information searching process, there is

the method which employs a word table in which separating characters are added to the beginning and the end of the keyword, as disclosed in Unexamined Japanese Patent Publication Hei 8-16617.

In addition to the above, as the prior art, the multiple language inputting system which can input effectively the texts in which multiple languages are mixed is disclosed in Unexamined Japanese Patent Publication Hei 5-282360. More particularly, the system converts the input text into a display text with consulting respective dictionaries, then defines the display text, and then decides the language in which the number of unknown words becomes the minimum as the input language or uses preferentially the precedingly utilized language as the input language.

In Unexamined Japanese Patent Publication Hei 7-262188, the language discrimination processing method which is able to perform the process of discriminating the language or genre of the stored or transmitted document is disclosed. Ordinary words of the object language are contained in the word occurring frequency table of the object language/genre, and the word codes have associated normalized occurring frequency values respectively, and the language/genre are detected by using such normalized occurring frequency values. Also, as the language/genre discriminating process, there is disclosed the system which receives a series of words from the input document, then compares the words with all words in the word occurring

frequency table, then adds the associated normalized occurring frequency values of the words to the existing total sum in the associated accumulator, and then discriminates the total value accumulated from all accumulators as the language discriminated value by reading the words of the document.

In Unexamined Japanese Patent Publication Hei 10-124513, there is disclosed the language discriminating method and system which discriminates the language described in the document by using sets of the words which are most frequently used in a plurality of candidate languages respectively. Respective sets of the most frequently used words are stored in respective sets of the word table for respective candidate languages according to character pairs in respective sets of the most frequently used words. Respective word tables are the N×N bit table. The bit denotes the given character pair at one location for the most frequently used words in one candidate language. The words from the document are compared with the most frequently used words being stored in the word table. The number of match between the words from the document and the words stored in respective sets in the word table is counted and then maintained for individual languages. The candidate language having the largest number of the match is selected as the language of the document.

However, in the method explained in the above in the prior art, for example, in Unexamined Japanese Patent Publication Hei 8-137886, the automatic discrimination can be achieved every

document, but the dictionaries must be provided to discriminate the words. In Unexamined Japanese Patent Publication Hei 8-160929, it is difficult to respond to different languages such as JCK integrated kanji by the same character codes such as the unicode, etc. In Unexamined Japanese Patent Publication Hei 10-171810, there is the problem that the discrimination is carried out every client.

In the method explained in the above in the prior art (e.g. JP 8-16617), there is the problem that the word table file must be prepared separately from the index used in the original document search.

In the method explained in the above in the prior art (e.g. JP 5-282360), there is the problem that, if the same character codes occurs in the text, it is difficult to decide which language such character code belong to. Also, there is the problem that, since the dictionaries used to discriminate the language must be prepared previously, it is impossible to discriminate the language against unknown terms. Further, there is the problem that the already known words, i.e., "recognition" or the like in the case of Japanese, can be recognized as the words because such words are contained in the dictionary as the dictionary data, nevertheless the words which are not contained in the dictionary but appear in the ordinary sentences cannot be discriminated because such words are not contained in the dictionary according to the conventional configuration. Furthermore, there is the

problem that, since rules for cutting out the word must be prepared for every discriminated language, no process is applied to the words unless there exists the rule to be applied to the particular language.

SUMMARY OF THE INVENTION

The present invention has been made to overcome the above-mentioned problems in the prior art. First, an object of the present invention is to discriminate a language of an input text described in a certain language by utilizing descriptive features of the language in the country, i.e., by detecting an occurring rate of the particular character of the text as a discriminated object, or detecting an average of the text length between the particular characters, or detecting an occurring rate of the character in a particular range, while observing particular characters (e.g., the space character in Korean) which are frequently occurred in the language or particular characters in the particular range (e.g., "hiragana", "kanji" in Japanese).

Second, another object of the present invention is to utilize the search system using two-character chains as the index prepared by the document search, and to enable the keyword search by adding information of separating characters to the index prepared at the time of registration, as they are, in the input text in which the separating characters are inserted into the beginning and the end of the keyword.

Third, still another object of the present invention is

to provide an information processing system which can discriminate automatically the language to which the given text belongs based on the given text, and to provide an information processing system which can discriminate automatically the language or the group type under the condition features of the text are decided as codes, particularly such features are classified by character pattern profiles, e.g., fonts, etc.

In order to overcome the above subjects, according to the present invention, firstly, means for detecting occurring rates of particular characters from occurring numbers of times of particular characters of an input text, means for storing standard occurring rates of the particular characters of a detected target language, and means for comparing the occurring rates of the particular characters of the input text with the standard occurring rates of the particular characters of the detected target language detected are provided in an information processing system. Secondly, means for detecting an average text length of unparticular texts between the particular characters of the input text, means for storing a standard average text length of the unparticular characters of the detected target language, and means for comparing the average text length of the unparticular characters of the input text with the standard average text length of the unparticular characters of the detected target language are provided in an information processing system. Thirdly, means for detecting occurring rates of characters-in-a-particular-

range based on occurring numbers of times of the characters-in-a-particular-range of the input text, means for storing standard occurring rates of the characters-in-a-particular-range of the detected target language, and means for comparing the occurring rates of the character-in-a-particular-range of the input text with the standard occurring rates of the character-in-a-particular-range of the detected target language are provided in an in formation processing system. Therefore, such an advantage can be achieved that the target language in the input text can be discriminated by comparing the occurring rate of the particular character in the input text derived from the occurring rate calculator with the standard occurring rate derived from the standard occurring rate memory to decide whether or not the input text corresponds to the text having the features which mate with the target language.

Also, according to the present invention, fourthly, means for inserting separating characters into library texts upon keyword search, means for detecting three-character chains in which the separating character is put in a middle if the separating characters are found in a separating character inserted text, means for generating a character chain consisting of characters located before and after the separating character, a character chain consisting of the separating character and a character located before the separating character, and a character chain consisting of the separating character and a character located

after the separating character, and means for inserting the separating character into a beginning of the input text or an end of the input text or the beginning and the end of the input text upon searching are provided. Therefore, such an advantage can be achieved that the search by using the text other than the keywords can be eliminated.

Also, according to the present invention, fifthly, in an information processing system which pattern-identifies uniquely a language, in which text data are contained, out of several predetermined language categories, a storing medium for recording programs which constitute pattern groups consisting of combinations of character code groups for character codes in which coded characters employed for pattern identification are classified into a plurality of one-dimensional code groups as code tables, and having a pattern storing means for storing a first data consisting of an n-character combination, and composed of plural character groups in any of which codes constituting combinations are contained, a second data having numerical values corresponding to combination patterns of the first data, and a third data assigned to predetermined pattern classification from combinations of the first data and the second data is provided. Further, a character chain extracting means for detecting two-character chains from a text, and deciding a character group in which characters constituting two-character chain are contained, a pattern calculating means for comparing whether or

not a combination of the character groups decided by the character chain extracting means coincides with the pattern in a pattern storing means, and calculating a numerical value of a coincident pattern, and a pattern deciding means for deciding a pattern having maximum numbers of values of all character chains calculated by the pattern calculating means and their combination patterns, which coincide with those in the first data, the second data, or the third data stored in the pattern storing means, as the pattern combination of a pattern storing means are provided. Furthermore, the pattern deciding means which calculates an average value of numerical values calculated by the pattern calculating means for all character chains, and decides the value which is closest to the second data of the pattern storing means as the pattern combination of the pattern storing means, or the pattern deciding means which calculates a most frequent value of numerical values calculated by the pattern calculating means for all character chains, and decides the value which is closest to the second data of the pattern storing means as the pattern combination of the pattern storing means is provided. Therefore, such an advantage can be achieved that the pattern to which the data should belong can be decided automatically from the sets of the patterns, in which the are assigned to all the patterns such as the character sets, etc. and the patterns are classified into several type meaning pattern groups and a plurality of meaning groups are expressed by a combination of pattern groups.

Also, according to the present invention, sixthly, a storing medium which constitutes pattern groups consisting of combinations of pattern groups for pattern identification and stores a first data consisting of two-character combination patterns for the pattern group in which pattern groups are previously known, a second data consisting of pattern types indicating the first data, and a third data showing occurring numbers of times of combinations of the first data and the second data, a character chain extracting means for extracting all two-character chain of the pattern data to which pattern types are previously assigned from the storing medium, a character chain frequency counting means for calculating occurring number of times of respective character chains and storing character chain patterns and the occurring number of times in the storing medium every pattern type, the character chain extracting means for extracting two-character chains from a given pattern, a frequency counting means for extracting occurring number of times of the character chain pattern corresponding to the extracted character chain from the storing medium every pattern type, and a collating means for collating the occurring number of times derived from the frequency counting means with all extracted character chains, and then deciding the pattern type which has a largest total occurring number of times as the pattern type in which the given pattern is contained are provided. In this case, the collating means decides the pattern type whose total occurring number of

times exceeds a certain threshold value as the pattern type in which the given pattern is contained, or the collating means decides the pattern type which has a largest total occurring frequency of the character chain pattern whose occurring number of times exceeds a certain threshold value over all character chains as the pattern type in which the given pattern is contained. Therefore, such an advantage can be achieved that the pattern to which the test data belongs can be automatically decided and, unless character types, fonts, etc. are coded, the pattern can be classified, and in addition the pattern can be classified without preparing the dictionary language by language.

