(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第4385119号
(P4385119)

(45)発行日　平成21年12月16日(2009.12.16)　(24)登録日　平成21年10月9日(2009.10.9)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 220Z
G06F 19/00 (2006.01)　G06F 17/30 170A
G06F 19/00 130

請求項の数 4　(全 17 頁)

(21)出願番号　特願2003-315129 (P2003-315129)
(22)出願日　平成15年9月8日 (2003.9.8)
(65)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)
審査請求日　平成18年8月23日 (2006.8.23)

特許法第３０条第１項適用　ＦＩＴ２００３講演論文集（平成１５年８月２５日）第５９，６０頁に発表

(73)特許権者　504145342
国立大学法人九州大学
福岡県福岡市東区箱崎六丁目１０番１号
(74)代理人　100103621
弁理士　林　靖
(72)発明者　池田　大輔
福岡県糟屋郡志免町大字南里４３３－１ピアザ橘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

最終頁に続く

(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)【特許請求の範囲】
【請求項１】
電子化された複数又は単数のテキスト情報を対象としてこのテキスト情報の中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、変化率が閾値以上の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と同一頻度の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、
前記テキスト情報に同一の部分文字列が存在する場合に、この部分文字列の頻度の大きさに比例して前記積の値の大きさを増し、頻度に関してピークを形成する分布にして、このピークの位置の頻度を有する部分文字列を基に前記複数又は単数のテキスト情報間で共通する配列をもつ文字列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。
【請求項２】
コンピュータを、電子化された複数又は単数のテキスト情報を対象としてこのテキスト情報の中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、変化率が閾値以上の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と同一頻度の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、
前記テキスト情報に同一の部分文字列が存在する場合に、この部分文字列の頻度の大きさに比例して前記積の値の大きさを増し、頻度に関してピークを形成する分布にして、前記情報抽出手段によって抽出された該ピークの位置の頻度を有する部分文字列を基に前記複数又は単数のテキスト情報間で共通する配列をもつ文字列情報を発見することを特徴とするプログラム。

【請求項３】

請求項２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４】

電子化された複数又は単数のテキスト情報を対象としてこのテキスト情報の中から部分文字列取り出し手段によって最大長さまでのすべての長さの部分文字列を抽出し、頻度カウント手段によっ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とともに該頻度を有する異なる部分文字列の数を部分文字列種類数カウント手段によってカウントし、総数計算手段によって前記頻度と前記異なる部分文字列の数との積を計算し、更にピーク発見手段によって前記積と前記頻度との関係から変化率が閾値以上の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情報抽出手段によって該ピークの位置の頻度と同一頻度の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方法であって、
前記テキスト情報に同一の部分文字列が存在する場合に、この部分文字列の頻度の大きさに比例して前記積の値の大きさを増し、頻度に関してピークを形成する分布にして、前記情報抽出手段によって抽出された該ピークの位置の頻度を有する部分文字列を基に前記複数又は単数のテキスト情報間で共通する配列をもつ文字列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。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、テキスト情報の中で共通する配列の文字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。

【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）が繰り返し表示される。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。

【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）。

【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。

【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。

【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×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、logＶ（ｆ）＝－ａ（logｆ）＋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。

【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
【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。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、複数又は単数のテキスト情報間で共通する配列の文字列情報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、複数又は単数のテキスト情報間で共通する配列の文字列情報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、複数又は単数のテキスト情報間で共通する配列の文字列情報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、複数又は単数のテキスト情報間で共通する配列の文字列情報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、電子化された複数又は単数のテキスト情報を対象としてこのテキスト情報の中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、変化率が閾値以上の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と同一頻度の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、テキスト情報に同一の部分文字列が存在する場合に、この部分文字列の頻度の大きさに比例して積の値の大きさを増し、頻度に関してピークを形成する分布にして、このピークの位置の頻度を有する部分文字列を基に複数又は単数のテキスト情報間で共通する配列をもつ文字列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、すべての長さの異なる部分文字列の数にその部分文字列の頻度を掛けることによりこの頻度に関して針状のピークを形成する分布とすることができ、このピークが出現する位置を探すことにより複数又は単数のテキスト情報間で共通する配列の文字列情報を抽出できる。また、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、テキスト情報の中で共通する配列の文字列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、電子化された複数又は単数のテキスト情報を対象としてこのテキスト情報の中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、変化率が閾値以上の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と同一頻度の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、テキスト情報に同一の部分文字列が存在する場合に、この部分文字列の頻度の大きさに比例して積の値の大きさを増し、頻度に関してピークを形成する分布にして、このピークの位置の頻度を有する部分文字列を基に複数又は単数のテキスト情報間で共通する配列をもつ文字列情報を発見する共通パターン発見装置である。テキスト情報においては、異なった部分文字列の数とその出現頻度の対応関係に規則性がある場合（ジップの第２法則に従う場合）と、この対応関係に規則性がない場合とが存在するが、その後者の中で共通パターンがある場合は、すべての長さの異なる部分文字列の数にその部分文字列の頻度を掛けることにより、部分文字列の頻度の大きさに比例して積の値の大きさを増し、後者の場合に頻度に関して針状のピークを形成する分布にする。このピークが出現する位置を探すことにより複数又は単数のテキスト情報間で共通する配列の文字列情報を抽出できる。また、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、テキスト情報の中で共通の配列を有する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。テキスト表記を利用することにより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。

