

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2005-84859

(P2005-84859A)

(43) 公開日 平成17年3月31日(2005.3.31)

(51) Int. Cl. <sup>7</sup>	F I	テーマコード (参考)
G06F 17/30	G06F 17/30 170A	4B024
// C12N 15/09	G06F 17/30 170F	5B075
	G06F 17/30 220Z	
	C12N 15/00 A	

審査請求 未請求 請求項の数 12 O L (全 17 頁)

(21) 出願番号	特願2003-315129 (P2003-315129)	(71) 出願人	800000035 株式会社産学連携機構九州 福岡県福岡市東区箱崎6丁目10番1号
(22) 出願日	平成15年9月8日(2003.9.8)	(74) 代理人	100103621 弁理士 林 靖
特許法第30条第1項適用申請有り		(72) 発明者	池田 大輔 福岡県糟屋郡志免町大字南里433-1ピアザ橋407
		(72) 発明者	山田 泰寛 福岡県福岡市東区千早1-7-11 レオパレス千早203号室
		(72) 発明者	廣川 佐千男 福岡県福岡市城南区荒江1丁目32-24-802
		Fターム(参考)	4B024 AA11 CA01 CA11 HA11 5B075 ND03 NS10 UU06 UU19

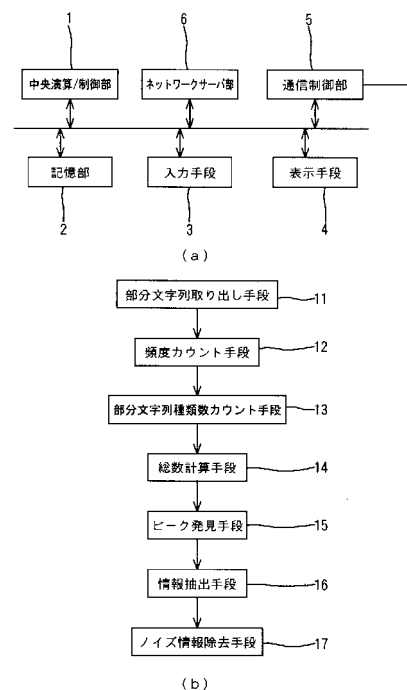
(54) 【発明の名称】 共通パターン発見装置とプログラム、記憶媒体、及び共通パターン発見方法

(57) 【要約】

【課題】本発明の解決しようとする問題点は、複数の情報間で共通のパターンを容易に発見することである。

【解決手段】本発明は、各テキスト情報から部分文字列を抽出する部分文字列取り出し手段11と、抽出した部分文字列の出現回数をカウントして同一の部分文字列ごとに出現回数の和をとって頻度とする頻度カウント手段12と、同一頻度ごとに部分文字列取り出し手段11が取り出した異なる部分文字列の数をカウントする部分文字列種類数カウント手段13と、頻度と異なる部分文字列の数との積を計算する総数計算手段14と、総数計算手段14によって計算された積と頻度との関係から、ピーンポイントで出現するピークに位置の頻度を探すピーク発見手段15と、ピークが存在するとき該ピークの位置で頻度がカウントされた部分文字列を含むテキスト情報を抽出する情報抽出手段16とを備えたことを主要な特徴とする。

【選択図】図7



## 【特許請求の範囲】

## 【請求項 1】

複数のテキスト情報を対象として、各テキスト情報から最大長さまでのすべての長さの部分文字列を抽出する部分文字列取り出し手段と、前記部分文字列取り出し手段が抽出した部分文字列の出現回数をカウントして同一の部分文字列ごとに出現回数の和をとって頻度とする頻度カウント手段と、同一頻度ごとに前記部分文字列取り出し手段が取り出した異なる部分文字列の数をカウントする部分文字列種類数カウント手段と、前記頻度カウント手段がカウントした頻度と前記部分文字列種類数カウント手段がカウントした異なる部分文字列の数との積を計算する総数計算手段と、前記総数計算手段によって計算された積と前記頻度との関係から、ピンポイントでピークが出現する位置の頻度を探すピーク発見手段と、ピークが存在するとき該ピークの位置の頻度がカウントされた部分文字列を含むテキスト情報を抽出する情報抽出手段とを備え、複数のテキスト情報の中からピークの存在で共通のパターンを有す情報を発見することを特徴とする共通パターン発見装置。

10

## 【請求項 2】

共通のパターンを有しないテキスト情報を除去するノイズ情報除去手段が設けられたことを特徴とする請求項 1 記載の共通パターン発見装置。

## 【請求項 3】

前記情報抽出手段が、共通のパターンを示す部分文字列を識別可能に表示して該部分文字列を含むテキスト情報を抽出することを特徴とする請求項 1 または 2 記載の共通パターン発見装置。

20

## 【請求項 4】

請求項 1 ~ 3 のいずれかに記載の共通パターン発見装置において、複数のテキスト情報を対象として部分文字列を抽出するのに代えて、単数のテキスト情報、複数の塩基配列情報または画像情報を対象としてそれぞれ繰り返しパターン、部分塩基配列または部分画素列を抽出することを特徴とする共通パターン発見装置。

## 【請求項 5】

コンピュータを、複数のテキスト情報を対象として各テキスト情報から最大長さまでのすべての長さの部分文字列を抽出する部分文字列取り出し手段、前記部分文字列取り出し手段が抽出した部分文字列の出現回数をカウントして同一の部分文字列ごとに出現回数の和をとって頻度とする頻度カウント手段、同一頻度ごとに前記部分文字列取り出し手段が取り出した異なる部分文字列の数をカウントする部分文字列種類数カウント手段、前記頻度カウント手段がカウントした頻度と前記部分文字列種類数カウント手段がカウントした異なる部分文字列の数との積を計算する総数計算手段、前記総数計算手段によって計算された積と前記頻度との関係から、ピンポイントでピークが出現する位置の頻度を探すピーク発見手段、ピークが存在するとき該ピークの位置の頻度がカウントされた部分文字列を含むテキスト情報を抽出する情報抽出手段として機能させるためのプログラムであって、複数のテキスト情報の中からピークの存在で共通のパターンを有す情報を発見することを特徴とするプログラム。

30

## 【請求項 6】

コンピュータを、共通のパターンを有しないテキスト情報を除去するノイズ情報除去手段として機能させることを特徴とする請求項 5 記載のプログラム。

40

## 【請求項 7】

前記情報抽出手段が、共通のパターンを示す部分文字列を識別可能に表示して該部分文字列を含むテキスト情報を抽出することを特徴とする請求項 5 または 6 記載のプログラム。

## 【請求項 8】

請求項 5 ~ 7 のいずれかに記載のプログラムにおいて、複数のテキスト情報を対象として部分文字列を抽出するのに代えて、単数のテキスト情報、複数の塩基配列情報または画像情報を対象としてそれぞれ繰り返しパターン、部分塩基配列または部分画素列を抽出することを特徴とするプログラム。

## 【請求項 9】

50

請求項 5 ~ 8 のいずれかに記載のプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項 10】

複数のテキスト情報を対象として各テキスト情報から最大長さまでのすべての長さの部分文字列を抽出し、同一の部分文字列ごとに出現回数の和をとって頻度とするとともに該頻度を有する異なる部分文字列の数をカウントし、前記頻度と前記異なる部分文字列の数との積を計算し、該積と前記頻度との関係からピンポイントでピークが出現する位置の頻度を探し、ピークが存在するとき該ピークの位置の頻度がカウントされた部分文字列を含むテキスト情報を抽出して、複数のテキスト情報の中から共通のパターンを有す情報を発見することを特徴とする共通パターン発見方法。