Also, according to the present invention, seventhly, a separating means for separating character types or at least one language into words when all character chains are extracted from pattern data to which pattern types are assigned previously in the storing medium which constitutes pattern groups consisting of combinations of pattern groups for pattern identification, a character chain extracting means for extracting character chains based on the separating means, a chain converting means for repeating a same character to form character chains if the character chain consists of one character, the character chain frequency counting means, the character chain extracting means, and the frequency counting means are provided. Therefore, such an advantage can be achieved that the pattern to which the test data belongs can be automatically decided, and there is such an

advantage that, unless character types, fonts, etc. are coded, the pattern can be classified, and in addition the pattern can be classified without the preparation of all rules for respective languages.

Also, according to the present invention, eighthly, a pattern storing means for storing document numbers of character chain patterns consisting of two-character combination, and character chain patterns of a text data every pattern classification or language is provided in the storing medium which constitute pattern groups consisting of combinations of character code groups employed in pattern identification. And, a character chain extracting means for extracting all two-character chains of a text data, a character chain frequency counting means for calculating occurring number of times of respective character chains and calculating the character chain patterns and document numbers every pattern type, and a character frequency comparing means for fetching frequencies of character chain patterns obtained by the character chain frequency counting means and character chain patterns in a storing medium every pattern classification, comparing total values of the frequencies of the character chain patterns for each pattern classification, and storing a character chain pattern, an occurring number of times, and a document number into the pattern classification of the storing medium having a larger pattern classification are provided. In this case, the character frequency comparing means

decides the pattern type whose total occurring number of times exceeds a certain threshold value as the pattern type in which the given pattern is contained, or the character frequency comparing means decides the pattern type which has a largest total occurring frequency of the character chain pattern whose occurring number of times exceeds a certain threshold value over all character chains as the pattern type in which the given pattern is contained. Therefore, such an advantage can be achieved that the pattern to which the test data belongs can be automatically decided, and, unless character types, fonts, etc. are coded, the pattern can be classified.

Also, according to the present invention, ninthly, in an information processing system which pattern-identifies uniquely a language, in which text data are contained, out of several predetermined language categories in which character codes are defined every character font, a storing medium which constitute pattern groups consisting of combinations of character code groups which are employed for pattern identification and stores two-character combinations constituting characters every language in an individual language distribution table in which an occurring rate of a first character are stored per character type of kanji, hiragana/katakana, symbol, Hangul, and other characters, an individual code designating means for reading an integrated code or a local code as an employed character code system, a character chain extracting means for extracting all

30 OCT 2002

two-character chain from input text data, an occurring number-of-times counting means for classifying two-character chains of kanji, hiragana/katakana, symbol, Hangul, and others in compliance with a designated code system every language, and then calculating their occurring rates respectively, an individual language interdistribution distance calculating means for a total of distances between the occurring rates of kanji, hiragana/katakana, symbol, Hangul every language according to designated code system and the occurring rates of kanji, hiragana/katakana, symbol, Hangul every language managed by the storing means, and a comparing means for causing the individual language interdistribution distance calculating means to calculate the smallest distance value, and then deciding a language category having the smallest distance value as a language of the input text are provided. Therefore, since distances between the occurring distributions of kanji, hiragana, symbol, etc. can be calculated, the language category can be easily decided.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG.1 is a block diagram showing a configuration of an information processing system according to a first method embodiment of the present invention;

FIG.2 is a conceptional view showing a first method of character code language discrimination according to the present invention;

FIG.3 is a block diagram showing a configuration of an

information processing system according to a second method embodiment of the present invention;

FIG.4 is a conceptional view showing a second method of character code language discrimination according to the present invention;

FIG.5 is a block diagram showing a configuration of an information processing system according to a third method embodiment of the present invention;

FIG.6 is a conceptional view showing a third method of character code language discrimination according to the present invention;

FIG.7 is a block diagram showing a configuration of an information processing system according to a fourth method embodiment of the present invention;

FIG.8 is a conceptional view showing a fourth method of searching a keyword according to the present invention;

FIG.9 is a view showing a code table according to a fifth method embodiment of the present invention;

FIG.10 is a view showing a configuration of a pattern storing device according to the fifth method embodiment of the present invention;

FIG.11 a conceptional view showing a text as a discrimination object according to the fifth method embodiment of the present invention;

FIG.12 is a view showing a data structure in the middle

of pattern discrimination according to the fifth method embodiment of the present invention;

FIG.13 is a block diagram showing a configuration of an information processing system according to the fifth method embodiment of the present invention;

FIG.14 is a conceptional view showing sample data according to a sixth method embodiment of the present invention;

FIG.15 is a view showing a character chain occurring rate according to the sixth method embodiment of the present invention;

FIG.16 is a conceptional view showing data structures of a frequency memory and the data according to the sixth method embodiment of the present invention;

FIG.17 is a conceptional view showing a test text according to the sixth method embodiment of the present invention;

FIG.18 is a block diagram showing a configuration of an information processing system according to the sixth method embodiment of the present invention;

FIG.19 is a conceptional view showing sample data according to a seventh method embodiment of the present invention;

FIG.20 is a conceptional view showing chain-converted data according to the seventh method embodiment of the present invention;

FIG.21 is a conceptional view showing data structures of a frequency memory and the data according to the seventh method embodiment of the present invention;

FIG.22 is a block diagram showing a configuration of an information processing system according to the seventh method embodiment of the present invention;

FIG.23 is a conceptional view showing a code table according to an eighth method embodiment of the present invention;

FIG.24 is a conceptional view showing character chain data according to the eighth method embodiment of the present invention;

FIG.25 is a conceptional view showing character chain data at the time of character chain collation according to the eighth method embodiment of the present invention;

FIG.26 is a conceptional view showing data structures of a frequency memory and the data according to the eighth method embodiment of the present invention;

FIG.27 is a block diagram showing a configuration of an information processing system according to the eighth method embodiment of the present invention;

FIG.28 is a view showing a configuration of an individual character set code table memory according to a ninth method embodiment of the present invention;

FIG.29 is a view showing a configuration of an individual language distribution table memory according to the ninth method embodiment of the present invention;

FIG.30 is a conceptional view showing input text data according to the ninth method embodiment of the present invention;

FIG.31 is a view showing a method of calculating distance decision according to the ninth method embodiment of the present invention; and

FIG.32 is a block diagram showing a configuration of an information processing system according to the ninth method embodiment of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

Embodiments of the present invention will be explained with reference to the accompanying drawings hereinafter.

(Embodiment 1)

FIG.1 shows a configuration of a character code language discriminating system according to a first method embodiment of the present invention.

In FIG.1, 101 denotes a particular character discriminator for discriminating whether or not the character in the input text (character string) corresponds to the particular character of the detected target language; 102, a particular character counter for counting the occurring number of time of the particular character discriminated by the particular character discriminator; 103, an input character counter for counting the occurrence of all characters of the input text; 104, an occurring rate calculator for calculating the occurring rate of the particular character based on the occurring number of time of the particular character counted by the particular character counter 102 and the number of character of the input text counted

by the input character counter 103; 105, a standard occurring rate memory for storing the standard occurring rate of the particular character of the detected target language; 106, a comparator for comparing the occurring rate detected by the occurring rate calculator 104 with the standard occurring rate stored in the standard occurring rate memory 105; and 107, an end-of-text detector for detecting the end of the input text.

FIG.2 shows a conception of a first method of the character code language discrimination according to the present invention.

In FIG.2, 201 denotes the input text to which the language discrimination is applied, and 202 denotes the particular character occurred in the input text. In the first method of the present invention, the target language is discriminated by detecting the occurring rate of the particular character in the input text. Assume that the total number of character of the input text 201 is m and the occurring number of time of the particular character 202 is n, the occurring rate k of the particular character in the input text 201 is given as n/m. Assume that the occurring rate of the particular character in the target language is a at a minimum and is b at a maximum, the first method of the present invention decides at this time that the input text corresponds to the target language if the occurring rate k of the particular character in the input text 201 is $a \leqq k \leqq b$. With the above, the language discrimination of the text can be achieved.

The present invention is effective when Korean in the

input text is discriminated by designating the "space character" as the particular character 202 like Korean. At this time, a=0.14, b=0.23 can be applied based on statistical data of the Korean newspaper account data consisting of one hundred and twenty thousand characters. According to this method, the document described in Korean Hangul can be discriminated from other document, e.g., the document described in Japanese.

