

(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2015年2月26日(26.02.2015)



(10) 国際公開番号
WO 2015/025751 A1

- (51) 国際特許分類:
G06F 17/30 (2006.01)
- (21) 国際出願番号: PCT/JP2014/071136
- (22) 国際出願日: 2014年8月5日(05.08.2014)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:
特願 2013-173134 2013年8月23日(23.08.2013) JP
- (71) 出願人: 日本電気株式会社(NEC CORPORATION) [JP/JP]; 〒1088001 東京都港区芝五丁目7番1号 Tokyo (JP).
- (72) 発明者: 楠村 幸貴(KUSUMURA, Yukitaka); 〒1088001 東京都港区芝五丁目7番1号日本電気株式会社内 Tokyo (JP). 村岡 優輔(MURAOKA, Yusuke); 〒1088001 東京都港区芝五丁目7番1号日本電気株式会社内 Tokyo (JP).
- (74) 代理人: 池田 憲保, 外(IKEDA, Noriyasu et al.); 〒1000011 東京都千代田区内幸町1丁目2番2号 日比谷ダイビル Tokyo (JP).
- (81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.
- (84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).
- 添付公開書類:
— 国際調査報告 (条約第21条(3))

(54) Title: FREQUENT SEQUENCE ENUMERATION DEVICE, METHOD, AND RECORDING MEDIUM

(54) 発明の名称: 頻出系列の列挙装置、方法および記録媒体

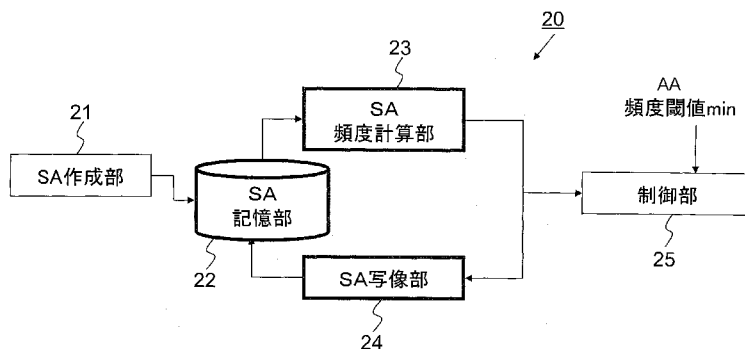


図2

21... SA CREATION UNIT
22... SA STORAGE UNIT
23... SA FREQUENCY CALCULATION UNIT
24... SA MAPPING UNIT
25... CONTROL UNIT
AA... FREQUENCY THRESHOLD (min)

(57) Abstract: In order to efficiently carry out a frequency calculation process, a frequent sequence enumeration device comprises: a suffix array storage unit which, when all suffixes for sequence data are arranged in dictionary order, stores a suffix array whereby it is possible to reference a location on the sequence data from a location on a dictionary; a suffix array frequency calculation unit which, taking a set of locations on the suffix array as input, counts the number of times a character is included in the suffixes referenced by the suffix array, using the fact that the suffixes are sequenced; and a suffix array mapping unit which, taking a set of locations in the suffix array and a specified character as input, returns an appearance location on the suffix array with respect to the inputted character by adding one to the value within the suffix array and referencing an inverted array of the suffix array.

(57) 要約: 頻度計算処理を効率よく行うために、頻出系列の列挙装置は、系列データについての全接尾辞を辞書順で並べた場合に、辞書上の位置

から系列データ上の位置を参照可能にする接尾辞配列を記憶する接尾辞配列記憶部と、接尾辞配列上の位置の集合を入力として、接尾辞配列によって参照される接尾辞に含まれる文字の数を、接尾辞が順序付けされていることを利用して数える接尾辞配列頻度計算部と、接尾辞配列内の位置の集合と、特定の文字を入力として、接尾辞配列内の値に1を加算し接尾辞配列の逆配列を参照することで、入力された文字に対する接尾辞配列上の登場位置を返す接尾辞配列写像部と、を備える。

明 細 書

発明の名称

頻出系列の列挙装置、方法および記録媒体

5

技術分野

本発明は、テキストデータ、購買データ、DNA (deoxyribonucleic acid) データなどのデータ列内から高頻度な系列を列挙する装置に関する。

10 背景技術

大規模なデータベースから利用者にとって有用なパターンを発見する手法として、頻出系列の列挙装置が知られている。ここで、「頻出系列の列挙装置」とは、購買データやDNAデータ、テストなどの順序関係に意味を持つデータ(系列データ)により構成されたデータベース(系列データベース)から、高
15 頻度な部分系列を列挙する装置である。「系列データ」とは順序関係の文字列である。なお、ここで「文字」とは、任意の種類数を持つ記号である。

よって例えば、購買データの場合、アイテム(商品)を文字、テキストデータの場合では、1単語を文字としても良い。

「部分系列」とは、系列データ内で順序を保った登場する部分列を指す。

20 例えば、4文字から成る系列データA B C Dの部分系列は、

A B, A C, A D, B C, B D, C D, A B C, A B D, A C D, B C D, A B C D
である。

頻出系列の列挙を実現する方法としては、PrefixSpanを用いた方法が知られている(非特許文献1参照)。PrefixSpanは、頻度計算と、データベース写
25 像と呼ばれる処理とを繰り返すアルゴリズムである。データベースは、D Bと略称される。

図1は、PrefixSpanを用いた列挙装置10の構成を示すブロック図を示す。列挙装置10は、系列D B管理部11と、頻度計算部12と、制御部13と、D B写像部14とを有する。

例えば、系列DB管理部11内の系列DBに{ACD, ABC, CBA, AAB}という4つの系列があり、頻度の閾値として2が指定されたとする。この場合に、この列挙装置10はまず、頻度計算部12を用いて頻度計算を行う。この処理では系列DB内の各文字の頻度が数えられる。この例では、A
5 が5回、Bが3回、Cが3回、Dが1回登場している。このうち系列データ{A, B, C}がそれぞれ頻度2以上であるため、制御部13はA, B, Cの三つの系列を出力する。さらに制御部13、DB写像部14を用いて最初の系列DBに対して、頻度2以上の文字を用いたDB写像を行う。

「文字xを用いたDB写像」とは、系列DBから、文字xを含むレコード
10 を取り出し、文字x以降の文字列(接尾辞)を取り出す処理を言う。

例えば、系列DB{ACD, ABC, CBA, AAB}に対し、文字Aを用いたDB写像を行うと、新たに{CD, BC, AB}という3つの系列からなる系列DBが得られる。

そして、この列挙装置10は、この系列DBに対して頻度計算を行い、A
15 が1回、Bが2回、Cが2回という結果を得る。この結果、制御部13は、ABが2回、ACが2回登場しているとしてABとACを出力する、この処理は深さ優先探索で再帰的に実施され、元の系列DB内で頻度2以上のすべての系列が列挙される。

PrefixSpanは、頻度計算と、DB写像と呼ばれる処理とを繰り返すことで、
20 頻度計算を行う。

非特許文献2は、接尾辞配列について開示している。

また、本発明に関連する先行技術文献も知られている。例えば、特許文献1は、射影にかかる処理の効率化を図ることが可能な「系列パターン抽出装置及び方法」を開示している。この特許文献1に開示された系列パターン抽出装
25 置では、系列データを構成するアイテムに対し、射影時において抽出の対象となるアイテムを特定するための射影位置を関連付けて記憶しておくことで、射影する位置を効率的に取得し、射影にかかる処理を効率的に行っている。

先行技術文献

特許文献

特許文献 1 : 特開 2009-169850 号公報

非特許文献

- 非特許文献 1 : J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen,
5 U. Dayal, and M.-C. Hsu, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth", Proc. of The 17th Int'l Conf. on Data Engineering, pp. 215-224 (2001)

- 非特許文献 2 : Klaus-Bernd Schurmann and Jens Stoye (2005). "An inc
omplex algorithm for fast suffix array construction". In Proceedings
10 of the 7th Workshop on Algorithm Engineering and Experiments and the 2nd Workshop on Analytic Algorithmics and Combinatorics (ALENEX/ANALCO 2005), 2005, pp. 77-85.

発明の概要

- 15 発明が解決しようとする課題

しかしながら、上記非特許文献 1 に開示された列挙装置 10 の第 1 の問題点は、処理時間が長いことである。従来の方法では、DB 写像後に繰り返し頻度計算処理が実施されるが、系列データの数が大きくなるとこれらの処理コストが大きくなってしまう。

- 20 なお、上記非特許文献 2 は接尾辞配列を開示しているに過ぎない。

一方、特許文献 1 は、射影時において抽出の対象となるアイテムを特定するための射影位置を関連付けて記憶する技術を開示しているに過ぎない。

[発明の目的]

- 本発明の目的は、頻度計算処理を効率よく行う高速な頻出系列の列挙装置
25 を実現することである。

課題を解決するための手段

本発明による頻出系列の列挙装置は、系列データについての全接尾辞ないし全接頭辞を辞書順で並べた場合に、辞書上の位置から元系列データ上の位置を参照可能にするソート済み配列を記憶するソート済み配列記憶部と；ソ

- ート済み配列とソート済み配列上の位置の集合を入力として、元系列データ上に存在する文字の数を、ソート済み配列が順序付けされていることを利用して数えるソート済み配列頻度計算部と；ソート済み配列内の位置の集合と、特定の文字を入力として、ソート済み配列内の値に対する加減算を行い、入力された文字に隣接する接尾辞ないし接頭辞の登場位置を参照し、対象とする位置の集合を絞り込むソート済み配列写像部と；を備える。