10

【請求項 11】

共通のパターンを有しないテキスト情報を除去するノイズ情報除去手段が設けられたことを特徴とする請求項 10 記載の共通パターン発見方法。

【請求項 12】

請求項 10 または 11 に記載された共通パターン発見方法において、複数のテキスト情報を対象とするのに代えて、単数のテキスト情報、複数の塩基配列情報または画像情報を対象とすることを特徴とする共通パターン発見方法。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、共通のパターンを含むテキストデータもしくは共通のパターンで記述されたウェブページ、アーカイブ、あるいは共通の特徴を含む画像、共通の DNA の塩基配列情報を簡単に収集することができる共通パターン発見装置とそのために使用するプログラム、記憶媒体、及び共通パターン発見方法に関する。

20

【背景技術】

【0002】

ウェブ上には、HTML や XML 等で記述された多種多様のウェブページや、メール、ニュース等のアーカイブなど、マークアップ言語で記述されたテキストデータが大量に存在している。そしてこれらのテキストデータには同種の表現を繰返して記述するものが多数存在する。例えば、オークションのリストは 1 つのウェブページ中に商品に関するデータ（製品名、型番、購入日、傷の有無、保証書の有無など）が繰返して表示される。また、新聞や株式に関するウェブサイト等では、分野や発刊日時、企業名等によって整理された記事や経済情報が整然とそれぞれ同一形式で表示されている。こうした共通のパターンを有する情報を発見するのは人間の判断以外には困難というのが現状である。唯一、ウェブページに関しては、共通のキーワードによって検索エンジンで探し、ブラウザで閲覧して要不要の判断を行い、抽出している。なお、多くのキーワードは、通常、自然言語から選ばれる。

30

【0003】

このウェブページに関して、本発明者らは、ウェブ上の同種ファイルを集めることができればデータベースのような使い方が可能になるとの考えから、構造の類似するウェブページを簡単に収集することができる類似構造ファイル収集方法を提案した（特願 2003-101944）。この際、自然言語の単語もしくは経験に基づく固定的な適宜の文字数で文字列を抽出するのでは、辞書の大きさや偶然に影響されるため、自然言語や偶然によらずに抽出する方法を採用した（非特許文献 1 参照）。

40

【0004】

すなわち、この類似構造ファイル収集方法は、複数のウェブページ情報を対象とし、マークアップ言語で記述されたそれぞれのテキストデータから所定の計算法で決定された文字数の文字列を抽出し、その出現頻度をカウントするとともに、カウントされたすべての出現頻度の中から高頻出文字列として評価するため所定の計算法で決定された所定の割合以上の出現頻度で出現する文字列の文字数をカウントし、各ウェブページ情報でカウント

50

された文字数を比較して同一クラスタに構成できるウェブページ情報同士を統合することによって、対象の全ウェブページ情報を複数のウェブページ情報群に分け、母数が少ないウェブページ情報群をノイズクラスタとして除去し、複数のウェブページ情報の中から類似構造のウェブページ情報を抽出する。なお、上記計算法はウェブページ情報の頻出部分と非頻出部分との境界の数が初期値の近くで極小となるときの文字数と割合を、抽出する文字数と高頻出文字列の割合に決定するものである。そして、この類似構造ファイル収集方法は遺伝子の塩基配列情報の解析にも利用できるものであった。

【0005】

しかし、本発明者らが提案したこの類似構造ファイル収集方法は、自然言語や偶然によらない画期的なものであったが、極小値の計算方法に課題が残るものであった。また、頻度を用いないものより計算時間は短くなったが、改善の余地があった。さらに、この方法は高頻度で出現するのは構造を示す記述部分と考えるため、タグ等が記述されたHTML等に適しており、文章表現などのあらゆる部分で共通のパターンを発見するものではなかった。

10

【0006】

ところで、従来テキスト情報中の文章表現に関して、使用されている単語と出現頻度との間に、ジップの法則 (Zipf's law) が成立することはよく知られている。これはこの法則の発見者が、英文テキストと単語を材料にして発見した関係であるが、現在では欧州系等の言語、ウェブページの被リンク数、都市の人口の偏在状態、論文の参照件数などの出現頻度が絡む多くのまとまりのあるデータでごく普通に拡張的に成立すると考えられている法則である。

20

【0007】

さて、このジップの第1法則は、テキスト中の単語を出現頻度順に並べたとき、順位  $r$  とその頻度  $f$  の積が定数  $C$  になるというもので、 $f \times r = C$  の関係が成立するというものである。また、ジップの第2法則は、テキスト中の単語の頻度分布、とくに低頻度部分において、頻度が  $f$  である単語の種類数  $V(f)$  は頻度  $f$  との間に、 $\log V(f) = -a(\log f) + b$  という関係が成立する、というものである。ここで  $a, b$  は情報ごとに存在する定数であり、 $a > 0$  である。図13はジップの第2法則を示す説明図である。

【0008】

しかしながら、このジップの法則は情報間で共通のパターンを有する情報を発見するのに寄与するものではない。さらに、ジップの法則は、本来、英文のように各単語がスペースを挟んで分離して配置されるような場合に成り立つ法則であるため、様々の助詞等を使って単語が次々と切れ目なく続く日本語や中国語等の言語、構造に関する記述を含むマークアップ言語、4つの塩基が様々のパターンで繰り返し並ぶDNA、さらには画像データ等の場合に、どのように文字列を抽出するかについては示唆するところがない。

30

【0009】

【非特許文献1】池田, 山田, 廣川「Eliminating Useless Parts in Semi-structured Documents using AlternationCounts」, In Proceedings of the4th International Conference on Discovery Science, Lecture Notes in Artificial Intelligence (ドイツ国), Springer-Verlag, 2001年11月, 第2226巻, p. 113 - 127

40

【発明の開示】

【発明が解決しようとする課題】

【0010】

従来、ウェブサイトの情報を構造化し、属性名のない表情報に変換する研究がすすめられている。しかし、これらはHTML等に限られ、一般的な手段ではなく、情報間で共通の表現やパターンを発見するものではない。

【0011】

また、本発明者らによる類似構造ファイル収集方法及び非特許文献1の提案は、文字数と高頻出文字列の割合を決定する極小値の計算の妥当性に関して、今後の検証に俟たなければならないものである。すなわち極小値の決定方法に改良の余地があるものであった。

50

さらにこの方法は上述の表情法に関する研究と同様、多数の情報において高頻度で出現するのは構造を示す記述部分と考えるため、HTML等以外の他のテキスト情報は共通のパターンを発見することはできない。そして、ジップの法則だけでは、文字列がファイル間で共通のパターンを示すものであるか否かの判断はできない。

【0012】

さらに、最近の遺伝子解析で多くの塩基配列情報が得られているが、解読した遺伝子情報をもとに類似した塩基配列情報を抜き出そうとしても、これが非常な難問であることが理論的に知られており、これを克服し簡単な計算で抜き出す方法は、現在のところ見当たらない。