As described above, when the end of the text is detected by the end-of-text detector 107, the target language in the input text can be discriminated by comparing the occurring rate of the particular character in the input text derived from the occurring rate calculator 104 with the standard occurring rate derived from the standard occurring rate memory 105 to decide whether or not the input text corresponds to the text having the features which mate with the target language.

(Embodiment 2)

FIG.3 shows a configuration of a character code language discriminating system according to a second method embodiment of the present invention.

In FIG.3, 301 denotes a particular character discriminator for discriminating whether or not the characters in the input text corresponds to the particular character of the detected target language; 302, a text length counter for counting the successive occurring number of the unparticular characters which are not discriminated by the particular character

discriminator 301; 303, an adder for counting a total sum of the text length output from the text length counter 302; 304, a particular character counter for counting the occurring number of times of the particular character discriminated by the particular character discriminator 301; 305, an average text length calculator for calculating an average text length by dividing the total sum of the text length calculated by the adder 303 by the number of the particular characters derived from the particular character counter 304; 306, a standard text length memory for storing the average text length of the detected target language; 307, a comparator for comparing the average text length obtained by the average text length calculator 305 with the standard average text length stored in the standard text length memory 306; and 308, an end-of-text detector for detecting the end of the input text.

result

FIG.4 shows a conception of a second method of the character code language discrimination according to the present invention.

In FIG.4, 401 denotes the input text to which the language discrimination is applied; 402, the particular character occurred in the input text; and 411, 412, 41n, texts of unparticular characters (unparticular texts) partitioned by the particular character. In the second method of the present invention, the target language is discriminated by detecting the average length

of the unparticular texts 411, 412, 41n, which is given by partitioning the input text 401 by the particular character 402. Assume that the unparticular characters of the input text 401 are partitioned by n-1 particular characters and thus the input text consists of the n unparticular texts and the unparticular text lengths are a1, a2, a3,...,an at this time, the average text length k of the unparticular texts in the input text 401 is given as (a1+a2+a3+...+an)/n. Assume that the average text length of the unparticular characters partitioned by the particular character in the target language is a at a minimum and is b at a maximum, the discriminating method according to the second method of the present invention decides at this time that the input text 401 corresponds to the target language if the average text length k of the unparticular character in the input text 401 is $a \leqq k \leqq b$. With the above, the language discrimination of the text can be achieved.

According to the present method, Korean in the input text can be discriminated by designating the "space character" as the particular character 402 like Korean, for example. At this time, a=3, b=5 can be applied based on statistical data of the Korean newspaper account data consisting of one hundred and twenty thousand characters.

Also, in the discrimination between Japanese and Chinese, for example, the language in the input text can be discriminated by designating "," ("、" in the case of Chinese), "的", and "。

" as the particular character 402. At this time, a=10, b=22 can be applied based on statistical data of the Japanese newspaper account data consisting of fifty-four thousand characters, and also a=4, b=9 can be applied based on statistical data of the Chinese statute document data consisting of eighty-four thousand characters. According to this method, it can be easily discriminated by comparing the above average text lengths that the descriptive language of the input text corresponds to either Japanese or Chinese.

As described above, when the end of the text is detected by the end-of-text detector 308, the target language in the input text can be discriminated by comparing the average text length of the input text derived from the average text length calculator 305 with the standard text length derived from the standard text length memory 306 to decide whether or not the input text corresponds to the text having the features which mate with the target language.

(Embodiment 3)

FIG.5 shows a configuration of a character code language discriminating system according to a third method embodiment of the present invention.

In FIG.5, 501 denotes a character-in-a-particular-range discriminator for discriminating whether or not the character in the input text corresponds to the character in a particular range of the detected target language; 502, a character-in-a-

particular-range counter for counting the occurring number of times of the character in the particular range discriminated by the character-in-a-particular-range discriminator; 503, an input character counter for counting the occurrence of all character codes of the input text; 504, an occurring rate calculator for calculating the occurring rate of the character-in-a-particular-range based on the occurring number of times of the character-in-a-particular-range counted by the character-in-a-particular-range counter 502 and the character number of the input text counted by the input character counter 503; 505, a standard occurring rate memory for storing the standard occurring rate of the character-in-a-particular-range of the detected target language; 506, a comparator for comparing the occurring rate derived from the occurring rate calculator 504 with the standard occurring rate stored in the standard occurring rate memory 505; and 507, an end-of-text detector for detecting the end of the input text.

FIG.6 shows a conception of a third method of the character code language discrimination according to the present invention. In FIG.6, 601 denotes the input text to which the language discrimination is applied; 602, a range of the character code of the character in the particular range occurred in the input text; and 611, 612, 613, characters in the particular range contained in the range 602. In the third method of the present invention, the target language is discriminated by detecting the occurring

rate of the characters in the particular range of the input text. Assume that the total number of the characters of the input text 601 is m and the occurring number of times of the character contained in the particular range is n, the occurring rate k of the character in the particular range of the input text 601 is given as n/m. Assume that the occurring rate of the character in the particular range in the target language is a at a minimum and is b at a maximum, the discriminating method according to the third method of the present invention decides at this time that the input text 601 corresponds to the target language if the occurring rate k of the character in the particular range of the input text 601 is $a \leqq k \leqq b$. With the above, the language discrimination of the text can be achieved.

According to the present method, Japanese in the input text can be discriminated by designating "hiragana" or "kanji" as the range 602 of the character in the particular range like Japanese, for example. At this time, a=0.1, b=0.5 of "hiragana" and a=0.2, b=0.6 of "kanji" can be applied based on statistical data of the Japanese newspaper account data consisting of fifty-four thousand characters. According to this method, it can be easily discriminated whether or not the descriptive language of the input text corresponds to Japanese.

As described above, when the end of the text is detected by the end-of-text detector 507, the target language in the input text can be discriminated by comparing the occurring rate of the

character in the particular range of the input text derived from the occurring rate calculator 504 with the standard occurring rate derived from the standard occurring rate memory 505 to decide whether or not the input text corresponds to the text having the features which mate with the target language.

(Embodiment 4)

FIG.8 shows a conception of a fourth method of searching a keyword according to the present invention. In FIG.8, 801 denotes a text "いろはに" which is input at the time of registration; 811, 812, 813, a keyword respectively; 802, a separating character inserted text in which a separating character "a" is inserted into boundaries between the keywords 811, 812, 813 of the input text 801; 803, a two-character chain "aい" registered first into the separating character inserted text 802; 804, a two-character chain "いろ" occurred after 803; 805, a two-character chain "ろa" which contains the separating character generated from a three-character string "ろaは" in which the separating character "a" is put in the middle; 806, a two-character chain "aは" which contains the separating character of the three-character string "ろaは"; 807, a two-character chain "ろは" which does not contain the separating character to spread over the separating character in the three-character string "ろaは"; 808, a two-character chain "はに" occurred next to 807;

and 809, a two-character chain "にa" occurred after 808.

Also, in FIG.8, the two-character chain 803 stores the occurring numbers of times n, n1 of "a" and "い"; the two-character chain 804 stores the occurring numbers of times n1, n2 of "い" and "ろ"; the two-character chain 805 stores the occurring numbers of times n3, n of "ろ" and "a"; the two-character chain 806 stores the occurring numbers of times n, n3 of "a" and "は"; the two-character chain 807 stores the occurring numbers of times n2, n3 of "ろ" and "は"; the two-character chain 808 stores the occurring numbers of times n3, n4 of "は" and "に"; and the two-character chain 809 stores the occurring numbers of times n4, n of "に" and "a" respectively.

In addition, in FIG.8, 821 denotes a searched text "いろはに" upon searching; 822, a searched text in which the separating character "a" is inserted into the beginning and the end of 821 to carry out the full coincidence search by using the searched text 821 as the key word; 823, a two-character chain "aい" searched first in the searched text 822; 824, a two-character chain "いろ" occurred after 823; 825, a two-character chain "ろは" occurred after 824; 826, a two-character chain "はに" occurred after 825; and 827, a two-character chain "にa" occurred after

826.

FIG.7 shows a configuration of an information processing system as a text searching method according to a fourth method embodiment of the present invention.