発明の効果

本発明では、頻度計算処理を効率よく行うことができる。

10 図面の簡単な説明

図1は非特許文献1に開示された関連の列挙装置の構造を示すブロック図である。

図2は本発明の第1の実施形態に係る、頻出系列の列挙装置の構成を示すブロック図である。

- 15 図3は接尾辞配列を説明する図である。

図4は接尾辞配列の例を示す図である。

図5は図2に示した頻出系列の列挙装置の処理フローを示す図である。

図6は図5における頻度計算処理f2の処理フローを示す図である。

図7は図6における2分検索処理c03の処理フローを示す図である。

- 20 図8は図5におけるSA写像処理f6のフローを示す図である。

図9は図8における接尾辞の拡張の処理フローを示す図である。

図10は本発明の第2の実施形態に係る、頻出系列の列挙装置の構成を示すブロック図である。

図11は接尾辞配列の別の例を示す図である。

- 25 図12は頻度算出部の出力の例を示す図である。

図13は対象系列AのSA写像により得られる対象接尾辞リストを擬似的に示す図である。

図14は頻度算出部の出力の別の例を示す図である。

図15は対象系列AAのSA写像により得られる対象接尾辞リストを擬似

的に示す図である。

図16は対象系列ABのSA写像により得られる対象接尾辞リストを擬似的に示す図である。

5 発明を実施するための形態

本発明の理解を容易にするために、接尾辞配列について説明する。「接尾辞配列」とは、上記非特許文献2で示されるデータ構造である。

接尾辞配列は、文字列中に登場するすべての接尾辞を辞書順で並べた場合の位置を保持した配列である。

- 10 図3に文字列a b r a c aの接尾辞配列の例を示す。図3の一番左のリストは文字列a b r a c aに対する全接尾辞を表す。なお、この図では便宜的に、末尾に辞書順で先頭に来る文字\$を付与している。ここで各接尾辞に S_i という記号を割り当てるものとする。 S_i は、i番目の文字以降の接尾辞を意味する記号である。図3の中央にあるリストは左のリストを辞書順にならべ
- 15 変えたものである。さらに図3の右にある整数値のリストは、中央の接尾辞リストの添え字を取り出したものであり、これが接尾辞配列となる。

- 接尾辞配列は、辞書順で任意の位置にある接尾辞が元文字列上のどこに存在するかを表したものである。接尾辞配列は、主に文書検索に用いられるデータ構造であり、任意の部分文字列を $\log(\text{文字列})$ で取り出せることが知られて
- 20 ている。以下では、接尾辞配列を、単に、SAと略称する場合がある。

また、接尾辞配列についての逆配列を保持することで、接尾辞配列上から一つ右に存在する接尾辞配列を取り出すことができる。なお、配列Aの逆配列Vとは

$$A[V[i]] = i$$

- 25 の関係が成立する配列である。

接尾辞配列を $SA[i]$ 、その逆配列を $INV[i]$ とする。

このとき、接尾辞配列 $SA[i]$ は、辞書順でi番目にある接尾辞が元文字列上で何文字目に登場するかを表す。

その逆配列 $INV[i]$ は、元文字列上でi番目に登場する接尾辞が、辞書

順で何番目に登場するかを表す。

この2つの配列を用いると、接尾辞配列上で i 番目の接尾辞の一つ右の接尾辞は、

$$\text{INV}[\text{SA}[i] + 1]$$

- 5 を計算することで算出できる。

例えば、図3のSA配列において、辞書順で上から三つ目にある $\text{a b r a c a \$}$ に対する $\text{SA}[2]$ (配列の添え字は0始まりであるとする) は1である。この値に1を足した S_2 が登場する位置を逆配列から算出すると、4となり、一つ右の接尾辞 $\text{b r a c a \$}$ の位置を算出できる。

- 10 また、接尾辞配列には、同じ接頭辞を持つ接尾辞の一つ右の接尾辞集合は接尾辞配列上で同じ順序で登場するという性質がある。例えば、文字列 $\text{a b r a c a d a b r a}$ に対する接尾辞配列を図4に示す。

- 15 例えば、この接尾辞集合において、同じ接頭辞 a で始まる接尾辞は1番から5番の接尾辞であり、元文字列上で10番目、7番目、0番目、3番目、5番目である。このリストの一つ右の接尾辞集合は、元文字列上で11番目、8番目、1番目、4番目、6番目の接尾辞の集合である。これらはそれぞれ、辞書順で、0番目、6番目、7番目、8番目、9番目に登場し、5つの接尾辞が、辞書順で同じ順番で並んでいることがわかる。

- 20 接尾辞配列はすべての接尾辞を辞書順で並べ替えたものであり、同じ接頭辞を持つ接尾辞は、その接頭辞の次の文字で比較される。この例では、 a が同じであるため、2文字目以降の順番で比較されるため、この順番は必ず保持される。

以下、本発明の実施形態について、詳細に説明する。

[第1の実施形態]

- 25 図2を参照すると、本発明の第1の実施形態に係る頻出系列の列举装置20は、SA作成部21と、SA記憶部22と、SA頻度計算部23と、SA写像部24と、制御部25と、から成る。

SA作成部21は、系列データ集合を入力とし、系列データの区切りを挿入しつつ、接尾辞配列と、接尾辞配列の逆配列と、元系列データとを生成す

る。

SA記憶部22は、SA作成部21が作成した接尾辞配列と、接尾辞配列の逆配列と、元系列データとを記憶する。

5 SA頻度計算部23は、SA記憶部22内のデータを参照し、データが辞書順に並んでいることを利用して頻度計算を行い、結果を制御部25に入力する。

SA写像部24は、制御部25に指定されたポインタを元に、次に参照すべき接尾辞のポインタを算出する。

10 SA記憶部22は、系列データについての全接尾辞を辞書順で並べた場合に、辞書上の位置から系列データ上の位置を参照可能にする接尾辞配列を記憶する。

SA頻度計算部23は、接尾辞配列上の位置の集合を入力として、接尾辞配列によって参照される接尾辞に含まれる文字の数を、接尾辞が順序付けされていることを利用して数える。

15 SA写像部24は、接尾辞配列内の位置の集合と、特定の文字を入力として、接尾辞配列内の値に1を加算し接尾辞配列の逆配列を参照することで、入力された文字に対する接尾辞配列上の登場位置を返す。

また、SA頻度計算部23は、接尾辞配列と接尾辞配列上の位置の集合を入力として、元系列データ上に存在する文字の数を、接尾辞配列が順序付けされていることを利用して数える際に、頻度によって2分探索を行うか否かを切り変えて用いる。

次に、図5参照して、本第1の実施形態の頻出系列の列举装置10の処理フローについて説明する。

25 本処理フローは、SA作成部21に系列データの集合が入力されることで処理を開始する。

SA作成部21はまず、系列データの集合を元に接尾辞配列を構築する(ステップf1)。

この処理は次の手順で行われる。

SA作成部21はまず、系列データに区切り文字を挿入しながら連結する

ことで一つの文字列を作成する。

例えば、

{CAABC, ABCB, CABC, ABBCA}

という4つ系列データに対し、区切り文字を「%」末尾文字を「\$」として

5 作成した文字列は

「CAABC%ABCB%CABC%ABBCA%\$」

となる。

次に、SA作成部21は、作成した文字列を元に接尾辞配列SAを構築する。そして、SA作成部21は、接尾辞配列SAを元に逆配列INVを構築

10 する。

SA作成部21は、結合して作成した文字列Tと、接尾辞配列SAと、逆配列INVの三つをSA記憶部21に格納する。

SA記憶部22に三つの情報が記憶されると、SA頻度計算部23はSAを参照して頻度計算処理を行う(ステップf2)。

15 この処理を図6に示す。

頻度計算処理は、接尾辞配列SA、文字列T、対象接尾辞リストP、対象系列TSを入力として処理される。このうち、対象接尾辞リストPは接尾辞配列SAの添え字を表すリストである。対象系列TSはその時点でSA写像している元となった系列を表す。

20 初回の実行時は、写像の元となった元が存在せず、すべての接尾辞を対象として実施されるため、それぞれ、0～全文字列長のidリストと、空文字列" "が入力となる。

初回以外はSA写像部24がSA記憶部22内に保存した値を設定するものとする。

25 この処理が開始されると、SA頻度計算部23はまず、P上のポインタを表す変数sを0として初期化する(初期化処理c01)。

次に、SA頻度計算部25は、Pのs番目の添え字に対応する文字をT[SA[s]]により取り出し、cとする(文字取り出し処理c02)。

そして、SA頻度計算部23は、P上で先頭文字がcでなくなる位置を2

分探索で取り出し、 e とする（2分探索処理c 0 3）。

2分探索処理c 0 3については、図7の処理によって実施される。

2分探索処理c 0 3は、文字 c と対象接尾辞リスト P を入力とし、 P 内の接尾辞において、先頭文字が c となる最後の位置を2分探索アルゴリズムで

5 実施する。

この処理ではまず2分探索の探索範囲を0から P の長さで開始するため、SA頻度計算部2 3は、範囲を表すポインタ l と u をそれぞれ0と P の長さにする（ステップb 0 1）。

次に、SA頻度計算部2 3は、その範囲の中間にある位置 m を計算し（ステップb 0 2）、その位置にある文字を $T[SA[P[m]]]$ によって取り出し、 t とする（ステップb 0 3）。

そして、SA頻度計算部2 3は、 c と t の比較処理 $comp(c, t)$ を実施する（ステップb 0 4）。

比較処理 $comp(a, b)$ とは、文字 a と文字 b を辞書順で比較し、文字 a と文字 b が一致する場合に0、辞書順で文字 $a < 文字b$ であれば1、辞書順で文字 $a > 文字b$ であれば-1を返す処理である。

$comp(c, t)$ の結果が0以上である場合、先頭文字が c でなくなる位置は探索範囲の後ろに存在するため、SA頻度計算部2 3は、探索範囲の後ろ半分にするために開始位置 l に m を代入する（ステップb 0 5）。

20 そうでない場合、SA頻度計算部2 3は、探索範囲の前半分にするために終了位置 u に m を代入する（ステップb 0 6）。

次に、SA頻度計算部2 3は、終了判定（ステップb 0 7）を行う。終了判定では探索範囲が1つの行に絞られているかどうかを判定するために、SA頻度計算部2 3は、 $l + 1$ が u と等しいかどうかを確認する。そうであれば、SA頻度計算部2 3は、 l を文字 c が登場する最後の位置として出力し、25 そうでない場合、SA頻度計算部2 3は、ステップb 0 2に戻り、同様の処理を行う。

図6に戻って、この結果、接尾辞上で文字 c が並ぶ位置は s から e と判定できるため、SA頻度計算部2 3は、文字 c の頻度を $(e - s)$ として算出

し、文字 c と頻度 $(e - s)$ 、その開始位置 s 、終了位置 e を出力する（出力処理 $c04$ ）。

なお、このとき文字 c が % ないし \$ である場合は出力しないものとする。

さらに SA 頻度計算部 23 は次の文字の頻度計算を行うため、 s を e とする（処理 $c05$ ）。

最後に SA 頻度計算部 23 はポインタ s が P の最後に到達したか否かを判定する（判定処理 $c06$ ）。

つまり $s = P.length$ であれば、SA 頻度計算部 23 は最後に到達したと判定し処理を終了する。

10 そうでなければ、SA 頻度計算部 23 は処理 $c02$ に戻り、同様の処理を行う。

なお、頻度計算の処理は通常、対象接尾辞リスト P のオーダ $O(P.length)$ 必要であるが、本アルゴリズムでは 2 分探索を用いているため、文字の種類数を W とした場合に $O(W \cdot \log(P.length))$ の処理時間で実施することができ
15 ける。