【0013】

そこで本発明は、複数の情報間で共通のパターンを容易に発見することができる共通パターン発見装置を提供することを目的とする。

【0014】

また本発明は、複数の情報間で共通のパターンを容易に発見することができるプログラムを提供することを目的とする。

【0015】

そして本発明は、複数の情報間で共通のパターンを容易に発見することができるプログラムを記録した記憶媒体を提供することを目的とする。

【0016】

さらに本発明は、複数の情報間で共通のパターンを容易に発見することができる共通パターン発見方法を提供することを目的とする。

【課題を解決するための手段】

【0017】

本発明は、複数のテキスト情報を対象として、各テキスト情報から最大長さまでのすべての長さの部分文字列を抽出する部分文字列取り出し手段と、部分文字列取り出し手段が抽出した部分文字列の出現回数をカウントして同一の部分文字列ごとに出現回数の和をとって頻度とする頻度カウント手段と、同一頻度ごとに部分文字列取り出し手段が取り出した異なる部分文字列の数をカウントする部分文字列種類数カウント手段と、頻度カウント手段がカウントした頻度と部分文字列種類数カウント手段がカウントした異なる部分文字列の数との積を計算する総数計算手段と、総数計算手段によって計算された積と頻度との関係から、ピンポイントでピークが出現する位置の頻度を探すピーク発見手段と、ピークが存在するとき該ピークの位置の頻度がカウントされた部分文字列を含むテキスト情報を抽出する情報抽出手段とを備え、複数のテキスト情報の中からピークの存在で共通のパターンを有す情報を発見することを主要な特徴とする。

【発明の効果】

【0018】

本発明の共通パターン発見装置とプログラム、記録媒体、共通パターン発見方法によれば、部分文字列を抽出してその頻度と同一頻度となる異なった部分文字列の数を数えて、両者の積を計算し、ピークの存在の有無をみるだけであるから、共通パターンを示す情報を簡単に発見できる。計算時間は格段に少なく、きわめてシンプルな構成、手法であるから、拡張、応用が容易であり、データベースの統合に有効となる。また、解読された遺伝子情報をもとに類似した塩基配列情報を簡単な計算で抜き出すことができる。

【発明を実施するための最良の形態】

【0019】

まず本発明を実施するための第1の形態は、複数のテキスト情報を対象として、各テキスト情報から最大長さまでのすべての長さの部分文字列を抽出する部分文字列取り出し手段と、部分文字列取り出し手段が抽出した部分文字列の出現回数をカウントして同一の部分文字列ごとに出現回数の和をとって頻度とする頻度カウント手段と、同一頻度ごとに部分文字列取り出し手段が取り出した異なる部分文字列の数をカウントする部分文字列種類数カウント手段と、頻度カウント手段がカウントした頻度と部分文字列種類数カウント手

10

20

30

40

50

段がカウントした異なる部分文字列の数との積を計算する総数計算手段と、総数計算手段によって計算された積と頻度との関係から、ピンポイントでピークが出現する位置の頻度を探すピーク発見手段と、ピークが存在するとき該ピークの位置の頻度がカウントされた部分文字列を含むテキスト情報を抽出する情報抽出手段とを備え、複数のテキスト情報の中からピークの存在で共通のパターンを有す情報を発見する共通パターン発見装置であり、部分文字列を抽出してその頻度と同一頻度となる異なる部分文字列の数を数えて、両者の積を計算して同一頻度ごとに文字列の総数を求め、この総数のピークの存在の有無をみるだけであるから、共通パターンを示す情報を簡単に発見できる。計算時間は格段に少なくなり、きわめてシンプルな構成であるから、拡張、応用が容易である。共通部分はテンプレートの部分であり、それ以外はコンテンツ部分と考えられ、データベースの統合に有効となる。

10

**【0020】**

本発明を実施するための第2の形態は、第1の形態において、共通のパターンを有しないテキスト情報を除去するノイズ情報除去手段が設けられた共通パターン発見装置であり、ノイズ情報を除いて共通のパターンを有する情報だけを集めることができる。

**【0021】**

本発明を実施するための第3の形態は、第1または2の形態において、情報抽出手段が、共通のパターンを示す部分文字列を識別可能に表示して該部分文字列を含むテキスト情報を抽出する共通パターン発見装置であり、テキスト情報の共通のパターンを示す部分文字列に文字修飾、例えば色づけ等を加えて表示するから、抽出された情報がどのような共通パターンをどこに有するか、直ちに確認することができる。

20

**【0022】**

本発明を実施するための第4の形態は、第1～3のいずれかの形態の共通パターン発見装置において、複数のテキスト情報を対象として部分文字列を抽出するのに代えて、単数のテキスト情報、複数の塩基配列情報または画像情報を対象としてそれぞれ繰り返しパターン、部分塩基配列または部分画素列を抽出する共通パターン発見装置であり、単数のテキスト情報の中から繰り返しのパターンを抽出でき、DNA等の塩基配列情報の中から共通の塩基配列をみつけることができ、画像情報の中で共通の画素配列を抽出して、同一の被写体を発見することができる。これにより、複数の画像の接合や、筆跡、指紋等の同一性判断を行うことができる。

30

**【0023】**

本発明を実施するための第5の形態は、コンピュータを、複数のテキスト情報を対象として各テキスト情報から最大長さまでのすべての長さの部分文字列を抽出する部分文字列取り出し手段、部分文字列取り出し手段が抽出した部分文字列の出現回数をカウントして同一の部分文字列ごとに出現回数の和をとって頻度とする頻度カウント手段、同一頻度ごとに部分文字列取り出し手段が取り出した異なる部分文字列の数をカウントする部分文字列種類数カウント手段、頻度カウント手段がカウントした頻度と部分文字列種類数カウント手段がカウントした異なる部分文字列の数との積を計算する総数計算手段、総数計算手段によって計算された積と頻度との関係から、ピンポイントでピークが出現する位置の頻度を探すピーク発見手段、ピークが存在するとき該ピークの位置の頻度がカウントされた部分文字列を含むテキスト情報を抽出する情報抽出手段として機能させるためのプログラムであって、複数のテキスト情報の中からピークの存在で共通のパターンを有す情報を発見するプログラムであり、コンピュータを使って、部分文字列を抽出してその頻度と同一頻度となる異なる部分文字列の数を数え、両者の積を計算して同一頻度ごとに文字列の総数を求め、この総数のピークの存在の有無をみるだけであるから、共通パターンを示す情報を簡単に発見できる。計算時間は格段に少なくなり、きわめてシンプルな構成であるから、プログラムの拡張、応用が容易である。

40

**【0024】**

本発明を実施するための第6の形態は、第5の形態において、コンピュータを、共通のパターンを有しないテキスト情報を除去するノイズ情報除去手段として機能させるプログ

50

ラムであり、ノイズ情報を除いて定型部分を有する情報だけを集めることができる。

【0025】

本発明を実施するための第7の形態は、第5または6の形態において、情報抽出手段が、共通のパターンを示す部分文字列を識別可能に表示して該部分文字列を含むテキスト情報を抽出するプログラムであり、テキスト情報の共通のパターンを示す部分文字列に文字修飾等を加えて表示するから、抽出された情報が共通パターンを有するか否か、直ちに確認することができる。