In FIGS.7 and 8, 701 denotes a text converter for inserting the particular separating character "a" into boundaries between the keywords of the library texts 801; 702, a separating character detector for detecting the separating character from the library text 802 which is output from the text converter 701 and into which the separating character is inserted; 703, a two-character chain generator for detecting the two-character chains 804, 808 which are generated and registered when no separating character is found from the text 802; 704, a separating character chain generator for detecting from the text 802 the two-character chains 805, 806, 807 which are generated from the three-character string "ろaは" having the separating character "a" in the middle and the two-character chains 803, 809 which have the separating character at the beginning and the end of the text respectively; 705, a two-character chain memory for storing the two- character chains 803, 804, 805, 806, 807, 808, 809 which are detected by the two-character chain generator 703 and the separating character chain generator 704, wherein the separating character indicates a constant value and other characters indicate their occurring number of times; 706, a text converter for inserting the particular separating character "a" into the beginning and the end of the

searched text 821; 707, a two-character chain detector for detecting two-character chains 823, 824, 825, 826, 827 from the searched text 822 which is generated by the text converter 706; 708, a comparator for detecting the two-character chains 823, 824, 825, 826, 827 which are detected by the two-character chain detector 707 from the two-character chain memory 705 and then deciding whether or not the occurring number of times of the preceding character of the detected two-character chains coincides with the occurring number of times of the character after the two-character chain which is detected immediately before; and 709, a controlling portion for deciding the coincidence of the text by causing the comparator 708 to decide all two-character chains detected from the two-character chain detector 707. At this time, according to the fourth method of the present invention, the two-character chain 803 equivalent to "aい" of the two-character chain 823 is detected and also the occurring number of times n1 of "い" at this time is detected, and the two-character chain 804 equivalent to "いろ" of the two-character chain 824 after the chain 823 is detected and also the occurring numbers of times n1, n2 of "い" and "ろ" at this time are detected. It is decided whether or not the occurring number of times of "い" coincides with each other as n1 between the two-character chains 803 and 804. If yes, the two-character chain 807 equivalent to "ろは" of the two-character chain 825 is then detected and also the occurring

numbers of times n2, n3 of "ろ" and "は" at this time are detected. It is decided whether or not the occurring number of times of "ろ" coincides with each other as n2 between the two-character chains 804 and 807. If yes, the two-character chain 808 equivalent to "はに" of the two-character chain 826 is then detected and also the occurring numbers of times n3, n4 of "は" and "に" at this time are detected. It is decided whether or not the occurring number of times of "は" coincides with each other as n3 between the two-character chains 807 and 808. If yes, the two-character chain 809 equivalent to "にa" of the two-character chain 827 is then detected. Then, it is decided whether or not the occurring number of times n4 of "に" at this time coincides with the occurring number of times of "に" of the two-character chain 808. If yes, it is decided based on complete coincidence of the keywords that the text 821 coincides with the text 801. With the above, the collation of the text can be achieved.

According to the present system, the search by using the text other than the keyword can be eliminated. For example, if the search should be carried out by using the text "いろは", which is not the keyword, like the text 831 in FIG.8, such search is carried out by using the searched text like 832 in which the separating character is inserted into the beginning and the end. In this case, since the text 831 does not correspond to the

30 OCT 2002

keywords 811, 812, 813, "はa" of the two-character chain 836 which cannot be detected from the searched text 832 is generated and is not searched.

In this case, the present system may be similarly applied to the case where positional information of the characters are employed in place of the occurring number of times of the character in respective two-character chain.

(Embodiment 5)

An example of a character patter discriminating method will be explained as a fifth embodiment of the present invention.

FIG.9 shows a structure of a code table for coding the character patterns and storing them. In this case, pattern codes are formed one-dimensionally, and a unicode, etc. are selected as objects of the text code, for example. An A-group 901 is a group which consists of so-called hiragana and katakana such as characters "あ",..., "力",... Code numbers are assigned to respective characters of the A group 901 such that #1 is assigned to "あ", #2 is assigned to "い", etc. Similarly, a B-group 902 is a group which consists of kanji and code number starting #100 are assigned respective kanjis, and a C-group 903 is a group which consists of symbol such as ×, ○, △, etc. and code number starting #200 are assigned respective symbols. In this case, above code numbers are assigned for the sake of convenience, and any code numbers may be employed if they are unique in the code table.

Next, FIG.10 shows a configuration of a pattern storing means for storing types of the pattern classifications consisting of the characters and their data. At this time, assume that the text data formed of a combination of character sets is constructed by any one of the texts in the patterns to be discriminated. As for two sets of character patterns, the patterns and their values are stored in sets of the code table in which the first character pattern and the second character pattern are contained. If (first character, second character) coincide with (character in the A-group, character in the A-group) (referred simply to as (A-group, A-group) hereinafter) or (B-group, B-group) respectively, the occurring pattern corresponds to a combination D1 and defined to be contained in a pattern 1. A numerical value 1 is allocated to this pattern. Also, the numerical value 1 is assigned to combinations of (A-group, B-group) or (B-group, A-group) in the pattern 1. The numerical values are assigned to a pattern 2 and a pattern 3 like patterns 2, 3 in FIG.10. For example, since Japanese contains hiragana, katakana, and kanji, the pattern 1 can be defined as Japanese. Similarly, if the C group is assumed as Hangul, the pattern 2 can be defined as Korean and the pattern 3 can be defined as Chinese.

Next, a method of deciding to which pattern the character string belong will be explained hereunder. In FIG.11, 1101 is character string data. Two-character combinations are extracted from the character string data 1101. Two-character chains (日、

本), (あ、い), (漢、力) are extracted from the two-character combinations. (B-group, B-group), (A-group, A-group), and (B-group, A-group) are detected respectively as the groups in which the extracted character chains are contained. As a result, as shown in FIG.12, (日、本) belongs to the pattern 1 or 2, (あ、い) belongs to the pattern 1, and (漢、力) belongs to the pattern 1.

In addition, scores (numerical values) corresponding to these patterns can be given as 1 or 2, 1, 1. The numerical value 1 is detected as the most frequent numerical value when the numbers of these numerical values are compared mutually, and the pattern 1 is detected as a pattern to mate with the numerical value 1. Accordingly, it is decided that the text data is contained in the pattern 1. Although the numerical value having the highest frequency is calculated in the fifth embodiment, the pattern which has the smallest difference between an average value of the numerical values and the numerical values of the patterns may be decided as the pattern in which the text data is contained. In addition, the most frequent value of the numerical values may be calculated instead of the average value.

According to the present method, after actual data have been analyzed, it has been found that, in the case of Japanese (newspaper account as the object), an occurring rate of the symbol is about 10 % of all character sets, an occurring rate of hiragana/katakana, etc. except kanji/symbol is about 41 %, and

an occurring rate of kanji is about 49 %. In the case of Chinese (statute book as the object), an occurring rate of the symbol is about 3 % of all character sets, an occurring rate of hiragana/katakana except kanji/symbol is about 22 %, and an occurring rate of kanji is about 74 %. In the case of Hangul (newspaper account as the object), an occurring rate of the symbol is about 15 % of all character sets, an occurring rate of hiragana/ katakana, etc. except kanji/symbol is about 3 %, an occurring rate of Hangul character is about 73 %, an occurring rate of kanji is about 4 %, and an occurring rate of others is about 7 %. According to this, the character codes are managed only in the form of symbol, hiragana/ katakana, kanji, and Hangul character, the language is Japanese if a rate of hiragana/katakana except the symbol and kanji is 41:49, the language is Chinese if such rate is 22:74, and the language is Hangul if such rate is 3:73. The language closest to one of these rates can be classified as the language to be discriminated.

FIG.13 shows a configuration of an information processing system according to the fifth method embodiment of the present invention. In FIG.13, 1301 denotes a character chain extractor for cutting out two-character chain from the text data; 1303, a code table in which codes of the patterns are stored; 1305, a pattern memory in which pattern sets and the numerical values are stored; 1302, a pattern calculator for collating sets of character code groups constituting the extracted two-character chain with

those in the pattern memory 1305 and then calculating the numerical values of the corresponding patterns for all extracted two-character chains; and 1304, a pattern discriminator for calculating the occurring frequency per numerical value or the average value or the most frequent value based on the numerical values calculated by the pattern calculator 1302 and then calculating the pattern whose occurring frequency coincides mostly with the numerical value of the pattern stored in the pattern memory 1305 or a difference from the average value or a difference from the most frequent value to then decide the pattern having the smallest value.

Therefore, according to this method, the pattern to which the data should belong can be decided automatically from the sets of the patterns, in which the are assigned to all the patterns such as the character sets, etc. and the patterns are classified into several type meaning pattern groups and a plurality of meaning groups are expressed by a combination of pattern groups.

In particular, this method is effective in sorting when character types and fonts as patterns are classified into a plurality of patterns. The unicode is thought of as an example of the codes. In this case, if the pattern 1 is "hiragana/katakana", the pattern 2 is "kanji (CJK kanji)", and the pattern 3 is "Hangul", the language can be sorted by defining, for example, the meaning group consisting of a combination of the pattern group 1 and the pattern group 2 as "Japanese", the meaning

group consisting of a combination of the pattern group 2 and the pattern group 3 as "Korean", and the meaning group consisting of the pattern group 2 only as "Chinese".