【００２３】

本発明を実施するための第２の形態は、コンピュータを、電子化された複数又は単数のテキスト情報を対象としてこのテキスト情報の中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、変化率が閾値以上の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

クの位置の頻度と同一頻度の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、テキスト情報に同一の部分文字列が存在する場合に、この部分文字列の頻度の大きさに比例して積の値の大きさを増し、頻度に関してピークを形成する分布にして、情報抽出手段によって抽出された該ピークの位置の頻度を有する部分文字列を基に複数又は単数のテキスト情報間で共通する配列をもつ文字列情報を発見することを特徴とするプログラムである。テキスト情報においては、異なった部分文字列の数とその出現頻度の対応関係に規則性がある場合（ジップの第２法則に従う場合）と、この対応関係に規則性がない場合とが存在するが、その後者の中で共通パターンがある場合は、すべての長さの異なる部分文字列の数にその部分文字列の頻度を掛けることにより、部分文字列の頻度の大きさに比例して積の値の大きさを増し、後者の場合に頻度に関して針状のピークを形成する分布にする。このピークが出現する位置を探すことにより複数又は単数のテキスト情報間で共通する配列の文字列情報を抽出できる。また、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、テキスト情報の中で共通する配列の文字列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。テキスト表記を利用することにより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。

【００２７】

本発明を実施するための第３の形態は、第２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４の形態は、電子化された複数又は単数のテキスト情報を対象としてこのテキスト情報の中から部分文字列取り出し手段によって最大長さまでのすべての長さの部分文字列を抽出し、頻度カウント手段によっ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とともに該頻度を有する異なる部分文字列の数を部分文字列種類数カウント手段によってカウントし、総数計算手段によって頻度と異なる部分文字列の数との積を計算し、更にピーク発見手段によって積と頻度との関係から変化率が閾値以上の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情報抽出手段によって該ピークの位置の頻度と同一頻度の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方法であって、テキスト情報に同一の部分文字列が存在する場合に、この部分文字列の頻度の大きさに比例して積の値の大きさを増し、頻度に関してピークを形成する分布にして、情報抽出手段によって抽出された該ピークの位置の頻度を有する部分文字列を基に複数又は単数のテキスト情報間で共通する配列をもつ文字列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である。テキスト情報においては、異なった部分文字列の数とその出現頻度の対応関係に規則性がある場合（ジップの第２法則に従う場合）と、この対応関係に規則性がない場合とが存在するが、その後者の中で、共通パターンがある場合は、すべての長さの異なる部分文字列の数にその部分文字列の頻度を掛けることにより、部分文字列の頻度の大きさに比例して積の値の大きさを増し、後者の場合に頻度に関して針状のピークを形成する分布にする。このピークが出現する位置を探すことにより複数又は単数のテキスト情報間で共通する配列の文字列情報を抽出できる。また、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、テキスト情報の中で共通する配列の文字列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。テキスト表記を利用することにより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。

【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。

【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。ここで共通のパターンとはテキスト情報の中で共通する配列をもつ文字列情報のことであり、以下、共通のパターン、共通パターンなどともいう。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・・から選ばれ、１０字，５字に限られるものではない。