【0026】

本発明を実施するための第8の形態は、第5～7のいずれかの形態のプログラムにおいて、複数のテキスト情報を対象として部分文字列を抽出するのに代えて、単数のテキスト情報、複数の塩基配列情報または画像情報を対象としてそれぞれ繰り返しパターン、部分塩基配列または部分画素列を抽出するプログラムであり、単数のテキスト情報の中から繰り返しのパターンを抽出でき、DNA等の塩基配列情報の中から共通の塩基配列を見つけることができ、画像情報の中で共通の画素配列を抽出して、同一の被写体を発見することができる。これにより、複数の画像の接合や、筆跡、指紋等の同一性判断を行うことができる。

10

【0027】

本発明を実施するための第9の形態は、第5～8のいずれかの形態のプログラムを記録したコンピュータ読み取り可能な記録媒体であり、プログラムの保存に適する。

【0028】

本発明を実施するための第10の形態は、複数のテキスト情報を対象として各テキスト情報から最大長さまでのすべての長さの部分文字列を抽出し、同一の部分文字列ごとに出現回数の和をとって頻度とするとともに該頻度を有する異なる部分文字列の数をカウントし、頻度と異なる部分文字列の数との積を計算し、該積と頻度との関係からピンポイントでピークが出現する位置の頻度を探し、ピークが存在するとき該ピークの位置の頻度がカウントされた部分文字列を含むテキスト情報を抽出して、複数のテキスト情報の中から共通のパターンを有す情報を発見する共通パターン発見方法であり、部分文字列を抽出してその頻度と同一頻度となる異なる部分文字列の数を数えて、両者の積を計算して同一頻度ごとに文字列の総数を求め、この総数のピークの存在の有無をみるだけであるから、共通パターンを示す情報を簡単に発見できる。計算時間は格段に少なくなり、きわめてシンプルな構成であるから、拡張、応用が容易である。共通部分はテンプレートの部分であり、それ以外はコンテンツ部分と考えられ、データベースの統合に効果的となる。

20

30

【0029】

本発明を実施するための第11の形態は、第10の形態において、共通のパターンを有しないテキスト情報を除去するノイズ情報除去手段が設けられた共通パターン発見方法であり、ノイズ情報を除いて定型部分を有する情報だけを集めることができる。

【0030】

本発明を実施するための第12の形態は、第10または11の形態の共通パターン発見方法において、複数のテキスト情報を対象とするのに代えて、単数のテキスト情報、複数の塩基配列情報または画像情報を対象とする共通パターン発見方法であり、単数のテキスト情報の中から繰り返しのパターンを抽出でき、DNA等の塩基配列情報の中から共通の塩基配列を見つけることができ、画像情報の中で共通の被写体を発見することにより、複数の画像の接合や、筆跡、指紋等の同一性判断を行うことができる。

40

【0031】

(実施の形態1)

以下、本発明の実施の形態1における共通パターン発見装置と、そのプログラム、またそれを記録したコンピュータ読み取り可能な記録媒体、さらにその共通パターン発見方法について説明する。実施の形態1の共通パターン発見装置と共通パターン発見方法、プログラム等は、情報間で、共通のパターンを示す定型部分を有する情報と、このような部分を有していない情報とを、情報に含まれる異なる部分文字列の数とその頻度とを利用して

50

抽出するものである。複数の情報間の場合を説明するが、単数の情報内で繰り返しパターンを抽出することもできる。図1は定型部分を有していない情報の部分文字列が出現する頻度  $f$ 、異なる部分文字列の数  $V(f)$ 、部分文字列長さ  $n$  の3次元説明図、図2は定型部分を有している情報の部分文字列が出現する頻度  $f$ 、異なる部分文字列の数  $V(f)$ 、部分文字列長さ  $n$  の3次元説明図、図3は定型部分を有していない情報の異なる部分文字列の数  $V(f)$  と頻度  $f$  の2次元説明図、図4は定型部分を有している情報の異なる部分文字列の数  $V(f)$  と頻度  $f$  の2次元説明図、図5は定型部分を有していない情報の頻度  $f$  と部分文字列の総数  $F(f)$  の関係図、図6は定型部分を有している情報の頻度  $f$  と部分文字列の総数  $F(f)$  の関係図、図7(a)は本発明における実施の形態1における共通パターン発見装置の構成図、図7(b)は(a)の共通パターン発見装置のプログラム構成図、図8は取り出す部分文字列の採取パターンを示す説明図、図9は本発明の実施の形態1における共通パターン発見装置が行う処理のフローチャートである。

10

#### 【0032】

実施の形態1においては、テキスト情報の代表例としてウェブページ情報を対象として共通のパターンを有する情報を発見して抽出する。しかし、ウェブページ情報に限らず、電子化されたテキスト情報であれば、共通のパターンを発見できるものである。さらに画像情報や塩基配列情報等に対しても共通のパターンの発見が可能である。抽出する部分文字列の文字数は固定されず、1文字から最大文字数(利用者が任意に設定できる)、例えば30文字までの間で変化させて部分文字列として取り出し、異なる部分文字列の数とそれぞれの出現回数の和をとって頻度としてカウントする。なお、共通パターンとして30文字を越えた部分文字列が繰り返して出現する場合でも、共通パターン以外には30文字を超えた部分文字列が繰り返して出現する可能性はほとんどない。このため、30文字以上の共通パターンは30文字の部分文字列の和として表すことができる。文字列長さ  $n$  の部分文字列取り出しは、図8に示すような採取パターンで行われる。ファイル最初の「<html><head><title>ABC sports</title>」から10字ずつ文字列を切り出す採取パターンと5字ずつ切り出す採取パターンを示している。もちろん、採取できる箇所はここだけに限られないし、ここでは10字、5字のみを示しているが、上述したとおり文字列長さ( $n$ 個)は、 $n = 1, 2, 3, 4 \dots$ から選ばれ、10字、5字に限られるものではない。

20

#### 【0033】

まず、本発明の共通パターン発見方法の原理について説明する。本発明は、定型部分を有する情報の異なる部分文字列の数とその頻度の関係、定型部分を有していない情報の異なる部分文字列の数とその頻度の関係の間には、顕著な相違が存在することに着目し、この異なる部分文字列の数と頻度の関係に基づいて情報間に存在する共通パターンを発見するものである。

30

#### 【0034】

この相違を検討するため、定型部分を有していない情報の典型である夏目漱石の作品「こころ」と、定型部分を有する情報の典型としてA新聞社のHTMLの50個の記事情報を使って検討する。図1、図2は、「こころ」とA新聞社記事情報の2種類の情報において、部分文字列が出現する頻度  $f$ 、異なる部分文字列の数  $V(f)$ 、部分文字列長さ  $n$  を3次的にプロットしたものである。このとき、3次元だけでは分かり辛いので2次的に捉え直したものが図3、図4である。「こころ」に関して、同一頻度ごとにすべての長さの文字列を取り込んだときの、異なる部分文字列の数  $V(f)$  と頻度  $f$  との関係を示したのが図3であり、これは図13で説明したジップの第2法則そのものである。同様に、定型部分を有する記事情報について、同一頻度ごとにすべての長さの文字列を取り込んだときの、異なる部分文字列の数  $V(f)$  と頻度  $f$  との関係を示したのが図4であり、これときには図13で説明したジップの第2法則とはまったく異なればらで別の傾向を示している。従って、定型部分を有する情報は、いわゆる拡張されたジップの第2法則が成立しないことが分かる。しかし、この  $V(f)$  と  $f$  との関係だけでは、定型部分を有する情報を抽出することはできない。