In the fifth embodiment, the patterns are not limited to the character patterns if they are coded patterns, and objects are not particularly limited if they are coded objects such as symbols, figures, etc. like ◎, ◇, etc. Although the characters are cut out every two characters, no limitation is imposed on the unit of cutting out the character if two characters can be adopted.

(Embodiment 6)

An example of a method of discriminating the character pattern will be explained as a sixth embodiment of the present invention. At first, a method of forming information to discriminate the patterns will be explained hereunder.

FIG.14 shows sample data to form information. In FIG.14, 1401 denotes Chinese sample data, and two-character chains are picked up from this sample data. Then, 1402 denotes a character chain "那些", and 1403 denotes a character chain "些都". These character chains are picked up adjacently. The record number 1 is assigned to the data of the sample data 1401. Subsequently all character chains are picked up from the sample data. In FIG.14, assume that four character chains 1402 to 1405 are shown by way of later explanation. In FIG.14, 1406, 1412 are Japanese sample data, and all two-character chains are extracted from the sample data like the above. Like the explanation of the Chinese sample

data, character chains 1407 to 1411, 1413 to 1416 are shown as two-character chains in the sixth embodiment. In this case, record numbers 2, 3 are assigned to the data 1406, 1412.

Next, the combination patterns, the record numbers, and the occurring numbers of times of the extracted character chain patterns are stored in the storing medium individually according to the pattern types like Chinese and Japanese.

As occurring patterns 1501 to 1504 in FIG.15, patterns and record numbers are stored when the pattern type is Chinese. Similarly, as occurring patterns 1505 to 1509 in FIG.15, patterns and record numbers are stored when the pattern type is Japanese. Since the occurring patterns 1504 to 1504 contain the sample data 1401, the record number 1 is stored therefor. Also, since the occurring patterns 1505 to 1509 contain the sample data 1406 or 1412, the record number 2 is stored if only one record is contained, and the record numbers "2, 3" are stored if two records are contained. The occurring number of the character chains is indicated by the number of the record number stored in FIG.15, and shown as a two-dimensional table shown in FIG.16.

In FIG.16, 1601 denotes the occurring number of Chinese, and 1602 denotes the occurring number of Japanese. In the tables in FIG.16, an ordinate denotes the first character of combinations of the pattern (first character, second character) and an abscissa denotes the second character thereof. The data in FIGS.15 and 16 are stored in the pattern discrimination storing medium. The

patterns having the record number in any number are stored in the following in the storing medium of the sixth embodiment. As a result, the occurring number of times can be stored in the table every pattern type.

Although the explanation of two pattern types of Chinese and Japanese has been made in the sixth embodiment, a plurality of pattern types may be stored every pattern type in the table by the similar method. Next, a method of discriminating the language by using the test pattern will be explained. In FIG.17, adjacent two-character chains are extracted from the test pattern 1701. The two-character chains are extracted like a chain 1702 "同事", a chain 1703 "事件", a chain 1704 "件と", a chain 1705 "と我", and a chain 1706 "我々". Then, the occurring numbers of times corresponding to the chains 1702, 1703, 1704, 1705, 1706 are detected from the table of the memory medium in FIG.16. In the Chinese table (1601 in FIG.16), the occurring number of times of the pattern (同, 事) is 1 and another occurring number of times does not exist. In contrast, in the Japanese table (1602 in FIG.16), the occurring number of times of the pattern (我、々) is 2, the occurring number of times of the pattern (同, 時) is 1, and the occurring number of times of the pattern (事, 件) is 2. If a total number of these occurring numbers of times is calculated every pattern type, such number is 1 in Chinese and also such number is 5 in Japanese. The pattern type having the

larger numerical value is detected by comparing these two results, and as a result it can be decided that such test pattern corresponds to Japanese.

Although the total sum of the occurring number of times is calculated in the sixth embodiment, such a method may be adopted that the pattern type whose total occurring number of times exceeds n can be calculated, or the total of the occurring numbers of times of the character chains in which the occurring number of times of each character chain exceeds n can be calculated.

According to the present approach, the number of the occurring character types which have the large occurring rate in the local code using the same character codes in Japanese/Chinese/Hangul have been checked with interest among the examples which are applied to the actual data cited in the embodiment 5. At that time, 3469 type kanjis (about 74 %) are employed in the Chinese data; 218 type Hanguls (about 73 %) are employed in the Hangul data; 4006 type kanjis (about 49 %) are employed in the Japanese data; and 130 type hiragana/katakanas (about 23 %) (510 type (about 41 %) if the symbols are contained) are employed in the Japanese data. From the above combination of the occurring rate and the character type, the 74 % occurring frequency and the 3469 occurring character type in the Chinese kanji code, the 73 % occurring frequency and the 218 occurring character type in the Hangul character code, the 23 % occurring frequency and the 130 occurring character type in the Japanese

hiragana/katakana code, and the 49 % occurring frequency and the 4006 occurring character type in the Japanese kanji code are compared mutually, and then the language can be decided as Chinese if the occurring frequency and the occurring character type of kanji are large in the Chinese code. Similarly, the language can be decided as Japanese or Hangul based on the occurring frequency and the occurring character type in respective language codes.

FIG.18 shows a configuration of an information processing system according to the sixth method embodiment of the present invention. In FIG.18, 1801 denotes a character chain extractor for picking up the adjacent two-character chains from the text data; 1802, a character chain frequency counter for reading the occurring number of times and the record number of the character chain and then storing them in a frequency memory 1803 every pattern (language) type; 1804, a frequency counter fetching the occurring number of times of the character chains being detected by the character chain extractor 1801 from the frequency memory 1803 every pattern type; and 1805, a collator for calculating a total of the occurring number of times detected from the frequency counter 1804 or a total of the occurring number of times of a certain threshold value every pattern type and then deciding the pattern type in which the test pattern is contained by calculating a pattern exceeding the certain threshold value every pattern type or comparing magnitudes of them every pattern type.

Therefore, according to this method, the pattern to which

the test data belongs can be automatically decided. Also, there is such an advantage that, unless character types, phonts, etc. are coded, the pattern can be classified. In addition, the pattern can be classified without preparing the dictionary language by language.

(Embodiment 7)

An example of a method of discriminating the character patterns will be explained as a seventh embodiment of the present invention. First of all, a method of forming information to discriminate the pattern is explained hereunder.

FIG.19 showing the sample data to form the information. In FIG.19, 1901 denotes a sample pattern in Chinese, and two-character chains are picked out from this sample data. The two-character chains, when picked up, are formed if the character types (e.g., character codes such as kanji, katakana, etc.) are identical. In this case, if the previously designated language, e.g., the dictionary includes the language, is cut out when two characters are picked up, the chain between the cut-out second character and the subsequent character is not formed. In the seventh embodiment, since the cutting-out for Japanese is employed, adjacent character chains are formed for the sample pattern 1901. Subsequently, similar operations to those in the sixth embodiment are applied to two-character chains 1902 to 1905. Then, in FIG.19, 1906 denotes a sample pattern in Japanese. After chain formation is checked every character type, the

character chains of hiragana such as "は", "な", etc. are not formed. If the Japanese dictionary exists like the character chain "事件", the preceding character "同" and the two-character chain "事件" are separated. As a result, the character chain is separated into the two-character chains 1907, 1911 and one characters 1908, 1909, 1910, etc. In turn, the two-character chains of one characters 1908, 1909, 1910 are formed by repeating the same character. FIG.20 shows character chains which are formed by converting into two-character chains when the patterns are decided as one character.

Next, according to the sixth embodiment, combination patterns, record numbers, and occurring numbers of times of all character chain patterns are stored into the storing medium separately every pattern type like Chinese, Japanese, etc. Examples of the data stored as a two-dimensional table are shown in FIG.21. In FIG.21, 2101 denotes the occurring number of Chinese and 2102 denotes the occurring number of Japanese, and the table structures are identical to those in the sixth embodiment. The patterns are stored subsequently up to any record number into the storing medium of the seventh embodiment, and as a result the occurring numbers of times for respective pattern types can be stored into the table. Like the sixth embodiment, a plurality of pattern (language) types can be stored into the table by the similar method every pattern type.

Next, a method of discriminating the patterns by using the test pattern will be explained hereunder. The test pattern in FIG.17 is employed. In this case, the data to be extracted in compliance with the character separating rule are extracted like the two-character chains 1703 "事件" and 1706 "我々". Then, the occurring numbers of times corresponding to two-character chains 1703, 1706 are detected from the table in the storing medium in FIG.21. In the Chinese table (2101 in FIG.21), the occurring number of times of the pattern (同,事) is 1 and does not take other numerical values.

In contrast, in the Japanese table (2102 in FIG.21), the occurring number of times of the pattern (我,々) is 2 and the occurring number of times of the pattern (事,件) is 2. When a total of these occurring numbers of times is calculated every pattern type, such total is 1 in Chinese and 4 in Japanese. As a result, it can be decided that the test pattern corresponds to Japanese.