【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。

【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次元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ため２次元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２次元で示したのが図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２次元で示したのが図４であり、これは図１３で説明したジップの第２法則とはまったく異なったばら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、異なる部分文字列の数Ｖ（ｆ）とその頻度ｆの対応関係が不規則な関係になり、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。

【００３５】

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×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が変化しても、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。

【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×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。すべての長さの異なる部分文字列の数Ｖ（ｆ）にその頻度ｆをそれぞれ掛けて頻度ｆに比例させ、頻度ｆとの関係においてピークが出現する分布とする。

【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はディスプレー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。

【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，・・）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×Ｖ（ｆ）の関係からピンポイントで変化率が閾値以上の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×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(ｓｔｅｐ１)。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(ｓｔｅｐ２)。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（ｓｔｅｐ３)。

【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(ｓｔｅｐ４)、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(ｓｔｅｐ５)、ｓｔｅｐ２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(ｓｔｅｐ６)。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×Ｖ（ｆ）を計算する(ｓｔｅｐ７)。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（ｓｔｅｐ８）。

【００４５】

ｓｔｅｐ８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（ｓｔｅｐ９）、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し（ｓｔｅｐ１０）、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（ｓｔｅｐ１１）、終了する。ｓｔｅｐ９において、ピンポイントのピークがない場合は、共通のパターンの情報を含まないとして終了する（ｓｔｅｐ１２）。

【００４６】

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×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】

【００４７】

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。

【００４８】

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。

【００４９】

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。

【実施例２】

【００５０】

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。

【００５１】

図１１によれば、頻度６１，６２においてＦ（ｆ）が2×10^{6}を示し、頻度１０３，１１０においてＦ（ｆ）が1.2×10^{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。

【００５２】

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発見することが可能になる。

【実施例３】

【００５３】

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４６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１２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。

【００５４】

図１２によれば、頻度４６においてＦ（ｆ）が３．８×１０^{6}を示し、頻度９１においてＦ（ｆ）が１．４×１０^{6}を示し、頻度９１３においてＦ（ｆ）が１．０×１０^{6}を示している。頻度４６でピークを示した検索結果は、４６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９１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９１３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、４４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１９個、１４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×ｎにおいてピークを示すことになる。

【００５５】

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】

【００５６】

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。

【００５７】

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，Ｇ４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ＡＴ，ＧＣ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線の構造をもつ。そして、このような遺伝子の塩基配列情報は、例えばＡＴＣＧＧＡ・・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。

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，・・）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らない配列であれば、高精度で類似配列を抽出することができる。

【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、ウェブページ情報などの電子化されたテキスト情報の処理を頻度の利用によってごく短時間に処理できる。テキスト表記を利用することにより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、アライン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、共通の配列を有する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

【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

【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

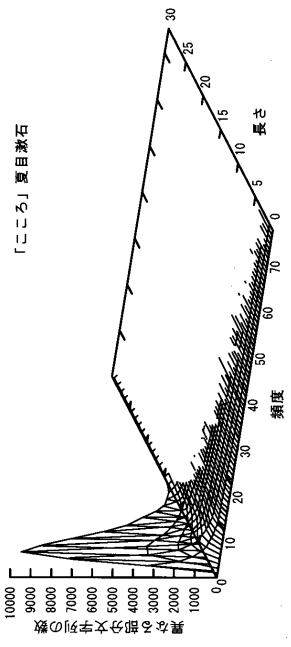
【図８】取り出す部分文字列の採取パターンを示す説明図

【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

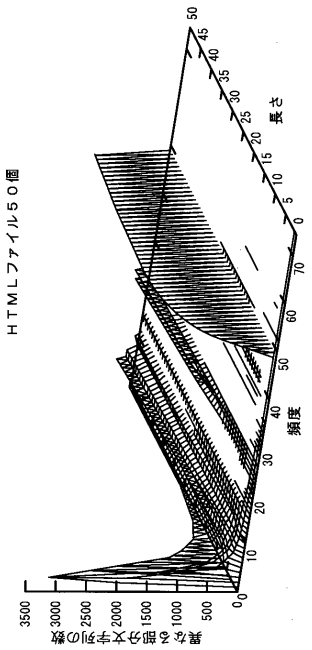
【図１０】本発明の実施例１における３新聞社の記事情報の頻度ｆと部分文字列の総数Ｆ