40

#### 【0035】

50



しかし、本発明者らは、図4のような一見ばらばらの $V(f)$ と $f$ との関係であるが、部分文字列の総数 $F(f) = f \times V(f)$ を計算すると、定型部分の特徴が現れることを発見した。すなわち、図5は、定型部分を有していない「こころ」の場合の頻度 $f$ と総数 $F(f)$ の関係を示しているが、頻度 $f$ が200程度で最小となるなだらかな曲線状のラインを示す。これはジップの法則を書き換えただけで、特段の特徴のないものとなる。単純に頻度が多くなるほど文字列の表れる頻度が低下することを示す。

#### 【0036】

これに対して、定型部分を有する記事情報の場合は、図6に示すように頻度 $f$ が50のところにピンポイントの針状のピークが現れ、同様 $f = 100$ の位置でもピンポイントのピークが現れる。これは、部分文字列の文字数も変化しても、この50個、100個が常に共通であることを示している。図2に示す部分文字列長さ方向（以下、長さ方向という）で文字列長さが増しても、これらの部分文字列を包含した形で増していくため、頻度は変化しないことから分かる。言い換えれば、記事情報に共通のパターンが50個、100個存在することを示している。

10

#### 【0037】

このように本発明は、対象の全情報から任意の長さの部分文字列（最大長さ $N_{max}$ は10～30に設定）を切り出して、同一の部分文字列ごとに出現回数の和をとって頻度 $f$ とし、この異なる部分文字列の数 $V(f)$ を数えて、総数 $F(f) = f \times V(f)$ を計算することを特徴とする。この $F(f)$ と頻度 $f$ の関係を求め（関数関係を示すグラフを作成し）、ピンポイントのピークがあれば、頻度 $f$ の情報が共通パターンを有していると判断するものである。きわめて容易に共通パターンを有する情報を抽出することができる。

20

#### 【0038】

本発明の実施の形態1における共通パターン発見装置は、図7(a)(b)に示すように構成される。図7(a)において、1は中央演算処理装置（以下CPU、本発明のコンピュータ）等から構成されプログラムをロードして演算を行いシステム制御し各種機能を実行する共通パターン発見装置の中央演算/制御部、2は中央演算/制御部1が実行するプログラムを記憶した記憶媒体から構成される記憶部、3はキーボードやマウス等の入力手段、4はディスプレイ等に表示させる表示手段、5はプロトコルTCP/IP等でインターネット等のネットワークと接続するための通信制御部、6はネットワークとの通信管理を行うネットワークサーバ部である。

30

#### 【0039】

実施の形態1の共通パターン発見装置は、通信制御部5やネットワークサーバ部6を備えているため、ネットワークからHTML等のマークアップ言語で記述したウェブページをダウンロードすることができる。しかし、その他の情報を入力手段3から入力することもできる。

#### 【0040】

次に、共通パターン発見装置が、共通パターン発見方法を実行し、中央演算/制御部1にこれを実行させるプログラム、またプログラムを記録した記録媒体について説明する。以下説明する各機能手段は、いずれも中央演算/制御部1を構成するCPU（コンピュータ）にプログラムを記憶媒体から読み込んで機能させる手段である。図7(b)において、11はHTML等で記述されたウェブページ情報から任意の文字列長さ $n$ （ $n = 1, \dots$ ）で部分文字列を取り出す部分文字列取り出し手段、12は部分文字列取り出し手段11が抽出した部分文字列の出現回数をカウントして同一の部分文字列ごとに出現回数の和をとって頻度とする頻度カウント手段、13は部分文字列取り出し手段11が取り出した部分文字列について同一頻度ごとに異なる部分文字列の数 $V(f)$ をカウントする部分文字列種類数カウント手段、14は頻度カウント手段12がカウントした頻度 $f$ と異なる部分文字列の数 $V(f)$ の積を計算する総数計算手段、15は頻度 $f$ と部分文字列の総数 $F(f) = f \times V(f)$ の関係からピンポイントでピークが出現する位置の頻度を発見するピーク発見手段、16はピークの位置の頻度をカウントした部分文字列を含むウェブペー

40

50

ジ情報を抽出する情報抽出手段、17はピーク発見手段15が発見したピークが存在する頻度 $f$ に該当しないノイズ情報を除去するノイズ情報除去手段である。情報抽出手段16は、記憶部2に部分文字列取り出し手段11が取り出したすべての部分文字列のデータが記憶されているから、ピークを示した頻度の文字列情報に基づいて、これらの文字列情報を含むウェブページ情報を抽出する。

#### 【0041】

なお、実施の形態1の表通パターン発見装置はウェブページ情報等のテキスト情報を対象とするものであるが、上述したとおり塩基配列情報または画像情報を対象とすることもできる。この場合は、部分文字列を抽出する代わりに、部分塩基配列または部分画素列を抽出することになる。ただ、塩基配列の場合は、実施例4で説明するように4つの塩基を示すA, T, C, Gの4文字の文字を並べて文字列で表現されるため、事実上テキスト情報から部分文字列を抽出する場合と差はない。そして、これらを対象とする場合、部分文字列取り出し手段11は、それぞれ部分塩基配列情報または部分画素配列の取り出し手段となり、部分文字列種類数カウント手段13も、それぞれ部分塩基配列情報または部分画素配列の種類カウント手段となる。

#### 【0042】

部分文字列取り出し手段13は、文字列長さ $n$ が $n = 1$ から最大の $N_{max}$  (任意に設定)まで全情報のあらゆる部分から部分文字列を取り出す。取り出し方は図8に示すとおり行われる。同一の部分文字列ごとに出現回数がカウントされ、頻度カウント手段12がこの出現回数の総和を頻度 $f$ として計算する。同様に、部分文字列種類数カウント手段13が、部分文字列取り出し手段11が取り出した異なる部分文字列の数 $V(f)$ を同一頻度ごとにカウントする。この結果から、関数計算手段14が総数 $F(f) = f \times V(f)$ を計算し、頻度 $f$ と $F(f)$ の関係を基にピーク発見手段15がピンポイントで出現するピークの位置の頻度 $f$ を探し、このときの頻度 $f$ から共通パターンが $f$ 個存在すること把握するとともに、情報抽出手段16が該当する情報の部分文字列を色付けなどして表示する。ピークは $F(f)$ の値の変化率が所定の大きさ(閾値)以上の場合にだけ抽出するのが好適である。ノイズ情報除去手段17は共通パターンを有さない情報を分離するものである。

#### 【0043】

続いて、本発明の実施の形態1における共通パターン発見装置が行う処理について、図9のフローチャートに基づいて説明する。分析対象のウェブページをダウンロードしたり、テキストデータや画像データを入力手段3から入力し、部分文字列長さ $n$ の最大値 $N_{max}$ を設定し、 $n$ の初期値を $n = 1$ とする(step 1)。なお、部分文字列長さ $n$ に代え、ファイルの最大長を設定するのでもよいし、適当な長さを入力して設定することもできる。文字列の長さ $n$ の部分文字列を $n = 1$ を初期値として取り出す(step 2)。部分文字列長さ $n$ のすべての部分文字列の出現回数をカウントして、同一の部分文字列ごとに出現回数の和をとって頻度 $f$ とする(step 3)。