しかしながら、文字の種類数 W に比べて対象接尾辞リスト P の数が小さい場合、各文字を上から順に調べ頻度算出した方が高速に動作することがある。

このため、SA 頻度計算部 23 は、図 5 の頻度計算処理 $f2$ においては、 $P.length$ に対して予め閾値を設け、 $P.length$ がある程度小さい場合に、単純に頻度計算を行っても良い。
20

図 5 に戻って、SA 頻度計算部 23 は、この結果得られた文字、頻度、開始位置、終了位置のリスト、対象接尾辞リスト P 、対象系列 TS を制御部 25 に入力する。

次に、制御部 25 は、各文字の頻度のリストを元に、あらかじめ設定されている頻度の閾値以上の文字が含まれるかを判定する出力判定処理 $f3$ を行う。
25

頻度の閾値以上の文字が含まれる場合、制御部 25 は、対象系列 TS にその文字を文字列結合したものを出力し、対象系列 TS にその文字をスタックに追加する（処理 $f4$ ）。

そして、制御部 25 は、スタックの一番上の系列を取り出す (処理 f 5)。

このとき、もしスタックが空である場合、処理を終えるものとする。

スタックから系列が得られた場合、制御部 25 は、対象接尾辞リストとその系列を SA 写像部 24 に入力する。

- 5 SA 写像部 24 に系列 S、対象接尾辞リストが入力されると、SA 写像部 24 は DB 写像処理を行う (処理 f 6)。

SA 写像部 24 は、系列 S と対象接尾辞リストを入力とし、SA 記憶部 22 内の接尾辞配列内のデータを参照して、系列 S の最後の文字 c 以降に登場する接尾辞に対する接尾辞配列を構築し、対象接尾辞リスト P を更新する処

- 10 理を行う。

この処理を、図 8 を用いて説明する。

SA 写像部 24 はまず、対象接尾辞リスト P の逆配列 P_INV を作成する (ステップ S 11)。

この処理は、対象接尾辞リストに対して次の式により計算される。

- 15 $P_INV[P[i]] = i$

次に、SA 写像部 24 は、出力対象接尾辞リスト P 内において、系列 S の最後の文字 c を接頭辞とする接尾辞の開始位置と終了位置に含まれる接尾辞に対して、接尾辞配列の拡張を行う (ステップ S 12)。

この処理を、図 9 を用いて説明する。

- 20 この処理は、文字 c の開始位置 s、終了位置 e、対象接尾辞リスト P、対象接尾辞リスト P_INV を用いて実施される。

SA 写像部 24 は、c で始まる接尾辞の添え字各 i ($s \leq i < e$) に対して、それぞれ次の処理を行う。

1. i をポインタ k に代入する (ステップ S 21)。
- 25 2. 一つ右にある文字の全接尾辞上の位置を $n = INV[SA[P[k]] + 1]$ により算出する (ステップ S 22)。
3. その位置にある文字を $T[SA[n]]$ により取り出し、区切り文字 '%' か否かを判定する (ステップ S 23)。
4. もしそうであれば処理を終える (ステップ S 24)。

5. そうでなければ、その位置の対象接尾辞リスト上での位置を、 $P_INV[S A[P[k]] + 1]$ により算出し（ステップS 2 5）、その位置を出力した上でステップS 2 2に戻る。

以上の処理により、現在の対象接尾辞リスト内において文字cで始まる文字列の次の位置をすべて取り出すことができる。

図8に戻って、SA写像部24は、この処理によって得られた位置を新たに対象接尾辞リストとし、SA記憶部22に保存する。さらに、SA写像部24は系列Sを対象系列として保存する（ステップS 1 3）。

図5に戻って、SA記憶部22内の対象接尾辞リストが更新されると、SA頻度計算部23は新しい対象接尾辞リストPと対象系列TSを元に頻度計算処理を行う（頻度計算処理f 2）。

本頻出系列の列挙装置20は、この処理（f 2からf 6）を制御部25内のスタックが空になるまで繰り返すことで動作する。

尚、頻出系列の列挙装置20の各部は、ハードウェアとソフトウェアの組み合わせを用いて実現すればよい。ハードウェアとソフトウェアとを組み合わせた形態では、RAM（random access memory）に列挙プログラムが展開され、該プログラムに基づいて制御部（CPU）等のハードウェアを動作させることによって、各部を各種手段として実現する。また、該プログラムは、記録媒体に記録されて頒布されても良い。当該記録媒体に記録されたプログラムは、有線、無線、又は記録媒体そのものを介して、メモリに読込まれ、制御部等を動作させる。尚、記録媒体を例示すれば、オプティカルディスクや磁気ディスク、半導体メモリ装置、ハードディスクなどが挙げられる。

上記実施の形態を別の表現で説明すれば、頻出系列の列挙装置20として動作させる情報処理装置（コンピュータ）を、RAMに展開された列挙プログラムに基づき、SA作成部21、SA記憶部22、SA頻度計算部23、SA写像部24、および制御部25として動作させることで実現することが可能である。

次に、本第1の実施形態の頻出系列の列挙装置20の効果について説明する。

本第1の実施形態の頻出系列の列举装置20では、接尾辞配列とその逆配列を用いてDB画像、頻度計算処理を行っている。接尾辞配列を用いると、ある文字より右に存在する接尾辞のリストを効率よく取り出せるため、繰り返しデータをコピーする必要がある。また、接尾辞は必ず辞書順で並んでいるため、この性質を利用して効率よく頻度を計算することができる。

[第2の実施形態]

なお、上記第1の実施形態に係る列举装置20では接尾辞配列とその逆配列を用い、一つ右の文字に移動するために2つの配列を参照し、 $INV[SA[i] + 1]$ として移動したが、同様の考え方で接頭辞配列を用いることもできる。以下では、接頭辞配列を、単にPAと略称する場合がある。

「接頭辞配列」とは、与えられた文字列内の全接頭辞を辞書順に並べ替えたものである。この場合、接頭辞配列PAとその逆配列INVを作成すると、 $INV[PA[i] - 1]$ を参照することで、一つ右の文字の辞書上の位置を知ることが可能である。

図10を参照すると、本発明の第2の実施形態に係る頻出系列の列举装置20Aは、PA作成部21Aと、PA記憶部22Aと、PA頻度計算部23Aと、PA画像部24Aと、制御部25Aと、から成る。

PA作成部21Aは、系列データ集合を入力とし、系列データの区切りを挿入しつつ、接頭辞配列と、接頭辞配列の逆配列と、元系列データとを生成する。

PA記憶部22Aは、PA作成部21Aが作成した接頭辞配列と、接頭辞配列の逆配列と、元系列データとを記憶する。

PA頻度計算部23Aは、PA記憶部22A内のデータを参照し、データが辞書順に並んでいることを利用して頻度計算を行い、結果を制御部25Aに入力する。

PA画像部24Aは、制御部25Aに指定されたポインタを元に、次に参照すべき接頭辞のポインタを算出する。

PA記憶部22Aは、系列データについての全接頭辞を辞書順で並べた場合に、辞書上の位置から系列データ上の位置を参照可能にする接頭辞配列を

記憶する。

PA頻度計算部23Aは、接頭辞配列上の位置の集合を入力として、接頭辞配列によって参照される接頭辞に含まれる文字の数を、接頭辞が順序付けされていることを利用して数える。

- 5 PA写像部24Aは、接頭辞配列内の位置の集合と、特定の文字を入力として、接頭辞配列内の値から1を減算し接頭辞配列の逆配列を参照することで、入力された文字に対する接頭辞配列上の登場位置を返す。

- また、PA頻度計算部23Aは、接頭辞配列と接頭辞配列上の位置の集合を入力として、元系列データ上に存在する文字の数を、接頭辞配列が順序付けされていることを利用して数える際に、頻度によって2分探索を行うか否かを切り変えて用いる。
- 10

- PA作成部21A、PA記憶部22A、PA頻度計算部23A、PA写像部24A、および制御部25Aの動作は、それぞれ、第1の実施形態におけるSA作成部21、SA記憶部22、SA頻度計算部23、SA写像部24、および制御部25と同様であるので、それらの詳細な動作説明については省略する。
- 15

- 尚、頻出系列の列挙装置20Aの各部は、ハードウェアとソフトウェアの組み合わせを用いて実現すればよい。ハードウェアとソフトウェアとを組み合わせた形態では、RAM (random access memory) に列挙プログラムが展開され、該プログラムに基づいて制御部(CPU)等のハードウェアを動作させることによって、各部を各種手段として実現する。また、該プログラムは、記録媒体に記録されて頒布されても良い。当該記録媒体に記録されたプログラムは、有線、無線、又は記録媒体そのものを介して、メモリに読み込まれ、制御部等を動作させる。尚、記録媒体を例示すれば、光学ディスクや磁気ディスク、半導体メモリ装置、ハードディスクなどが挙げられる。
- 20
- 25

上記実施の形態を別の表現で説明すれば、頻出系列の列挙装置20Aとして動作させる情報処理装置(コンピュータ)を、RAMに展開された列挙プログラムに基づき、PA作成部21A、PA記憶部22A、PA頻度計算部23A、PA写像部24A、および制御部25Aとして動作させることで実

現することが可能である。

次に、本第2の実施形態の頻出系列の列挙装置20Aの効果について説明する。

- 本第2の実施形態の頻出系列の列挙装置20Aでは、接頭辞配列とその逆
- 5 配列を用いてDB写像、頻度計算処理を行っている。接頭辞配列を用いると、ある文字より右に存在する接頭辞のリストを効率よく取り出せるため、繰り返しデータをコピーする必要がある。また、接頭辞は必ず辞書順で並んでいるため、この性質を利用して効率よく頻度を計算することができる。

- なお、「接尾辞配列」および「接頭辞配列」は、「ソート済み配列」と総称
- 10 される。

したがって、上記第1の実施形態および上記第2の実施形態を上位概念で表した、頻出系列の列挙装置は、図示はしないが、ソート済み配列作成部と、ソート済み配列記憶部と、ソート済み配列頻度計算部と、ソート済み配列写像部と、制御部と、から成る。