（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
【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
１１　部分文字列取り出し手段
１２　頻度カウント手段
１３　部分文字列種類数カウント手段
１４　総数計算手段
１５　ピーク発見手段
１６　情報抽出手段
１７　ノイズ情報除去手段

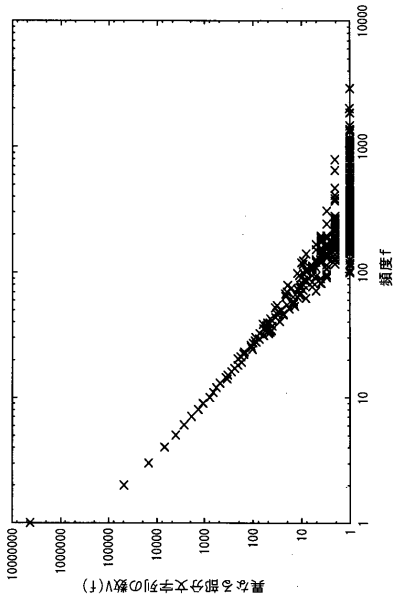
【図１】



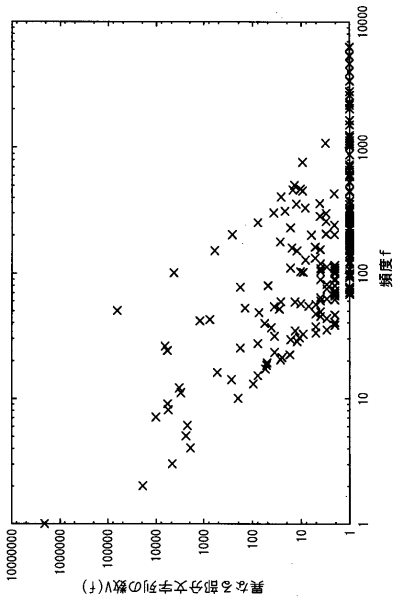
【図２】



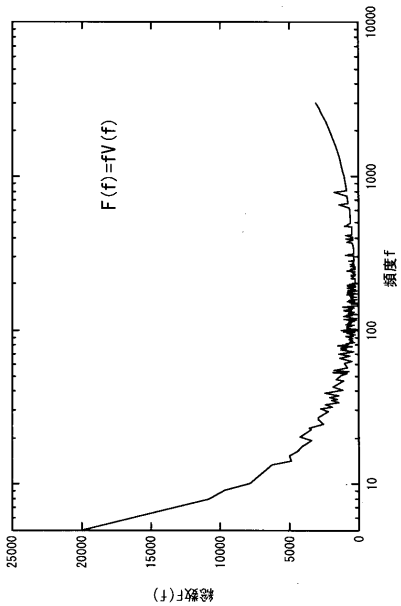
【図３】


異なる部分文字列の数V(f)
10000000
1000000
100000
10000
1000
100
10
1
1
10
100
1000
10000
頻度f

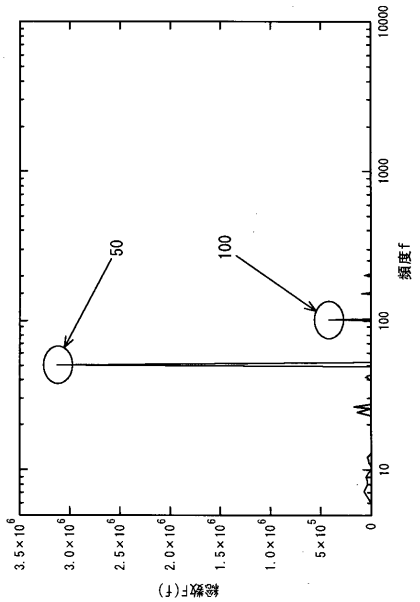
【図４】


異なる部分文字列の数V(f)
10000000
1000000
100000
10000
1000
100
10
1
1
10
100
1000
10000
頻度f