#### 【0044】

頻度 $f$ をカウントした後、部分文字列長さ $n$ が最大値 $N_{max}$ と一致したか否かをチェックし(step 4)、一致していない場合は、部分文字列長さ $n$ を $n = n + 1$ としてインクリメントして(step 5)、step 2に戻り、一致した場合には、すべての頻度 $f$ に対して異なる部分文字列の数 $V(f)$ をカウントする(step 6)。次いで、すべての頻度 $f$ に対して部分文字列の総数 $F(f) = f \times V(f)$ を計算する(step 7)。この結果から頻度 $f$ と総数 $F(f)$ の関係のグラフを作成する(step 8)。

#### 【0045】

step 8において、グラフにピンポイントのピークがあるか否かを探し(step 9)、ピンポイントのピークがある場合、共通のパターンの情報を得るため、ピークの位置で頻度をカウントした部分文字列を識別可能に表示してウェブページ情報を抽出する(step 10)、共通のパターンを有しない情報をノイズ情報として除去して(step 11)、終了する。step 9において、ピンポイントのピークがない場合は、共通のパタ

10

20

30

40

50

ーンの情報を含まないとして終了する ( s t e p 1 2 ) 。

【 0 0 4 6 】

このように実施の形態 1 の共通パターン発見装置と共通パターン発見方法は、任意の部分文字列を取り出して、同一の部分文字列ごとに出現回数の和をとって頻度  $f$  とするとともに該頻度  $f$  における異なる部分文字列の数  $V(f)$  をカウントし、総数  $F(f) = f \times V(f)$  を計算するだけで共通のパターンを有する情報を発見でき、短時間で共通パターンを発見することができる。

【 実施例 1 】

【 0 0 4 7 】

本発明の実施例 1 で検出した共通パターンについて説明する。実施例 1 は、A 新聞社と B 新聞社、C 新聞社の HTML の記事情報の母集団について、共通パターンの発見が行えるか否か検討したものである。A 新聞社の HTML の記事情報は 50 件、B 新聞社の HTML の記事情報は 104 件、C 新聞社の HTML の記事情報は 140 件である。図 10 は本発明の実施例 1 における 3 新聞社の記事情報の頻度  $f$  と部分文字列の総数  $F(f)$  の関係図である。

10

【 0 0 4 8 】

図 10 によれば、3 新聞社の記事情報 294 件に対して、頻度 49, 50 で  $F(f)$  が 80, 000 のピンポイントのピーク、頻度 103, 104 で  $F(f)$  が 130, 000 のピンポイントのピーク、頻度 140 で  $F(f)$  が 170, 000 のピンポイントのピークが出現している。これは、A, B, C 新聞社記事情報はそれぞれ別の定型のフォーマットを有しているからと考えられ、A 新聞社の 50 件が頻度 50 で共通のパターンを示し、B 新聞社の 104 件が頻度 104 で、C 新聞社の 140 件が頻度 140 で共通のパターンを示しているものである。なお、その他のピークの検討を行った結果、頻度 49 は B 新聞社と C 新聞社の独立のテンプレートで偶然に共通のパターンを示したものであり、頻度 103 も同様に B, C 新聞社の独立のテンプレートで偶然の共通のパターンを示したものであった。

20

【 0 0 4 9 】

これからも分かるように、定型部分が別の複数の情報源の情報を母集団にしたときでも、実施例 1 における共通パターン発見方法によれば、別々に分離することが可能になる。言い換えれば、パターンが異なれば、パターンごとに分離して取り出せる。

30

【 実施例 2 】

【 0 0 5 0 】

本発明の実施例 2 で検出した共通パターンについて説明する。実施例 1 は、D 大学内の複数サイトの 598 ファイルを母集団としたとき、サイトの中に共通パターンの発見が行えるか否か検討したものである。図 11 は本発明の実施例 2 における大学内サイトの頻度  $f$  と部分文字列の総数  $F(f)$  の関係図である。

【 0 0 5 1 】

図 11 によれば、頻度 61, 62 において  $F(f)$  が  $2 \times 10^6$  を示し、頻度 103, 110 において  $F(f)$  が  $1.2 \times 10^6$  を示している。この頻度 61, 62 で示した共通パターンが何か示しているか調査するため、D 大学内のサイトを確認したところ、D 大学内のホームページは大学の総合のトップページを上位階層とし、各学部や学科等の下位階層へのリンクをもつものであった。各学部や学科等は独立にサイトを構築するため、本来、通常共通のパターンやフォーマット、テンプレートは存在しないと予想される。しかし、D 大学のトップページから最大 3 階層リンクを辿り 598 個のファイルを収集し、実施の形態 1 の共通パターン発見方法により共通パターンを探したところ、62 のページが大学のトップページを基礎にして利用していたため、頻度 62 でピークを示したものであった。頻度 103, 110 においても同様であった。頻度 61 でピークを示したのは、1 サイトだけトップページが余分に編集されていたことによる。頻度 103, 110 にピークが出現したのは、編集されたトップページの中に 2 つの部分文字列を含むものがあったことを意味している。

40

50

## 【 0 0 5 2 】

このように、本発明の実施例 2 によれば、まったく関連付けの情報をもたない多数の未知の情報の中から共通パターンを有する情報を抽出でき、共通のパターンをチェックすれば、権限なく他人の情報を改ざんしたものの発見することが可能になる。

## 【 実施例 3 】

## 【 0 0 5 3 】

本発明の実施例 3 で検出した共通パターンについて説明する。実施例 3 は、インターネットの検索エンジンを使い適当な検索語を用いて検索したときの検索結果 4 6 ファイルの中に、共通パターンが発見できるか否か検討したものである。図 1 2 は本発明の実施例 2 における検索エンジンによる検索結果の頻度  $f$  と部分文字列の総数  $F(f)$  の関係図である。

10

## 【 0 0 5 4 】

図 1 2 によれば、頻度 4 6 において  $F(f)$  が  $3.8 \times 10^6$  を示し、頻度 9 1 において  $F(f)$  が  $1.4 \times 10^6$  を示し、頻度 9 1 3 において  $F(f)$  が  $1.0 \times 10^6$  を示している。頻度 4 6 でピークを示した検索結果は、4 6 ファイルが共通のフォーマットで表示されるため、同じ文字列が複数存在し、共通のパターンを有するものとして検出されたものである。頻度 9 1 においてピークを示したのも、共通のテンプレートが存在したことによる。また、頻度 9 1 3 でピークが出現したのは、この検索エンジンでは検索結果が 20 件ずつ表示するフォーマットを有しており、4 4 個のファイルに 20 個の同一文字列が存在し、残りの 2 個のファイルではこれが少なく、それぞれ 1 9 個、1 4 個の同一文字列が存在したためである。このように、 $M$  個のファイルに対し、1 ファイルに  $n$  個の同一文字列が含まれる場合には、頻度  $M \times n$  においてピークを示すことになる。

20

## 【 0 0 5 5 】

複数の検索エンジンの検索結果であっても、検索エンジンごとに情報をまとめて、他の検索エンジンの情報との間でパターンの変換が可能になるので、ウェブ上の情報を 1 つのデータベースのように利用することが可能になる。