- 15 ソート済み配列記憶部は、系列データについての全接尾辞ないし全接頭辞を辞書順で並べた場合に、辞書上の位置から元系列データ上の位置を参照可能にする。

- ソート済み配列頻度計算部は、ソート済み配列とソート済み配列上の位置の集合を入力として、元系列データ上に存在する文字の数を、ソート済み配
- 20 列が順序付けされていることを利用して数える。

ソート済み配列写像部は、ソート済み配列内の位置の集合と、特定の文字を入力として、ソート済み配列内の値に対する加減算を行い、入力された文字に隣接する接尾辞ないし接頭辞の登場位置を参照し、対象とする位置の集合を絞り込む。

- 25 また、ソート済み配列頻度計算部は、ソート済み配列とソート済み配列上の位置の集合を入力として、元系列データ上に存在する文字の数を、ソート済み配列が順序付けされていることを利用して数える際に、頻度によって2分探索を行うか否かを切り変えて用いる。

実施例1

次に、図2に示した第1の実施形態に係る、頻出系列の列挙装置10の実施例について説明する。

ここでは、頻度の閾値2において、系列データベースとして次の4つ系列が得られたとする。

- 5 CAABCC
- ABCB
- CABC
- ABBCA

このとき、SA作成部21が区切り文字%と末尾文字\$を挿入し、次の文字列

「CAABC%ABCB%CABC%ABBCA%\$」

を作成し、図11の接尾辞配列を作成する(図5のSA作成処理f1)。

なお、図11では分かりやすくするために各接尾辞を示しているが、実施には接尾辞自体は保持されていないものとする。

15 図11の接尾辞配列SAが得られると、SA頻度計算部23は2分探索処理を行い、各文字の頻度計算を行う(図5の頻度計算処理f2)。

この場合、対象接尾辞リストPは0~23、対象系列TSは空文字列""となる。

この結果を図12に示す。図12は文字A、B、Cについてそれぞれ頻度
20 と全SA上での開始位置、終了位置が出力されている。

SA頻度計算部23は、対象接尾辞リストPは0~23、対象系列TSは空文字列""、図12のデータを制御部25に出力する。

次に、制御部25は、図12のデータを調べ、文字A、B、Cを出力する(図5の出力判定処理f3)と共に、スタックに「C」、「B」、「A」を追加する(図
25 5のステップf4)。

制御部25はさらに、この状態でスタックの上位にあるAを取り出し(図5のステップf5)、SA写像部24に対して対象接尾辞リストP(0~23)、系列(A)を入力する。

SA写像部24は、これらを元にSA写像を行う(図5のSA写像処理f

6)。

この処理ではまず、SA写像部24は、Pに対する逆配列を作成する（図8のステップS11）。

なお、現状のPは $P[i] = i$ であるため、 $P_INV[i] = P[i]$ である。

- 5 この上で、SA写像部24は、接尾辞配列の拡張処理を行う（図8のステップS12）。この処理において、SA写像部24は、現状のPにおいて、Aで始まる接尾辞配列のリストを作成する。

従って、SA写像部24は、図11における6～12番目の位置から右方向への拡張を行う。

- 10 例えば、 $i = 6$ に対する次の位置 n は、

$$n = INV[SA[P[6]] + 1] = INV[SA[6] + 1] = INV[22] = 1$$

となる。

$i = 6$ の一つ右の接尾辞は $i = 1$ にあり、 $T[SA[1]] = '%'$ であるため、

- 15 以降の処理は行われない。

次に、 $i = 7$ についての例を説明する。

$i = 7$ に対する次の位置 n は、

$$n = INV[SA[P[7]] + 1] = INV[SA[7] + 1] = INV[3] = 10$$

- 20 となる。

$T[SA[10]] = A$ であり、このとき、SA写像部24は、

$$k = P_INV[10] = 10$$

を出力する。

同様の処理を繰り返すことで、対象系列Aを元とした対象接尾辞配列リス

- 25 トとして

{6, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 23}

がSA記憶部22に出力される。

図13に、これらの行のみを取り出した擬似的な部分接尾辞配列図を示す。

次に、SA頻度計算部23はこれらのデータを元に頻度計算を行う（図5の

頻度計算処理 f 2)。

この結果を図 1 4 に示す。図 1 4 から、対象系列 A に続く文字としては、A が 2 回、B が 6 回、C が 6 回であることがわかる。よって制御手段 2 5 はこれらと対象系列 A を結合し、AA, AB, AC を出力する。

- 5 さらに、制御部 2 5 はスタック [C, B] にこれらを追加し、スタックを [C, B, AC, AB, AA]

とする (図 5 の出力判定処理 f 3)。

そして、制御部 2 5 は一番上にある系列 AA を取り出し、再び同様の処理を行う。

- 10 すなわち、次の SA 写像により、図 1 5 の対象接尾辞リストが得られる。

しかし、この中には B が 1 回、C が 1 回しか含まれないため、制御部 2 5 は何も出力せず、スタックの次の系列 AB を対象とした処理を行う。

この場合、SA 写像の結果は図 1 6 になり、ABB が 3 回、ABC が 4 回という出力を得る。

- 15 本列挙装置 2 0 は、このような再帰的計算を DB 写像ではなく、接尾辞配列を用いて実施する。写像自体の計算時間は従来と同じくデータ長であるが、写像後の頻度計算を、2 分探索を行って実施できるため、効率よく動作させることができる。

- 20 以上、実施例を参照して本発明を説明したが、本発明は上記実施例に限定されるものではない。本発明の構成や詳細には、本発明のスコップ内で当業者が理解し得る様々な変更をすることができる。

上記の実施形態の一部又は全部は、以下の付記のようにも記載されうるが、以下には限られない。

- 25 (付記 1) 系列データについての全接尾辞ないし全接頭辞を辞書順で並べた場合に、辞書上の位置から元系列データ上の位置を参照可能にするソート済み配列を記憶するソート済み配列記憶部と、

前記ソート済み配列と前記ソート済み配列上の位置の集合を入力として、前記元系列データ上に存在する文字の数を、前記ソート済み配列が順序付けされていることを利用して数えるソート済み配列頻度計算部と、

前記ソート済み配列内の位置の集合と、特定の文字を入力として、前記ソート済み配列内の値に対する加減算を行い、前記入力された文字に隣接する接尾辞ないし接頭辞の登場位置を参照し、対象とする位置の集合を絞り込むソート済み配列写像部と、

5 を備える、頻出系列の列举装置。

（付記 2） 前記ソート済み配列頻度計算部は、前記ソート済み配列と前記ソート済み配列上の位置の集合を入力として、前記元系列データ上に存在する文字の数を、前記ソート済み配列が順序付けされていることを利用して数

10 える際に、頻度によって 2 分探索を行うか否かを切り変えて用いる、付記 1

（付記 3） 前記ソート済み配列は接尾辞配列から成り、

系列データ集合を入力として、前記系列データの区切りを挿入しつつ、前記接尾辞配列と、該接尾辞配列の逆配列と、前記元系列データとを生成する接尾辞配列作成部を更に備える、付記 1 又は 2 に記載の頻出系列の列举装置。

15 （付記 4） 前記ソート済み配列記憶部は、前記接尾辞配列作成部が生成した、前記接尾辞配列と、前記接尾辞配列の逆配列と、前記元系列データとを記憶する接尾辞配列記憶部から成る、付記 3 に記載の頻出系列の列举装置。

（付記 5） 前記ソート済み配列は接頭辞配列から成り、

20 系列データ集合を入力として、前記系列データの区切りを挿入しつつ、前記接頭辞配列と、該接頭辞配列の逆配列と、前記元系列データとを生成する接頭辞配列作成部を更に備える、付記 1 又は 2 に記載の頻出系列の列举装置。

（付記 6） 前記ソート済み配列記憶部は、前記接頭辞配列作成部が生成した、前記接頭辞配列と、前記接頭辞配列の逆配列と、前記元系列データとを記憶する接頭辞配列記憶部から成る、付記 5 に記載の頻出系列の列举装置。

25 （付記 7） 列举装置を用いて、データ列内から高頻度な系列を列举する方法であって、

系列データについての全接尾辞ないし全接頭辞を辞書順で並べた場合に、辞書上の位置から元系列データ上の位置を参照可能にするソート済み配列を、ソート済み配列記憶部に格納する格納ステップと、

前記ソート済み配列と前記ソート済み配列上の位置の集合を入力として、前記元系列データ上に存在する文字の数を、前記ソート済み配列が順序付けされていることを利用して数える計算ステップと、

- 5 前記ソート済み配列内の位置の集合と、特定の文字を入力として、前記ソート済み配列内の値に対する加減算を行い、前記入力された文字に隣接する接尾辞ないし接頭辞の登場位置を参照し、対象とする位置の集合を絞り込む写像ステップと、
- を備える、頻出系列の列举方法。

- (付記 8) 前記計算ステップは、前記ソート済み配列と前記ソート済み配列上の位置の集合を入力として、前記元系列データ上に存在する文字の数を、前記ソート済み配列が順序付けされていることを利用して数える際に、頻度によって 2 分探索を行うか否かを切り変えて用いる、付記 7 に記載の頻出系列の列举方法。
- 10

- (付記 9) 前記ソート済み配列は接尾辞配列から成り、
- 15 系列データ集合を入力として、前記系列データの区切りを挿入しつつ、前記接尾辞配列と、該接尾辞配列の逆配列と、前記元系列データとを生成する作成ステップを更に備える、付記 7 又は 8 に記載の頻出系列の列举方法。

- (付記 10) 前記格納ステップは、前記ソート済み配列記憶部である接尾辞配列記憶部に、前記作成ステップで生成した、前記接尾辞配列と、前記接尾辞配列の逆配列と、前記元系列データとを格納する、付記 9 に記載の頻出系列の列举方法。
- 20

- (付記 11) 前記ソート済み配列は接頭辞配列から成り、
- 系列データ集合を入力として、前記系列データの区切りを挿入しつつ、前記接頭辞配列と、該接頭辞配列の逆配列と、前記元系列データとを生成する作成ステップを更に備える、付記 7 又は 8 に記載の頻出系列の列举方法。
- 25

- (付記 12) 前記格納ステップは、前記ソート済み配列記憶部である接頭辞配列記憶部に、前記作成ステップで生成した、前記接頭辞配列と、前記接頭辞配列の逆配列と、前記元系列データとを格納する、付記 11 に記載の頻出系列の列举方法。

(付記 1 3) データ列内から高頻度な系列を、コンピュータに列挙させるプログラムを記録したコンピュータ読取可能な記録媒体であって、前記プログラムは前記コンピュータに、