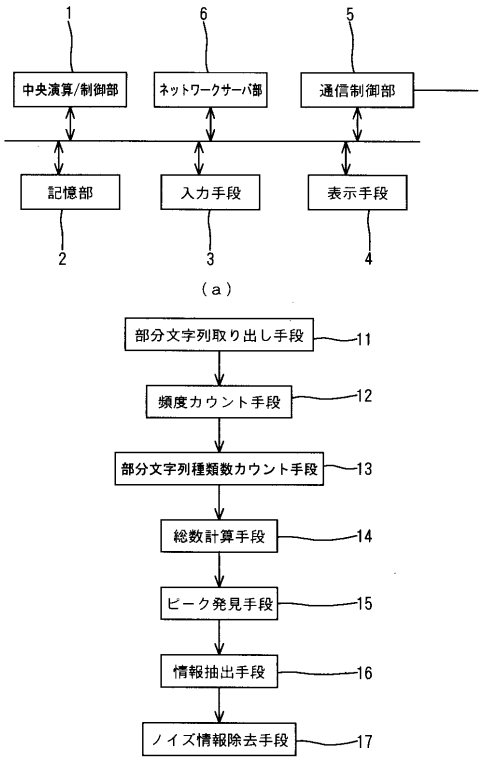
【図５】


総数F(f)
25000
20000
15000
10000
5000
0
F(f)=fV(f)
1
10
100
1000
10000
頻度f

【図６】


総数F(f)
3.5×10^6
3.0×10^6
2.5×10^6
2.0×10^6
1.5×10^6
1.0×10^6
5×10^5
0
50
100
1
10
100
1000
10000
頻度f

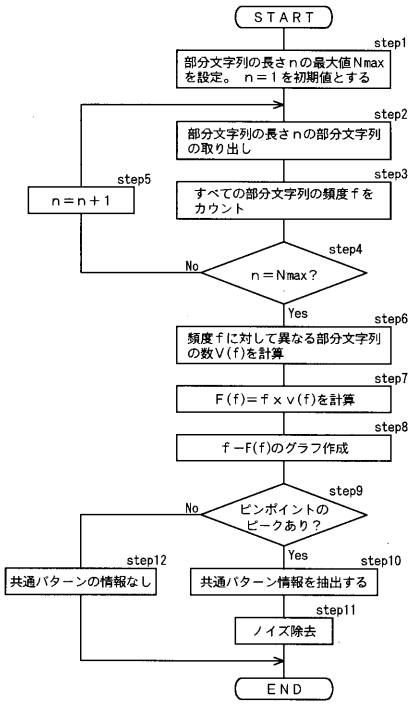
【図７】


1
6
5
中央演算/制御部
ネットワークサーバ部
通信制御部
記憶部
入力手段
表示手段
2
3
4
（ａ）
部分文字列取り出し手段
11
頻度カウント手段
12
部分文字列種類数カウント手段
13
総数計算手段
14
ピーク発見手段
15
情報抽出手段
16
ノイズ情報除去手段
17
（ｂ）