Although the total of the occurring number of times is calculated in the seventh embodiment, there may be adopted a method which either calculates the pattern types whose total of the occurring numbers of times exceeds n or calculates a total of the occurring numbers of times of the character chains in which the occurring numbers of times of each character chain exceeds n.

Although the separating rule is applied as the method of extracting the character chain corresponding to the test pattern,

all character chains of the neighboring two characters may be picked up like the sixth embodiment without the application of the separating rule.

According to this approach, based on the example which is applied to the actual data cited in the fifth embodiment or the sixth embodiment, the pattern can be decided by setting the character number, which mates with the occurring frequency %, to n as the criterion of the occurring numbers of times. For example, in the case of Chinese, a total sum n of the occurring numbers of times corresponding to the kanji code can be set to 74 if the occurring frequency in the Chinese kanji codes is 74 % and the occurring character types are 3469 types to 100 characters.

FIG.22 shows a configuration of an information processing system according to the seventh method embodiment of the present invention. In FIG.22, 2202 denotes a break discriminator for cutting out the sample data from the text data according to the separating rule; 2201, a character chain extractor for acquiring two-character chain or one character chain and reading the occurring number of times and the record number when the result is obtained from the break discriminator 2202, and extracting all neighboring character chains when no result is obtained from the break discriminator 2202; 2203, a chain converter for forming two-character chain by repeating one character when the character chain extractor 2201 extracts the one character; 2204, a character chain frequency counter for storing the two-character chains, the

record numbers, and the occurring number of times into a frequency memory 2205 every pattern (language) type; 2206, a frequency counter for fetching the occurring number of times of the character chains, which are obtained by the character chain extractor 2201 or the chain converter 2203, from the frequency memory 2205 every pattern type; 2207, a collator for calculating a total of the occurring number of times detected from the frequency counter 2206 or a total of the occurring number of times of a certain threshold value every pattern type and then deciding the pattern type in which the test pattern is contained by calculating a pattern exceeding the certain threshold value every pattern type or comparing magnitudes of them every pattern type.

Therefore, according to this method, the pattern to which the test data belongs can be automatically decided. Also, there is such an advantage that, unless character types, fonts, etc. are coded, the pattern can be classified. In addition, the pattern can be classified without the preparation of all rules for respective languages.

(Embodiment 8)

An example of a method of discriminating the character pattern will be explained as an eighth embodiment of the present invention. First, a method of forming information to discriminate the pattern is explained hereunder. Assume that classified patterns are set to 2 (Japanese and Chinese, etc.). FIG.23 shows a configuration of a code table which stores coded character

patterns. An A group 2301 is a group which consists of so-called hiragana and katakana such as characters "と",..., etc. Code numbers are assigned to respective characters of the A group 2301 such that #10 is assigned to "と". Similarly, a B group 2302 is a group which consists of kanji and code numbers are assigned respective kanjis such that #100 is assigned to "我", #101 is assigned to "同",... In this case, above code numbers are assigned for the sake of convenience, and any code numbers may be employed if they are unique in the code table.

While taking the same data as the sample patterns indicated by 1401, 1406, 1412 in FIG.14 as the sample data which are used to form the information to discriminate the patterns, the eighth embodiment will be explained. In this case, in the eighth embodiment, assume that the pattern classifications (languages) to which the sample patterns 1401, 1406, 1412 in FIG.14 belong are not known. FIG.24 shows examples of the character chain pattern extracted from the sample table in FIG.14. In answer to the character chains, for example, 1404 "我同", 1405 "同事" in FIG.14, the document number 1, the character code 101 of the first character, the character code 102 of the second character, and the occurring number of times 1 of the character chain pattern 2401 "我同" are stored. Similarly, 2402 denotes the character chain pattern of the character chain "同事". Then,

there are character chain patterns 2403, 2404, 2405, etc. in FIG.24 as the character chain pattern 1406 in FIG.14. Similarly, there are character chain patterns 2406, 2407, etc. in FIG.24 as the character chain pattern 1412 in FIG.14.

Next, character chain patterns are put together every character chain pattern. Here, the patterns in which the same character chain pattern is contained are collected on the basis of the document number, and the number of such patterns is calculated. FIG.25 shows examples in which the character chain patterned are collected respectively. As indicated by 2501 in FIG.25, in the document number 1, "同事" are common to other documents, the associated document number is 1, and the number of the common character chain patterns is 2 (where the number 1 indicates that there is no common character chain pattern).

While, as indicated by 2502 in FIG.25, "我々", "事件" are common to other documents, and the number of the common character chain patterns is 4. Similarly, in the document number 3, the number of the common character chain patterns is 4. According to this, it is decided that the document numbers 2 and 3 are different from the document number 1, and then the document number 1 is classified into the pattern 2 and the document numbers 2, 3 are classified into the pattern 1. Then, the character chain patterns are stored. FIG.26 shows examples in which the documents are stored in the storing medium after they are classified into the patterns 1, 2. Subsequently, two- character chains are formed

from the sample data, then the character chain patterns are picked up from two patterns (pattern 1, pattern 2) stored in the storing medium, then character chains of the sample data are compared with the character chain patterns stored in the storing medium every pattern type, and then the character chain patterns of the pattern type having the larger number in the sample data are stored into the storing medium.

The character chain patterns of the pattern type having the larger number are stored in the storing medium in the eighth embodiment, but the pattern type may be decided by fetching the character chain patterns which are larger than a certain threshold value or comparing the case number when types of the fetched character chain patterns are larger than a certain threshold value.

According to this approach, when the actual data are analyzed, distributions consisting of the occurring number of times, the character codes, and their combinations are different in Chinese/ Hangul/ Japanese, like the examples cited in the embodiments 5, 6, 7. Particularly, as explained in the embodiment 6, the frequency is concentrated onto few Hangul characters in Hangul, while the frequency is concentrated onto the combination of hiragana/katakana in Japanese, and the frequency is distributed uniformly over a variety of kanji characters in Chinese. Based on the smallest difference between the occurring frequency and the occurring character type in three language codes,

it can be decided which distribution the language is close to.

Next, a method of discriminating the pattern type in which the test pattern is contained by using the test pattern will be explained hereunder. This discrimination method can be achieved by the method similar to those in the fourth embodiment to the seventh embodiment.

FIG.27 shows a configuration of an information processing system according to the eighth method embodiment of the present invention. In FIG.27, 2701 denotes a character chain extractor for picking up the adjacent two-character chains from the text data; 2702, a character chain frequency counter for reading the occurring number of times and the record number of the character chains; 2703, a frequency counter fetching the occurring number of times of the character chains being detected by the character chain extractor 2701 from a frequency memory 2705 every pattern type, or calculating the frequency of the character chains obtained by the character chain frequency counter 2702 if no occurring number of times exists; and 2704, a pattern sorter for calculating a total of the occurring number of times of the character chain patterns or a total of the occurring number of times of a certain threshold value to detect the pattern type in which the character chains or the character chain patterns obtained from the frequency counter 2703 are contained every pattern type, then comparing their magnitudes every pattern type, then storing the character chain pattern into respective pattern

types of the frequency memory 2705 in the order of larger total, and then informing the test pattern of the pattern type as the decision result.

Therefore, according to this method, the pattern to which the test data belongs can be automatically decided. Also, there is such an advantage that, unless character types, fonts, etc. are coded, the pattern can be classified. In addition, the pattern can be classified without the preparation of the dictionaries for respective languages. Further, the language or the pattern to which the input text belongs can be decided by forming individual language files with referring the rule and reconstructing the classification of the language or the pattern based on their occurring relation.

(Embodiment 9)

Then, the case where the text data used in the integrated code (unicode) or the local code (EUC, etc.) are used commonly will be discussed as an example of the text language discrimination. An embodiment will be explained in accordance with the flow in a block configurational diagram of FIG.32.

First, examples of the character codes when the input texts are managed by the integrated code or the local code are stored in an individual character set code table memory shown in FIG.28 (3203 in FIG.32). In FIG.28, 2801 to 2804 denote tables of the integrated code wherein 2801 denotes a hiragana/katakana code table; 2802, a kanji code table; 2803, a symbol code table;

30 OCT 2002

2804, a Hangul code table. Code numbers are assigned uniquely to respective tables. In FIG.28, 2805 to 2809 denote local code tables wherein 2805 denotes a hiragana/ katakana local code table in Japanese; 2806, a hiragana/katakana local code table in Chinese; and 2807, a hiragana/katakana local code table in Hangul. Their code numbers are different according to the character types. Further, 2808 denotes a kanji local code table, and 2809 denotes a symbol local code table. The input data is the text shown in FIG.30.

The example using the integrated code is explained in the ninth embodiment, and its differences from the case where the local code is used will be explained appropriately in the ninth embodiment.