- 系列データについての全接尾辞ないし全接頭辞を辞書順で並べた場合に、
- 5 辞書上の位置から元系列データ上の位置を参照可能にするソート済み配列を、ソート済み配列記憶部に格納する格納手順と、

前記ソート済み配列と前記ソート済み配列上の位置の集合を入力として、前記元系列データ上に存在する文字の数を、前記ソート済み配列が順序付けされていることを利用して数える計算手順と、

- 10 前記ソート済み配列内の位置の集合と、特定の文字を入力として、前記ソート済み配列内の値に対する加減算を行い、前記入力された文字に隣接する接尾辞ないし接頭辞の登場位置を参照し、対象とする位置の集合を絞り込む写像手順と、

を実行させる、コンピュータ読取可能な記録媒体。

- 15 (付記 1 4) 前記計算手順は、前記コンピュータに、前記ソート済み配列と前記ソート済み配列上の位置の集合を入力として、前記元系列データ上に存在する文字の数を、前記ソート済み配列が順序付けされていることを利用して数える際に、頻度によって2分探索を行うか否かを切り変えて用いさせる、ことを特徴とする付記 1 3に記載のコンピュータ読取可能な記録媒体。

- 20 (付記 1 5) 前記ソート済み配列は接尾辞配列から成り、

前記プログラムは前記コンピュータに、系列データ集合を入力として、前記系列データの区切りを挿入しつつ、前記接尾辞配列と、該接尾辞配列の逆配列と、前記元系列データとを生成する作成手順を更に実行させる、付記 1 3又は1 4に記載のコンピュータ読取可能な記録媒体。

- 25 (付記 1 6) 前記格納手順は、前記コンピュータに、前記ソート済み配列記憶部である接尾辞配列記憶部に、前記作成ステップで生成した、前記接尾辞配列と、前記接尾辞配列の逆配列と、前記元系列データとを格納させる、付記 1 5に記載のコンピュータ読取可能な記録媒体。

(付記 1 7) 前記ソート済み配列は接頭辞配列から成り、

前記プログラムは前記コンピュータに、系列データ集合を入力として、前記系列データの区切りを挿入しつつ、前記接頭辞配列と、該接頭辞配列の逆配列と、前記元系列データとを生成する作成手順を更に実行させる、付記 13 又は 14 に記載のコンピュータ読取可能な記録媒体。

- 5 (付記 18) 前記格納手順は、前記コンピュータに、前記ソート済み配列記憶部である接頭辞配列記憶部に、前記作成ステップで生成した、前記接頭辞配列と、前記接頭辞配列の逆配列と、前記元系列データとを格納させる、付記 17 に記載のコンピュータ読取可能な記録媒体。

10 産業上の利用可能性

本発明は、購買ログの分析、DNAの分析、テキストデータの分析、ログデータにおいて、頻出する特徴的なパターンを高速に算出するのに利用可能である。

15 符号の説明

- 20、20A 頻出系列の列挙装置
- 21 SA作成部
- 21A PA作成部
- 22 SA記憶部（ソート済み配列記憶部）
- 20 22A PA記憶部（ソート済み配列記憶部）
- 23 SA頻度計算部（ソート済み配列頻度計算部）
- 23A PA頻度計算部（ソート済み配列頻度計算部）
- 24 SA写像部（ソート済み配列写像部）
- 24A PA写像部（ソート済み配列写像部）
- 25 25、25A 制御部

この出願は、2013年8月23日に提出された日本出願特願2013-173134号を基礎とする優先権を主張し、その開示の全てをここに取り込む。

請 求 の 範 囲

〔請求項 1〕

- 系列データについての全接尾辞ないし全接頭辞を辞書順で並べた場合に、
- 5 辞書上の位置から元系列データ上の位置を参照可能にするソート済み配列を記憶するソート済み配列記憶部と、

前記ソート済み配列と前記ソート済み配列上の位置の集合を入力として、前記元系列データ上に存在する文字の数を、前記ソート済み配列が順序付けされていることを利用して数えるソート済み配列頻度計算部と、

- 10 前記ソート済み配列内の位置の集合と、特定の文字を入力として、前記ソート済み配列内の値に対する加減算を行い、前記入力された文字に隣接する接尾辞ないし接頭辞の登場位置を参照し、対象とする位置の集合を絞り込むソート済み配列写像部と、
- を備える、頻出系列の列举装置。

15 〔請求項 2〕

前記ソート済み配列頻度計算部は、前記ソート済み配列と前記ソート済み配列上の位置の集合を入力として、前記元系列データ上に存在する文字の数を、前記ソート済み配列が順序付けされていることを利用して数える際に、頻度によって 2 分探索を行うか否かを切り変えて用いる、請求項 1 に記載の

- 20 頻出系列の列举装置。

〔請求項 3〕

前記ソート済み配列は接尾辞配列から成り、

- 系列データ集合を入力として、前記系列データの区切りを挿入しつつ、前記接尾辞配列と、該接尾辞配列の逆配列と、前記元系列データとを生成する
- 25 接尾辞配列作成部を更に備える、請求項 1 又は 2 に記載の頻出系列の列举装置。

〔請求項 4〕

前記ソート済み配列記憶部は、前記接尾辞配列作成部が生成した、前記接尾辞配列と、前記接尾辞配列の逆配列と、前記元系列データとを記憶する接

尾辞配列記憶部から成る、請求項 3 に記載の頻出系列の列举装置。

[請求項 5]

前記ソート済み配列は接頭辞配列から成り、

- 5 系列データ集合を入力として、前記系列データの区切りを挿入しつつ、前記接頭辞配列と、該接頭辞配列の逆配列と、前記元系列データとを生成する接頭辞配列作成部を更に備える、請求項 1 又は 2 に記載の頻出系列の列举装置。

[請求項 6]

- 10 前記ソート済み配列記憶部は、前記接頭辞配列作成部が生成した、前記接頭辞配列と、前記接頭辞配列の逆配列と、前記元系列データとを記憶する接頭辞配列記憶部から成る、請求項 5 に記載の頻出系列の列举装置。

[請求項 7]

- 15 列举装置を用いて、データ列内から高頻度な系列を列举する方法であって、系列データについての全接尾辞ないし全接頭辞を辞書順で並べた場合に、辞書上の位置から元系列データ上の位置を参照可能にするソート済み配列を、ソート済み配列記憶部に格納する格納ステップと、

前記ソート済み配列と前記ソート済み配列上の位置の集合を入力として、前記元系列データ上に存在する文字の数を、前記ソート済み配列が順序付けされていることを利用して数える計算ステップと、

- 20 前記ソート済み配列内の位置の集合と、特定の文字を入力として、前記ソート済み配列内の値に対する加減算を行い、前記入力された文字に隣接する接尾辞ないし接頭辞の登場位置を参照し、対象とする位置の集合を絞り込む写像ステップと、

を備える、頻出系列の列举方法。

- 25 [請求項 8]

前記計算ステップは、前記ソート済み配列と前記ソート済み配列上の位置の集合を入力として、前記元系列データ上に存在する文字の数を、前記ソート済み配列が順序付けされていることを利用して数える際に、頻度によって 2 分探索を行うか否かを切り変えて用いる、請求項 7 に記載の頻出系列の列

挙方法。

[請求項 9]

データ列内から高頻度な系列を、コンピュータに列挙させるプログラムを記録したコンピュータ読取可能な記録媒体であって、前記プログラムは前記

5 コンピュータに、

系列データについての全接尾辞ないし全接頭辞を辞書順で並べた場合に、辞書上の位置から元系列データ上の位置を参照可能にするソート済み配列を、ソート済み配列記憶部に格納する格納手順と、

10 前記ソート済み配列と前記ソート済み配列上の位置の集合を入力として、前記元系列データ上に存在する文字の数を、前記ソート済み配列が順序付けされていることを利用して数える計算手順と、

15 前記ソート済み配列内の位置の集合と、特定の文字を入力として、前記ソート済み配列内の値に対する加減算を行い、前記入力された文字に隣接する接尾辞ないし接頭辞の登場位置を参照し、対象とする位置の集合を絞り込む写像手順と、

を実行させる、コンピュータ読取可能な記録媒体。

[請求項 10]

20 前記計算手順は、前記コンピュータに、前記ソート済み配列と前記ソート済み配列上の位置の集合を入力として、前記元系列データ上に存在する文字の数を、前記ソート済み配列が順序付けされていることを利用して数える際に、頻度によって2分探索を行うか否かを切り変えて用いさせる、請求項9に記載のコンピュータ読取可能な記録媒体。