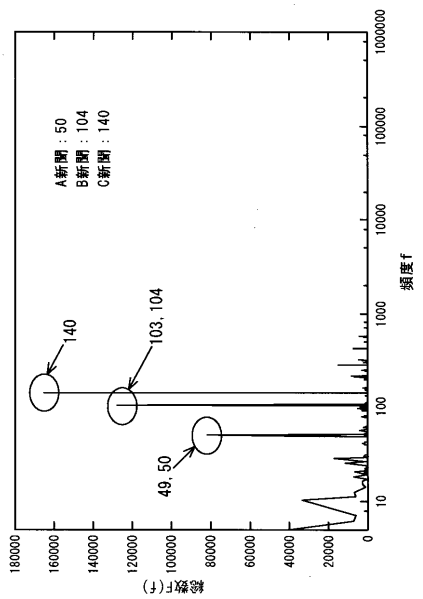
【図８】


10字
10字
<html><head><title>ABC sports</title>
5字
5字

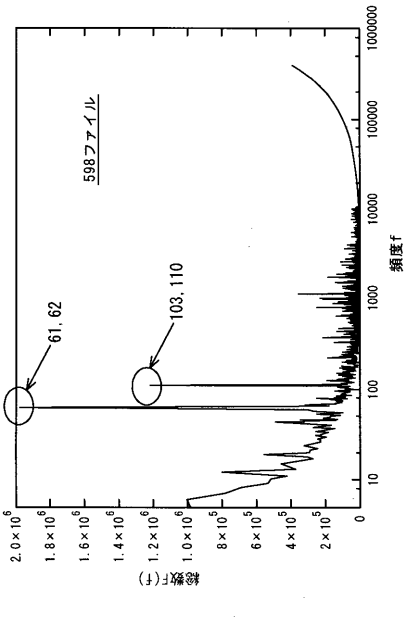
【図９】


START
step1
部分文字列の長さｎの最大値Nmaxを設定。 ｎ＝１を初期値とする
step2
部分文字列の長さｎの部分文字列の取り出し
step3
すべての部分文字列の頻度ｆをカウント
step4
No
ｎ＝Nmax？
Yes
step5
ｎ＝ｎ＋１
step6
頻度ｆに対して異なる部分文字列の数V(f)を計算
step7
F(f)＝ｆｘｖ(f)を計算
step8
ｆ－F(f)のグラフ作成
step9
No
ピンポイントのピークあり？
Yes
step10
共通パターン情報を抽出する
step11
ノイズ除去
step12
共通パターンの情報なし
END

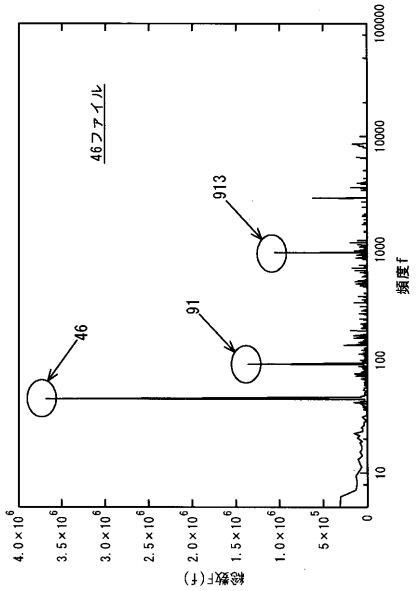
【図１０】


A新聞：50
B新聞：104
C新聞：140
140
103,104
49,50
1000000
100000
10000
1000
100
10
0
頻度f
180000
160000
140000
120000
100000
80000
60000
40000
20000
総数F(f)

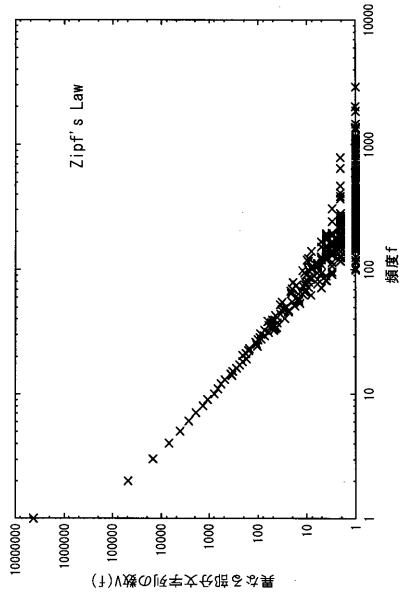
【図１１】


598ファイル
61.62
103.110
総数F(f)
頻度f

【図１２】


46ファイル
46
91
913
総数F(f)
頻度f

【図１３】


Zipf's Law
異なる部分文字列の数V(f)
頻度f

フロントページの続き

審査官　鈴木　和樹

(56)参考文献　相澤彰子，低頻度語の利用によるテキスト分類性能の改善と評価，情報処理学会論文誌，日本，社団法人情報処理学会，２００３年　7月１５日，第４４巻，第7号，ｐ．１７２０－１７３０
風間一洋、外２名，Ｉｎｇｒｉｄ　ＮｅｗｓＣａｓｔ－自立型ニュース配信システム，電子情報通信学会技術研究報告（ＣＰＳＹ９７－５１～６３），日本，社団法人電子情報通信学会，１９９７年　8月２０日，第９７巻，第２２６号，ｐ．１７－２４
池田大輔、外２名，部分文字列増幅法による共通パタン発見アルゴリズム，情報処理学会研究報告，日本，社団法人情報処理学会，２００３年１２月１２日，第２００３巻，第１２２号，ｐ．４５－４８
池田大輔、外２名，文字列の頻度分布による共通パタン発見，情報処理学会研究報告（２００３－ＦＩ－７２），日本，社団法人情報処理学会，２００３年　9月３０日，第２００３巻，第９８号，ｐ．２５－３２

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／３０
Ｇ０６Ｆ　１９／００
ＪＳＴＰｌｕｓ（ＪＤｒｅａｍＩＩ）