## 【 実施例 4 】

## 【 0 0 5 6 】

本発明の実施例 4 は、遺伝子解析によって得られた塩基配列情報を対象として、複数の塩基配列情報から類似の塩基配列を抽出したものである。

30

## 【 0 0 5 7 】

最近の遺伝子解析により、遺伝子による遺伝の仕組みがかなり正確に解明されてきている。この遺伝子は共通の 4 つの塩基から成り立っており、この塩基の配列によって様々なタンパク質が作られ、各生物特有の生命活動が行われている。全ての生物に共通する 4 つの塩基とは、アデニン (A と表記される)、グアニン (G と表記される)、チミン (T と表記される)、シトシン (C と表記される) である。ところで、この A, T, C, G 4 つの塩基は互いに A と T、G と C がそれぞれ水素結合し易い性質をもち、DNA の二本鎖においては A T, G C で対をなして、相補関係を充たす二本鎖を形成して二重螺旋の構造をもつ。そして、このような遺伝子の塩基配列情報は、例えば A T C G G A . . . のような記述方法によって、A, T, C, G の 4 文字のテキスト表記による配列データとして記述される。

40

さて、実施例 4 の共通パターン発見装置と共通パターン発見方法は、このように記述された遺伝子の塩基配列データを解析対象とする。A, T, C, G の 4 文字で記述された塩基配列データから、所定の文字数 ( $n = 1, \dots$ ) の部分文字列を抽出し、同一の部分文字列ごとに出現回数の和をとって頻度  $f$  をカウントするとともに、頻度  $f$  に対して異なる部分文字列の数  $V(f)$  をカウントする。次いで頻度  $f$  と異なる部分文字列の数  $V(f)$  の積  $F(f)$  をとり、頻度  $f$  と  $F(f)$  の関係からピンポイントのピークが出現するか否かを検討する。ピークがある場合、共通のパターンの情報を得るため、ピークの位置で頻度をカウントした部分文字列を識別可能に表示して該当する情報を抽出する。SNPs のように数塩基しか異なる配列であれば、高精度で類似配列を抽出することができる。

50

## 【0058】

このように実施例4の共通パターン発見装置と共通パターン発見方法は、共通パターンを示す塩基配列情報を収集することができ、遺伝子工学に対してきわめて大きなツールを提供することができることになる。

## 【産業上の利用可能性】

## 【0059】

本発明の共通パターン発見装置と、そのプログラム、またそれを記録したコンピュータ読み取り可能な記録媒体、さらにその共通パターン発見方法は、情報の中に隠れている有用な法則を発見するために情報抽出を行うウェブマイニングに有効で、データベースの統合にも有力な手段となる。自然言語処理を頻度の利用でごく短時間に処理できる。DNA等の塩基配列情報の中からモチーフとなる共通の塩基配列をみつけることができ、アライメントの類似性も確認できる。遺伝子の重要な機能を司る部分は、遺伝情報解析により正例と負例の頻度分布の差から調査されるが、本発明によれば、正例のみで共通な塩基配列部分の抽出によりこれが可能になる。また、データ処理において、共通パターンを示す部分をキャッシュし、圧縮を行うことができ、効率的なデータ処理が可能になる。また、画像情報の中で共通の被写体を発見することにより、複数の画像の接合や、筆跡、指紋等の同一性判断を行うことができる。

## 【図面の簡単な説明】

## 【0060】

【図1】本発明の実施の形態1における定型部分を有していない情報の部分文字列が出現する頻度  $f$ 、異なる部分文字列の数  $V(f)$ 、部分文字列長さ  $n$  の3次元説明図 20

【図2】本発明の実施の形態1における定型部分を有している情報の部分文字列が出現する頻度  $f$ 、異なる部分文字列の数  $V(f)$ 、部分文字列長さ  $n$  の3次元説明図

【図3】本発明の実施の形態1における定型部分を有していない情報の異なる部分文字列の数  $V(f)$  と頻度  $f$  の2次元説明図

【図4】本発明の実施の形態1における定型部分を有している情報の異なる部分文字列の数  $V(f)$  と頻度  $f$  の2次元説明図

【図5】本発明の実施の形態1における定型部分を有していない情報の頻度  $f$  と部分文字列の総数  $F(f)$  の関係図

【図6】本発明の実施の形態1における定型部分を有している情報の頻度  $f$  と部分文字列の総数  $F(f)$  の関係図 30

【図7】(a)本発明における実施の形態1における共通パターン発見装置の構成図、(b)(a)の共通パターン発見装置のプログラム構成図

【図8】取り出す部分文字列の採取パターンを示す説明図

【図9】本発明の実施の形態1における共通パターン発見装置が行う処理のフローチャート

【図10】本発明の実施例1における3新聞社の記事情報の頻度  $f$  と部分文字列の総数  $F(f)$  の関係図

【図11】本発明の実施例2における大学内サイトの頻度  $f$  と部分文字列の総数  $F(f)$  の関係図 40

【図12】本発明の実施例2における検索エンジンによる検索結果の頻度  $f$  と部分文字列の総数  $F(f)$  の関係図

【図13】ジップの第2法則を示す説明図

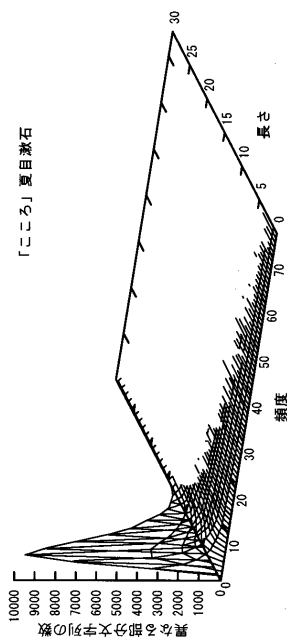
## 【符号の説明】

## 【0061】

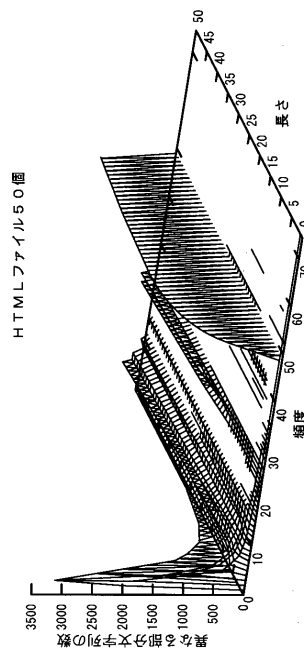
- 1 中央演算 / 制御部
- 2 記憶部
- 3 入力手段
- 4 表示手段
- 5 通信制御部

- 6 ネットワークサーバ部
- 1 1 部分文字列取り出し手段
- 1 2 頻度カウント手段
- 1 3 部分文字列種類数カウント手段
- 1 4 総数計算手段
- 1 5 ピーク発見手段
- 1 6 情報抽出手段
- 1 7 ノイズ情報除去手段