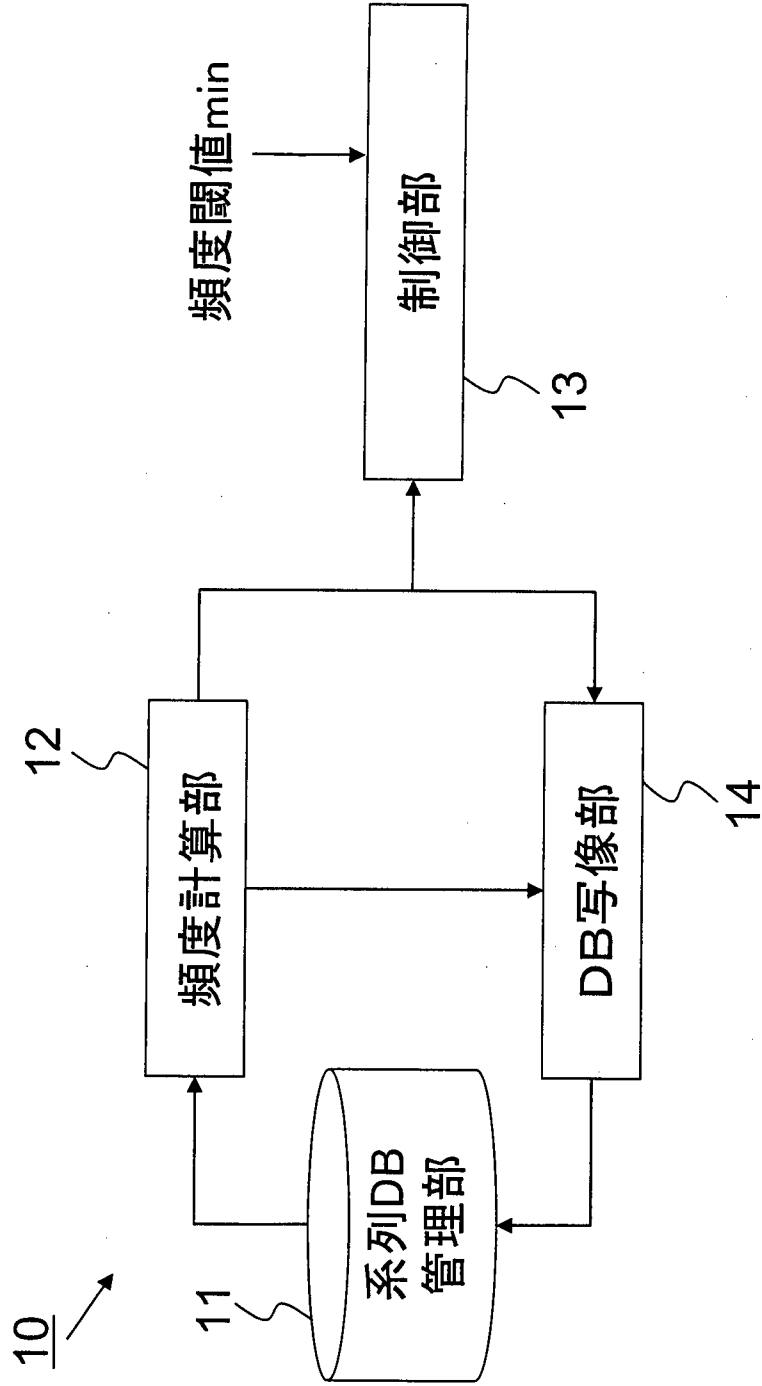


図1 関連技術

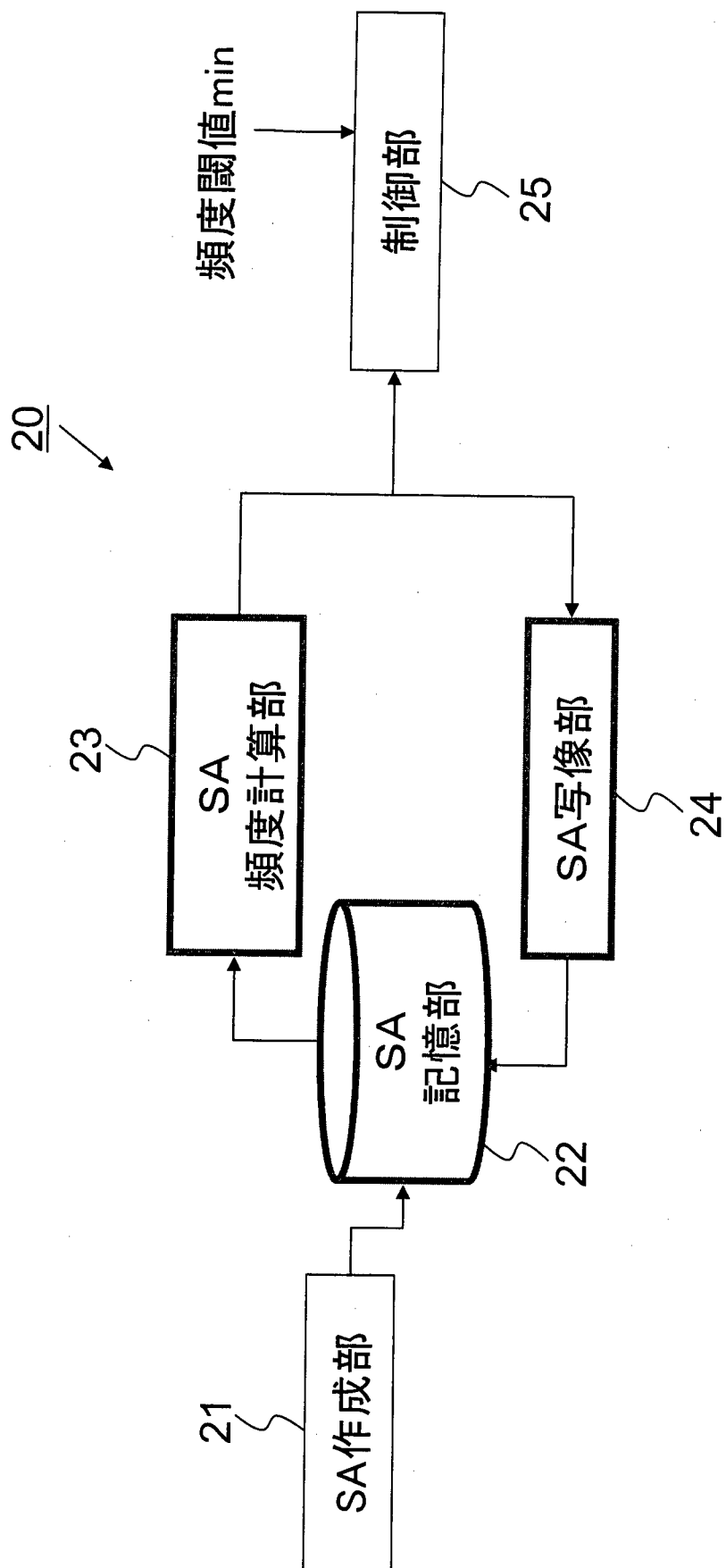


図2

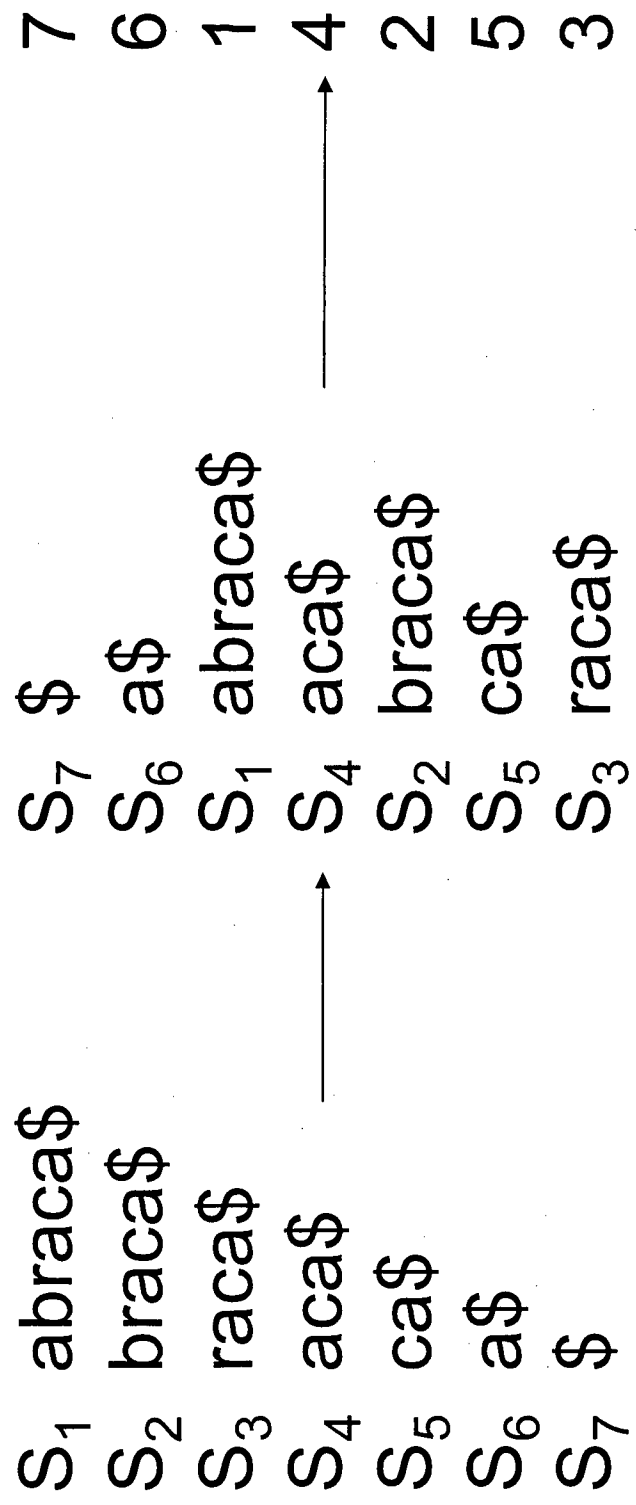


図3

i SA 接尾辞配列
0 11 \$
1 10 a\$
2 7 abra\$
3 0 abracadabra\$
4 3 acadabra\$
5 5 adabra\$
6 8 bra\$
7 1 bracadabra\$
8 4 cadabra\$
9 6 dabra\$
10 9 ra\$
11 2 racadabra\$