The code numbers are correlated with respective characters in "日本にあう漢字" in FIG.30. At this time, a code discrimination designator 3201 reads a designated character code system (the integrated code or the local code), and a two-character chain extractor 3202 cuts out two character chains "日本", "本に", "にあ", "あう", "う漢", "漢字" from the input data. Then, an individual language occurring frequency counter 3204 detects the code number corresponding to the text in unicode based on an individual character set code table memory 3203, then checks the code number of the first character of the two-character chain, and then calculates the occurring number of the character type of kanji, hiragana,... For example, in the case of the input text

in FIG.30, the number of two-character chains is 6, the chain having Kanji as the first character is 3, the chain having hiragana/katakana as the first character is 3, and others are 0. Accordingly, the occurring degree of kanji is 50 %, and the occurring degree of hiragana/katakana is 50 %. Then, the individual language occurring frequency counter 3204 sends the occurring degree of two character chain to an individual language interdistribution distance calculator 3206.

Then, character occurring rates % are read from an individual language distribution table memory 3205. The individual language distribution table memory 3205 is a learning table in which rates of the occurring patterns of two-character chains are calculated from the data described in respective languages. The occurring % for each character type such as kanji, hiragana/katakana, symbol, Hangul, etc. in Japanese, Chinese, Hangul, etc. are stored as the table. The individual language interdistribution distance calculator 3206 measures an interdistribution distance between the already extracted data and the above learning table in compliance with the distance calculating rule in FIG.31. For instance, in the case of Japanese standard, the distance is given by (50 %(kanji occurring %)-49 %)2+(50 %(hiragana occurring %)-41 %)2+(0 %(symbol occurring %)-10 %)2 +(0 % (Hangul occurring %)-0 %)2=182. Similarly, in the case of Chinese standard, the distance is given by 1376. In the case of Hangul standard, the distance is given by 9895. These

results are sent to a comparator 3207. The comparator 3207 checks the minimum standard of the Japanese, Chinese, and Hangul standards and then detects the Japanese standard of 182. It is decided based on the above result that the input text is Japanese.

In this case, the distance in FIG.31 is calculated as (occurring % of respective character types-occurring % in the learning table)2. However, any distance may be adopted if such distance can satisfy a triangle inequality (i.e., $d(x,z) \leqq d(x,y)+d(y,z)$ where d is a distance function and x, y are any values). Also, in the ninth embodiment, the pattern is not limited to the character pattern if such pattern is the coded pattern. In addition, the object should not be particularly limited if such object is the coded object such as symbol, figure,.. like ◎, ◇, etc.

As described above, the present invention has such an advantage that the language of the input text described in a certain language can be discriminated by utilizing descriptive features of the language in the country, i.e., by detecting the occurring rate of the particular character of the text as the discriminated object, or detecting the average of the text length between the particular characters, or detecting the occurring rate of the character in the particular range, while observing particular characters which are frequently occurred in the language or particular characters in the particular range.

Second, the present invention has such another advantage

that, while utilizing the search system using two-character chains as the index prepared by the document search, the keyword search can be achieved by adding information of separating characters to the index prepared at the time of registration, as they are, in the input text in which the separating characters are inserted into the beginning and the end of the keyword.

Third, the present invention has such still another advantage that the language to which the given text belongs can be discriminated automatically based on the given text, and that the language or the group type can be discriminated automatically under the condition features of the text are decided as codes, particularly such features are classified by character pattern profiles, e.g., fonts, etc.

WHAT IS CLAIMED IS:

1. An information processing system comprising:

particular character counting means for detecting and counting particular characters having a particular character code of a detected target language from all characters of an input text;

occurring rate calculating means for calculating a particular character occurring rate based on a number of the particular characters detected by the particular character counting means and a number of all characters in the input text;

standard occurring rate storing means for storing a standard occurring rate of the particular characters of the target language in advance; and

a comparator for comparing the particular character occurring rate of the input text derived from the occurring rate calculating means with the standard occurring rate detected from the standard occurring rate storing means;

wherein it is decided whether or not the input text corresponds to a text having features which mate with the target language.

2. An information processing system comprising:

particular character detecting means for detecting and counting one or plural particular characters in a character code of a detected target language from all characters of an input text;

unparticular text detecting means for extracting texts

30 OCT 2002

consisting of unparticular characters between the particular characters derived from the particular character detecting means and detecting a number of the unparticular characters;

average text length calculating means for calculating an average text length of unparticular texts derived from the unparticular text detecting means;

standard text length storing means for storing a standard average text length of the unparticular characters of the target language in advance; and

comparing means for comparing an unparticular character length of the input text derived from the average text length calculating means with the standard average text length derived from the standard text length storing means;

wherein it is decided whether or not the input text corresponds to a text having features which mate with the target language.

3. An information processing system comprising:

character-in-a-particular-range counting means for detecting and counting characters-in-a-particular-range as characters in a character code in a particular range of a detected target language from all characters of an input text;

occurring rate calculating means for calculating a character-in-a-particular-range occurring rate based on a number of the characters-in-a-particular-range detected by the

character-in-a-particular-range counting means and a number of characters in the input text;

standard occurring rate storing means for storing a standard occurring rate of the characters-in-a-particular-range of the target language in advance; and

comparating means for comparing the character-in-a-particular-range occurring rate of the input text derived from the occurring rate calculating means with the standard occurring rate detected from the standard occurring rate storing means;

wherein it is decided whether or not the input text corresponds to a text having features which mate with the target language.

4. An information processing system according to claim 3, wherein said comparating means more than two ranges which specifies the character-in-a-particular-range to compare occurring rates of more than two characters-in-a-particular-range.

5. An information processing system comprising:

first separating character inserting means for inserting separating characters as special control character codes which do not exist in an input text into boundaries between keywords as a text by which meaning of the input text is separated when the text is registered;

detecting means for detecting three-character chains in which the separating character is put in a middle if the separating characters are found in a separating character inserted text;

character chain generating means for generating a character chain consisting of characters located before and after the separating character, a character chain consisting of the separating character and a character located before the separating character, and a character chain consisting of the separating character and a character located after the separating character; and

second separating character inserting means for inserting the separating character into a beginning of the input text or an end of the input text or the beginning and the end of the input text upon searching.

6. An information processing system comprising:

separating character inserting means for inserting separating characters into boundaries between keywords when a text is registered;

first character chain generating means for generating a character chain consisting of the separating character and a character located before the separating character, a character chain consisting of the separating character and a character located after the separating character, and a character chain consisting of characters located before and after the separating

character if characters of the input text generated by the separating character inserting means are particular separating characters;

second character chain generating means for generating two- character chain consisting of a first character and a subsequent character if input characters are not the separating characters;

character chain storing means for storing occurring numbers of times of two-character chains and respective characters in the two-character chains;

searched text generating means for inserting particular separating characters into a beginning and an end of a collated text;

third character chain generating means for generating two- character chain consisting of a first character and a subsequent character of a searched text generated by the searched text generating means;

detecting means for detecting a same character chain as the character chain generated by the third character chain generating means from the character chain storing means and detecting its occurring number of times; and

comparing means for deciding whether or not the occurring number of times of a second character of a former chain of two successive chains, in which the second character of the former chain corresponds to a first character of a latter chain, is equal

to the occurring number of times of the first character of the latter chain;

wherein it is decided by the comparing means upon collation over all chains extracted by the third character chain generating means whether or not the occurring number of times of the second character of the former chain is equal to the occurring number of times of the first character of the latter chain, and thus keyword search is carried out by detecting the searched text and boundaries formed by inserting the separating characters in the searched text.

7. An information processing system comprising:

separating character inserting means for inserting separating characters into boundaries between keywords when a text is registered;

first character chain generating means for generating a character chain consisting of the separating character and a character located before the separating character, a character chain consisting of the separating character and a character located after the separating character, and a character chain consisting of characters located before and after the separating character if characters of the input text generated by the separating character inserting means are particular separating characters;

second character chain generating means for generating

two- character chain consisting of a first character and a subsequent character if input characters are not the separating characters;

character chain storing means for storing two-character chains and occurring positions of respective characters in the two- character chains;

searched text generating means for inserting particular separating characters into a beginning and an end of a collated text;

third character chain generating means for generating two- character chain consisting of a first character and a subsequent character of a searched text generated by the searched text generating means;

detecting means for detecting a same character chain as the character chain generated by the third character chain generating means from the character chain storing means and detecting its occurring number of times; and

comparing means for deciding whether or not the occurring position of a second character of a former chain of two successive chains, in which the second character of the former chain corresponds to a first character of a latter chain, is equal to the occurring position of the first character of the latter chain;

wherein it is decided by the comparing means upon collation over all chains extracted by the third character chain generating means whether or not the occurring position of the

second character of the former chain is equal to the occurring position of the first character of the latter chain, and thus keyword search is carried out by detecting the searched text and boundaries formed by inserting the separating characters in the searched text.