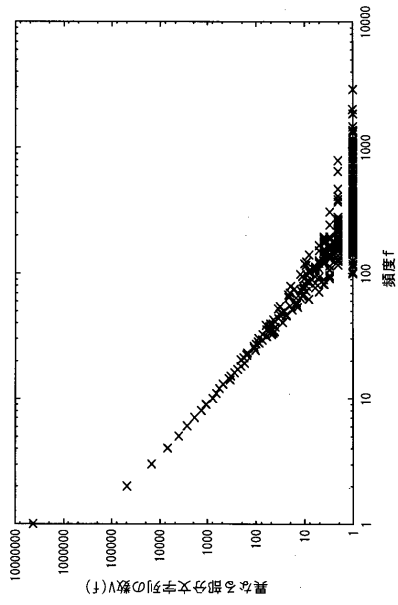
【図 1】



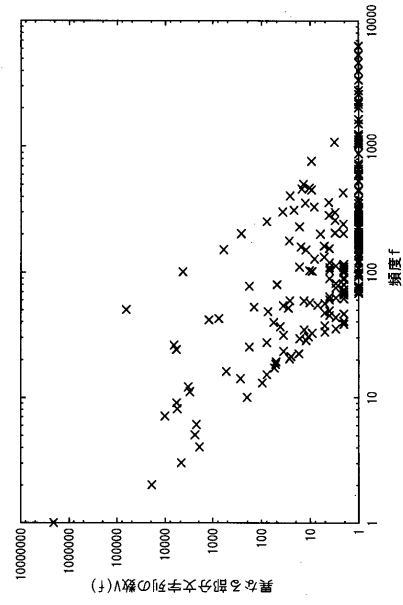
【図 2】



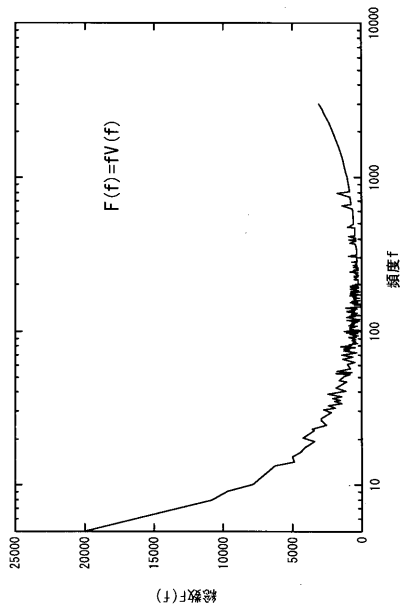
【 図 3 】



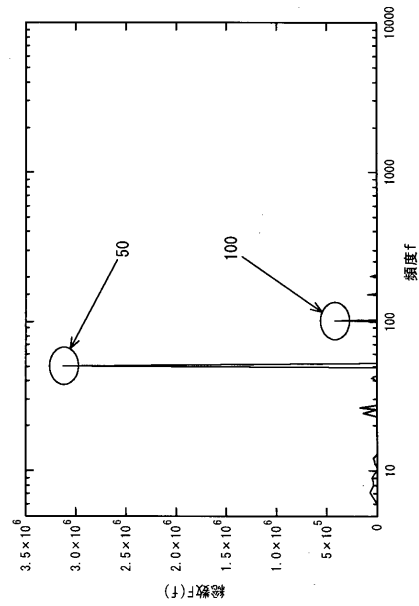
【 図 4 】



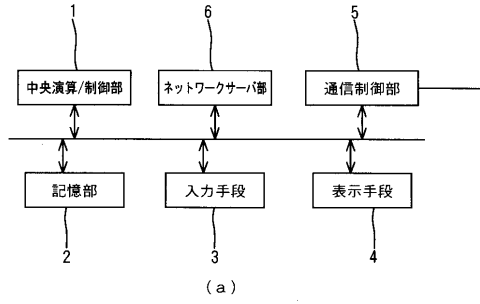
【 図 5 】



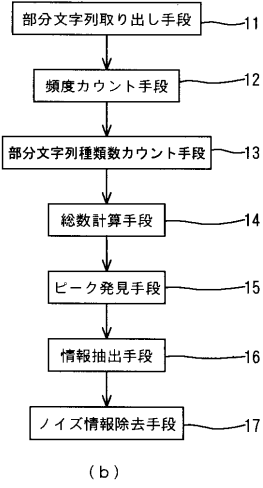
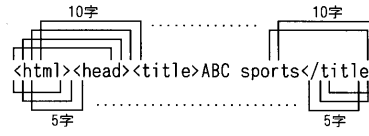
【 図 6 】



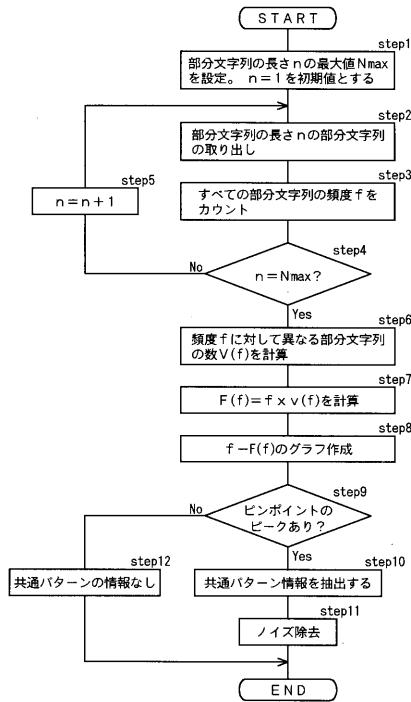
【 図 7 】



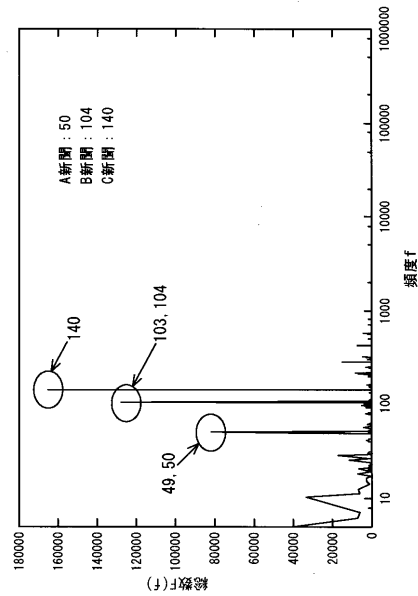
【 図 8 】



【 図 9 】

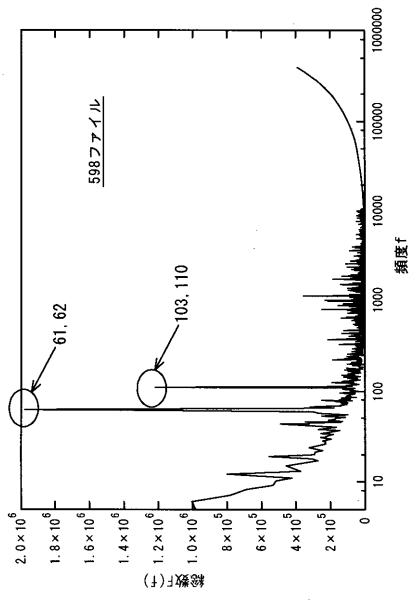


【 図 10 】

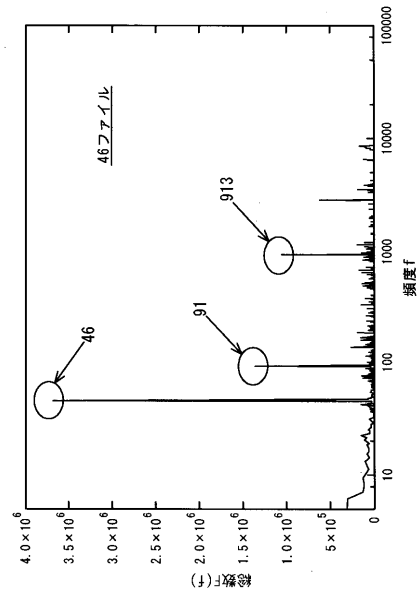




【図 1 1】



【図 1 2】



【図 1 3】