図4

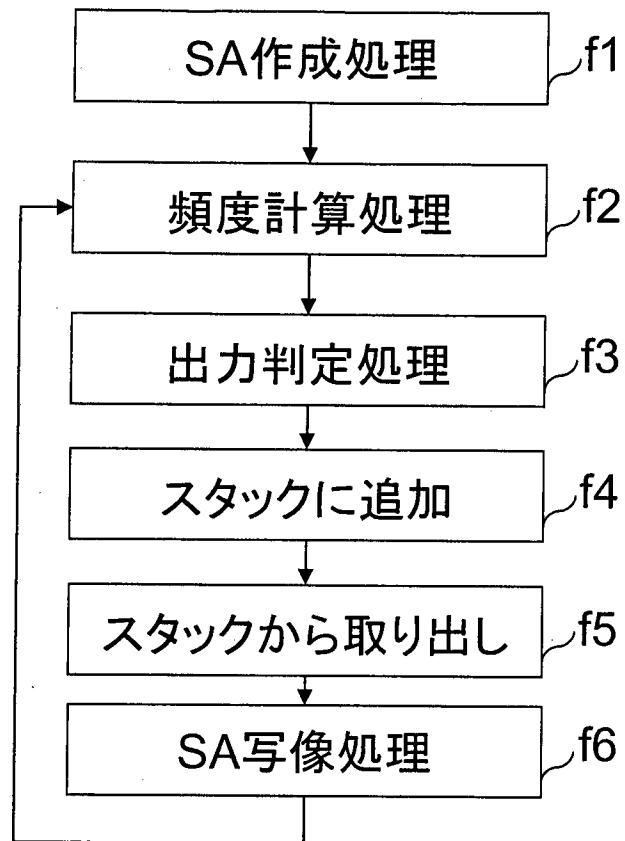


図5

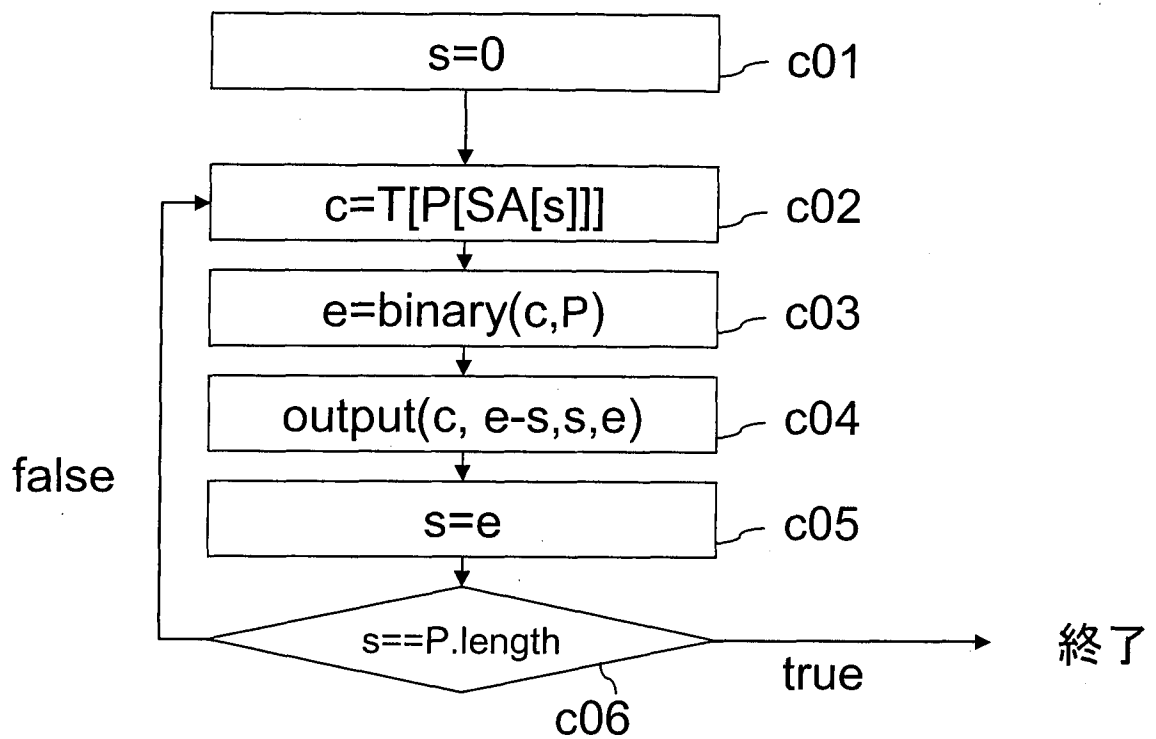


図6

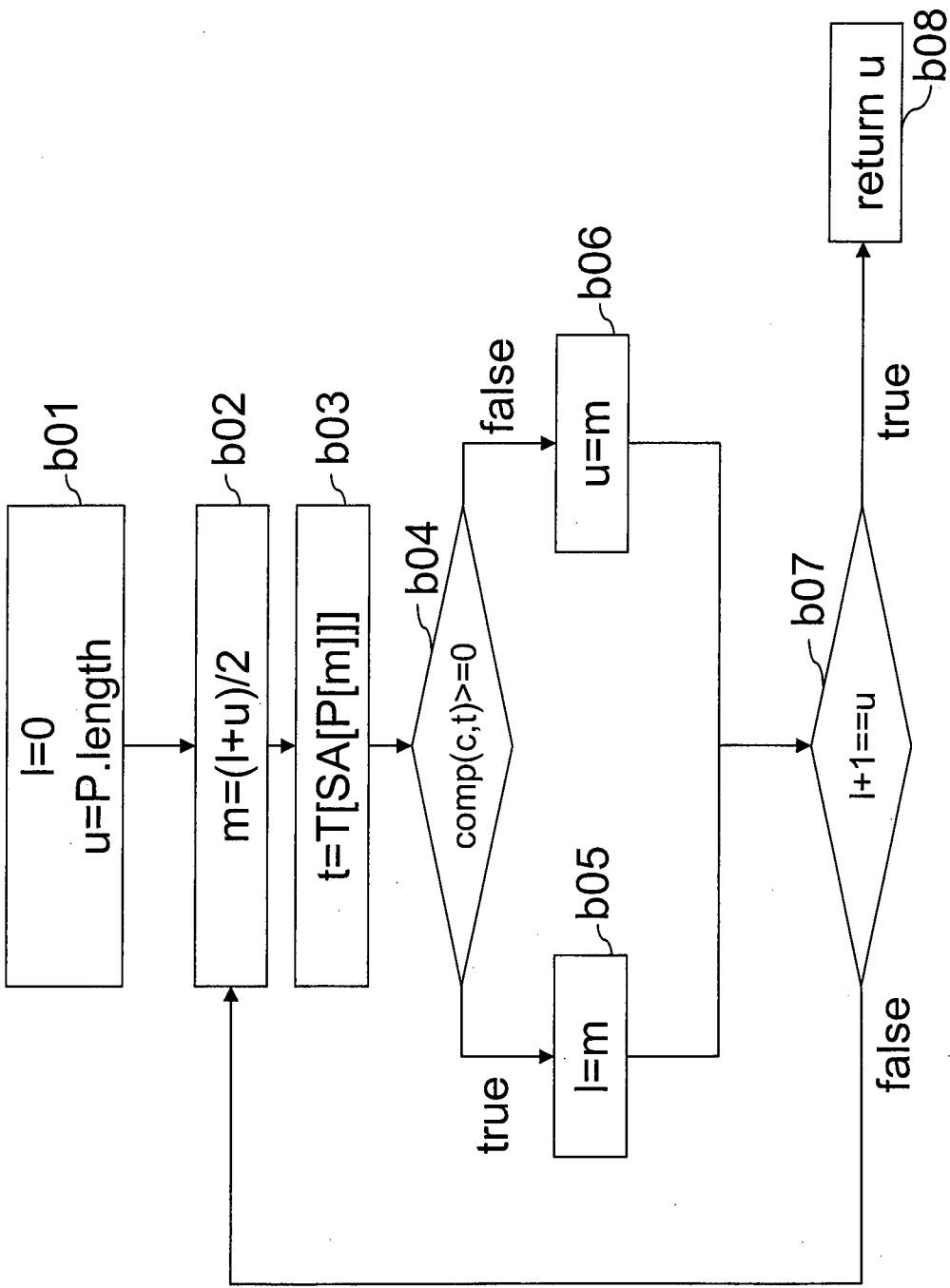


図7

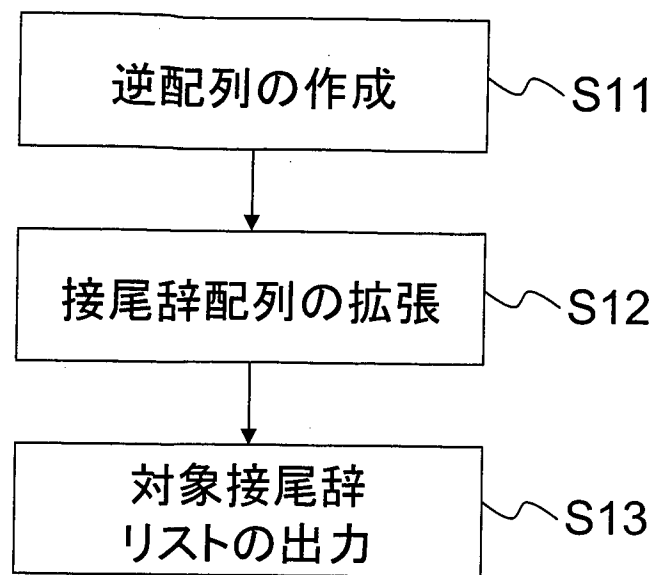


図8

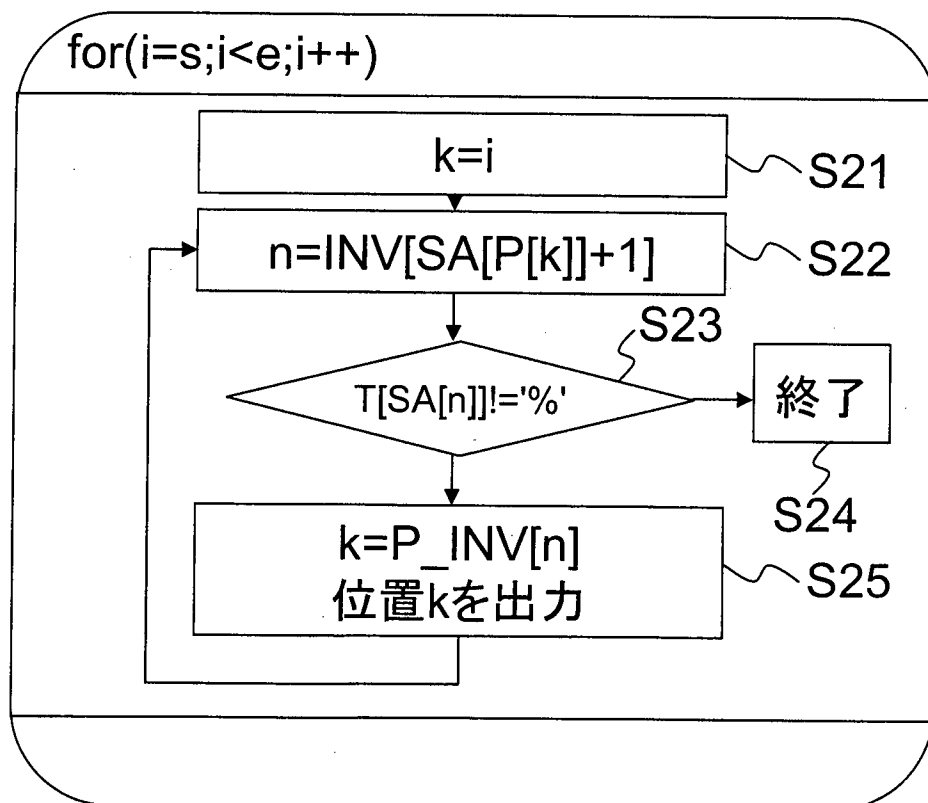


図9

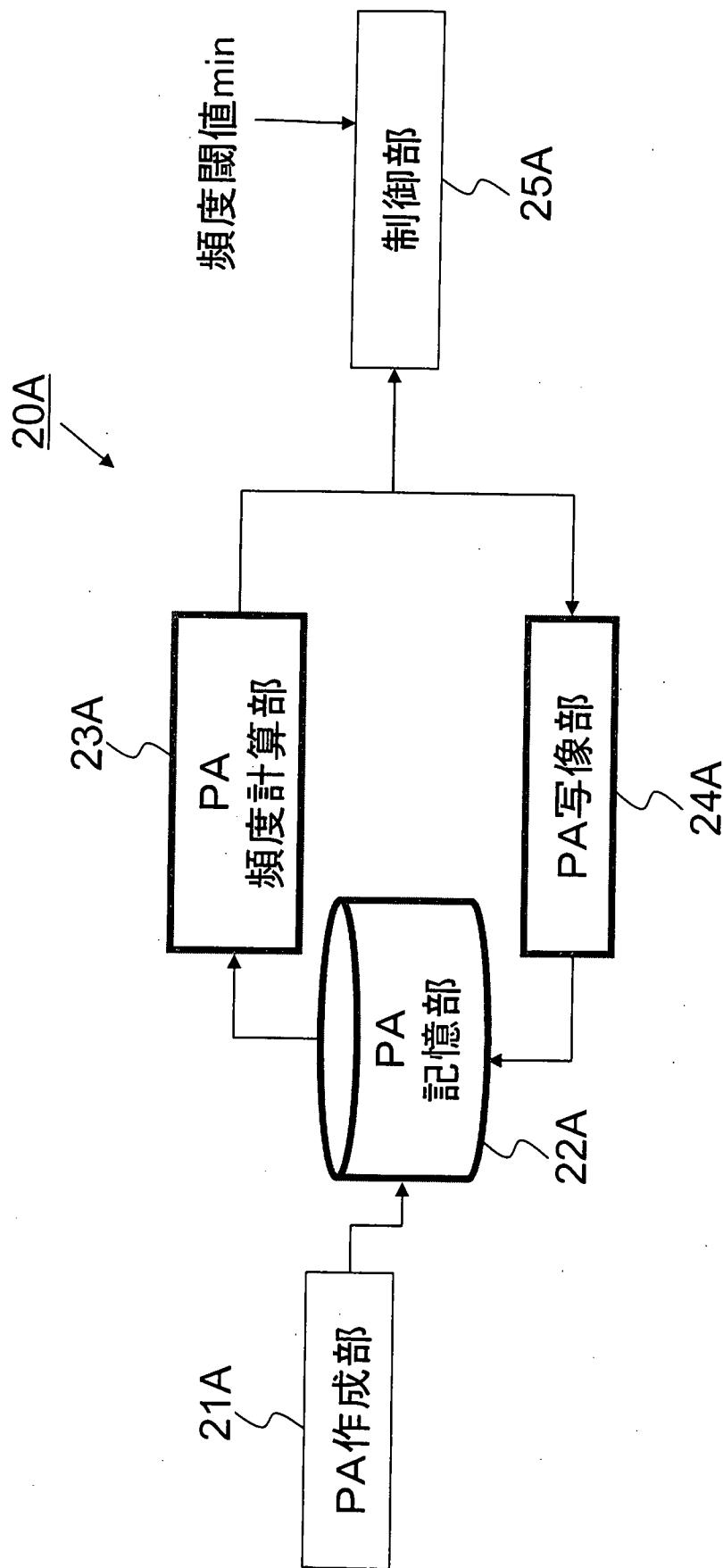


図10

1	SA	接尾辞
0	23	\$
1	22	% \$
2	16	%ABBCA%\$
3	6	%ABCB%ABC%ABBCA%\$
4	0	%CAABC%ABCB%ABC%ABB
5	11	%CABC%ABBCA%\$
6	21	A%\$
7	2	AABC%ABCB%ABC%ABBCA
8	17	ABBCA%\$
9	13	ABC%ABBCA%\$
10	3	ABC%ABCB%ABC%ABBCA%
11	7	ABCB%ABC%ABBCA%\$
12	10	B%ABC%ABBCA%\$
13	18	BBCA%\$
14	14	BC%ABBCA%\$
15	4	BC%ABCB%ABC%ABBCA%\$
16	19	BCA%\$
17	8	BCB%ABC%ABBCA%\$
18	15	C%ABBCA%\$
19	5	C%ABCB%ABC%ABBCA%\$
20	20	CA%\$
21	1	CAABC%ABCB%ABC%ABBC
22	12	CABC%ABBCA%\$
23	9	CB%ABC%ABBCA%\$

文字	頻度	開始位置	終了位置
A	6	6	12
B	6	12	18
C	6	18	24

図12

P[i]	SA	接尾辞
6	21	A % \$
10	3	ABC % ABCB % CAB % CABBCA %
12	10	B % CAB % CABBCA % \$
13	18	BB % CA % \$
14	14	BC % ABBCA % \$
15	4	BC % ABCB % CAB % CABBCA % \$
16	19	BCA % \$
17	8	BCB % CAB % CABBCA % \$
18	15	C % ABBCA % \$
19	5	C % ABCB % CAB % CABBCA % \$
20	20	CA % \$
23	9	CB % CAB % CABBCA % \$

図13

文字	頻度	開始位置	終了位置
A	2	0	2
B	6	2	8
C	6	8	12

図14

P[i]	SA	接尾辞
15	4	BC%ABCB%CABC%ABBCA%\$
19	5	C%ABCB%CABC%ABBCA%\$

図15

P[i]	SA	接尾辞
6	21	A%\$
12	10	B%CABC%ABBCA%\$
13	18	BBCA%\$
16	19	BCA%\$
18	15	C%ABBCA%\$
19	5	C%ABCB%CABC%ABBCA%\$
20	20	CA%\$
23	9	CB%CABC%ABBCA%\$

図16

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2014/071136

A. CLASSIFICATION OF SUBJECT MATTER

G06F17/30(2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F17/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2014
Kokai Jitsuyo Shinan Koho	1971-2014	Toroku Jitsuyo Shinan Koho	1994-2014

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	JP 2004-272639 A (Nippon Telegraph and Telephone Corp.), 30 September 2004 (30.09.2004), paragraphs [0017], [0019] to [0021], [0027], [0029] to [0031], [0035]; fig. 9, 10 (Family: none)	1-10
Y	JP 2002-229987 A (International Business Machines Corp.), 16 August 2002 (16.08.2002), paragraphs [0017], [0028], [0055] & US 2002/0123995 A1	1-10

☒ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
24 October, 2014 (24.10.14)

Date of mailing of the international search report
04 November, 2014 (04.11.14)

Name and mailing address of the ISA/
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2014/071136

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	Akiko AIZAWA, "Ta-class Bunsho Bunrui ni Okeru Ziv-Merhav Crossparsing no Tekiyo to Hyoka", Transactions of Information Processing Society of Japan Ronbunshi Journal, vol.52, no.10, [CD-ROM], 31 October 2011 (31.10.2011) (received date), vol.52, no.10, pages 2953 to 2964	2, 8, 10