8. An information processing system according to claim 6 or 7, wherein a second searched text generating means for inserting particular separating characters into a beginning of a collated text is provided in place of said searched text generating means, and thus front coincidence search of the searched text is carried out by detecting a boundary of a head character of the searched text.

9. An information processing system according to claim 6 or 7, wherein a third searched text generating means for inserting particular separating characters into an end of a collated text is provided in place of the searched text generating means, and thus rear coincidence search of the searched text is carried out by detecting a boundary of a end character of the searched text.

10. A computer-readable storing medium employed in an information processing system which pattern-identifies uniquely a language, in which text data are contained, out of several

predetermined language categories, for recording programs which constitute pattern groups consisting of combinations of character code groups for character codes in which coded characters employed for pattern identification are classified into a plurality of one-dimensional code groups as code tables,

the program comprising the steps of pattern-storing,

a first data consisting of an n-character combination, and composed of plural character groups in any of which codes constituting combinations are contained,

a second data having numerical values corresponding to combination patterns of the first data, and

a third data assigned to predetermined pattern classification from combinations of the first data and the second data.

11. A storing medium according to claim 10, wherein the first data consist of a two-character combination, and character groups in which codes constituting the combination are contained in any one of the plurality of character groups are combined mutually.

12. An information processing system comprising:

a storing medium in which text data are contained, out of several predetermined language categories, for recording programs which constitute pattern groups consisting of

combinations of character code groups for character codes in which coded characters employed for pattern identification are classified into a plurality of one-dimensional code groups as code tables, the program comprising the steps of pattern-storing, a first data consisting of an n-character combination, and composed of plural character groups in any of which codes constituting combinations are contained, a second data having numerical values corresponding to combination patterns of the first data, and a third data assigned to predetermined pattern classification from combinations of the first data and the second data;

character chain extracting means for detecting two-character chains from a text, and deciding a character group in which characters constituting two-character chain are contained;

pattern calculating means for comparing whether or not a combination of the character groups decided by the character chain extracting means coincides with the pattern in a pattern storing means, and calculating a numerical value of a coincident pattern; and

pattern deciding means for deciding a pattern having maximum numbers of values of all character chains calculated by the pattern calculating means and their combination patterns, which coincide with those in the first data, the second data, or the third data stored in the pattern storing means, as the pattern combination of a pattern storing means.

13. An information processing system according to claim 12, wherein the pattern deciding means calculates an average value of numerical values calculated by the pattern calculating means for all character chains, and decides the value which is closest to the second data of the pattern storing means as the pattern combination of the pattern storing means.

14. An information processing system according to claim 12, wherein the pattern deciding means calculates a most frequent value of numerical values calculated by the pattern calculating means for all character chains, and decides the value which is closest to the second data of the pattern storing means as the pattern combination of the pattern storing means.

15. A computer-readable storing medium employed in an information processing system which pattern-identifies uniquely a language, in which text data are contained, out of several predetermined language categories, for recording a program which constitutes pattern groups consisting of combinations of pattern groups for pattern identification if character codes are overlapped respectively in the pattern groups in which patterns are composed of combination of character codes,

the program comprising the steps of pattern-storing a first data consisting of two-character combination patterns for the pattern group in which pattern groups are

previously known,

a second data consisting of pattern types indicating the first data, and

a third data showing occurring numbers of times of combinations of the first data and the second data.

16. An information processing system comprising:

a character chain extracting means for extracting all two-character chain of the pattern data recorded in a computer-readable storing medium, in which text data are contained, out of several predetermined language categories, for recording a program which constitutes pattern groups consisting of combinations of pattern groups for pattern identification if character codes are overlapped respectively in the pattern groups in which patterns are composed of combination of character codes, the program comprising the steps of pattern-storing a first data consisting of two-character combination patterns for the pattern group in which pattern groups are previously known, a second data consisting of pattern types indicating the first data, and a third data showing occurring numbers of times of combinations of the first data and the second data;

a character chain frequency counting means for calculating occurring number of times of respective character chains and storing character chain patterns and the occurring number of times in the storing medium every pattern type;

the character chain extracting means for extracting two-character chains from a given pattern;

a frequency counting means for extracting occurring number of times of the character chain pattern corresponding to the extracted character chain from the storing medium every pattern type; and

a collating means for collating the occurring number of times derived from the frequency counting means with all extracted character chains, and then deciding the pattern type which has a largest total occurring number of times as the pattern type in which the given pattern is contained.

17. An information processing system according to claim 16, wherein the collating means decides the pattern type whose total occurring number of times exceeds a certain threshold value as the pattern type in which the given pattern is contained.

18. An information processing system according to claim 17, wherein the collating means decides the pattern type which has a largest total occurring frequency of the character chain pattern whose occurring number of times exceeds a certain threshold value over all character chains as the pattern type in which the given pattern is contained.

19. An information processing system according to

claim 10, 11, or 15, further comprising:

a separating means for separating character types or at least one language into words when all character chains are extracted from pattern data;

a character chain extracting means for extracting character chains based on the separating means;

a chain converting means for repeating a same character to form character chains if the character chain consists of one character;

a character chain frequency counting means for calculating occurring number of times of respective character chains and then storing character chain patterns and the occurring number of times in the storing medium every pattern type;

the character chain extracting means for extracting two-character chains from a given pattern;

a frequency counting means for extracting occurring number of times of the character chain pattern corresponding to the extracted character chain from the storing medium every pattern type; and

a collating means for collating the occurring number of times derived from the frequency counting means with all extracted character chains, and then deciding the pattern type which has a largest total occurring number of times as the pattern type in which the given pattern is contained.

20. A computer-readable storing medium employed in an information processing system which pattern-identifies uniquely a language, in which text data are contained, out of several predetermined language categories, for recording programs which constitute pattern groups consisting of combinations of character code groups for character codes in which coded characters employed for pattern identification,

the program comprising the step of storing document numbers of character chain patterns consisting of two-character combination, and character chain patterns of a text data every pattern classification or language.

21. An information processing system comprising:

a character chain extracting means for extracting all two-character chains of a text data recorded in a computer-readable storing medium, in which text data are contained, out of several predetermined language categories, for recording programs which constitute pattern groups consisting of combinations of character code groups for character codes in which coded characters employed for pattern identification, the program comprising the step of storing document numbers of character chain patterns consisting of two-character combination, and character chain patterns of a text data every pattern classification or language;

a character chain frequency counting means for

calculating occurring number of times of respective character chains and calculating the character chain patterns and document numbers every pattern type;

a character frequency comparating means for fetching frequencies of character chain patterns obtained by the character chain frequency counting means and character chain patterns in a storing medium every pattern classification, comparing total values of the frequencies of the character chain patterns for each pattern classification, and storing a character chain pattern, an occurring number of times, and a document number into the pattern classification of the storing medium having a larger pattern classification;

the character chain extracting means for extracting two-character chains from a given pattern;

a frequency counting means for extracting the occurring number of times of the character chain pattern corresponding to the extracted character chain from the storing medium every pattern type; and

a collating means for collating the occurring number of times derived from the frequency counting means with all extracted character chains, and then deciding the pattern type which has a largest total occurring number of times as the pattern type in which the given pattern is contained.

22. An information processing system according to

claim 21, wherein the character frequency comparating means set forth in claim 21 decides the pattern type whose total occurring number of times exceeds a certain threshold value as the pattern type in which the given pattern is contained.

23. An information processing system according to claim 22, wherein character frequency comparating means set forth in claim 22 decides the pattern type which has a largest total occurring frequency of the character chain pattern whose occurring number of times exceeds a certain threshold value over all character chains as the pattern type in which the given pattern is contained.

24. A computer-readable storing medium employed in an information processing system which pattern-identifies uniquely a language, in which text data are contained, out of several predetermined language categories in which character codes are defined every character font, for recording programs which constitute pattern groups consisting of combinations of character code groups which are employed for pattern identification,

the program comprising the step of storing two-character combinations constituting characters every language in an individual language distribution table in which an occurring rate of a first character are stored per character type of kanji, hiragana/katakana, symbol, Hangul, and other characters.

25. An information processing system comprising:

an individual code designating means for reading an integrated code or a local code as an employed character code system;

a character chain extracting means for extracting all two-character chain from input text data;

an occurring number-of-times counting means for classifying two-character chains of kanji, hiragana/katakana, symbol, Hangul, and others in compliance with a designated code system every language, and then calculating their occurring rates respectively;

an individual language interdistribution distance calculating means for a total of distances between the occurring rates of kanji, hiragana/katakana, symbol, Hangul every language according to designated code system and the occurring rates of kanji, hiragana/ katakana, symbol, Hangul every language managed by the storing means; and

a comparing means for causing the individual language interdistribution distance calculating means to calculate the smallest distance value, and then deciding a language category having the smallest distance value as a language of the input text.