A	JP 2013-061745 A (Nippon Telegraph and Telephone Corp.), 04 April 2013 (04.04.2013), entire text; all drawings (Family: none)	1-10
A	JP 2001-067378 A (Sumitomo Electric Industries, Ltd.), 16 March 2001 (16.03.2001), entire text; all drawings (Family: none)	1-10
A	JP 2003-228571 A (Kyoji UMEMURA), 15 August 2003 (15.08.2003), entire text; all drawings (Family: none)	1-10
A	Naoya ITO, Recent Perl World Dai 18 Kai Suffix Array (Setsubiji Hairetsu) Algorithm & Data Kozo Series 1, WEB+DB PRESS, 15 May 2009 (15.05.2009), vol.50, 1st edition, pages 185 to 193	1-10
A	Naoki TANIDA et al., "Hardware Acceleration of Full-Text Search Using Succinct Data Structure", IEICE Technical Report, 29 July 2008 (29.07.2008), vol.108, no.180, pages 7 to 12	1-10

A. 発明の属する分野の分類（国際特許分類（I P C））

Int.Cl. G06F17/30(2006.01)i

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（I P C））

Int.Cl. G06F17/30

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1 9 2 2 - 1 9 9 6 年
日本国公開実用新案公報	1 9 7 1 - 2 0 1 4 年
日本国実用新案登録公報	1 9 9 6 - 2 0 1 4 年
日本国登録実用新案公報	1 9 9 4 - 2 0 1 4 年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
Y	JP 2004-272639 A（日本電信電話株式会社） 2004.09.30, 段落【0017】，【0019】－【0021】，【0027】， 【0029】－【0031】，【0035】，【図9】，【図10】 （ファミリーなし）	1-10
Y	JP 2002-229987 A（インターナショナル・ビジネス・マシーンズ・ コーポレーション） 2002.08.16, 段落【0017】，【0028】，【0055】 & US 2002/0123995 A1	1-10

☒ C欄の続きにも文献が列挙されている。☐ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー

「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「O」口頭による開示、使用、展示等に言及する文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願

の日の後に公表された文献

「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」同一パテントファミリー文献

国際調査を完了した日

2 4 . 1 0 . 2 0 1 4

国際調査報告の発送日

0 4 . 1 1 . 2 0 1 4

国際調査機関の名称及びあて先

日本国特許庁（I S A / J P）
郵便番号 1 0 0 - 8 9 1 5
東京都千代田区霞が関三丁目4番3号

特許庁審査官（権限のある職員）

齊藤 貴孝

5 M

4 7 7 4

電話番号 0 3 - 3 5 8 1 - 1 1 0 1 内線 3 5 9 9

C (続き) . 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
Y	相澤 彰子, 多クラス文書分類における Z i v - M e r h a v C r o s s p a r s i n g の適用と評価, 情報処理学会論文誌 論 文誌ジャーナル V o l . 5 2 N o . 1 0 [C D - R O M] , 2011.10.31 (受入日), 第52巻, 第10号, p. 2953-29 64	2, 8, 10
A	JP 2013-061745 A (日本電信電話株式会社) 2013.04.04, 全文, 全図 (ファミリーなし)	1-10
A	JP 2001-067378 A (住友電気工業株式会社) 2001.03.16, 全文, 全図 (ファミリーなし)	1-10
A	JP 2003-228571 A (梅村 恭司) 2003.08.15, 全文, 全図 (ファミリーなし)	1-10
A	伊藤 直也, Recent Perl World 第18回 S u f f i x A r r a y (接尾辞配列) アルゴリズム&データ構造シリーズ1, WEB+DB PRESS, 2009.05.15, 第50巻, 初版, p. 185-193	1-10
A	谷田 直輝、外5名, 簡潔データ構造による全文検索のハードウェ アを用いた高速化, 電子情報通信学会技術研究報告, 2008.07.29, 第108巻, 第180号, p. 7-12	1-10