(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2006-350950
(P2006-350950A)
(43)公開日　平成18年12月28日(2006.12.28)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 17/21	(2006.01)	G06F 17/21	530P	5B009
G06F 3/048	(2006.01)	G06F 17/21	564P	5B075
G06F 17/30	(2006.01)	G06F 3/00	654D	5E501
		G06F 17/30	170A	
		G06F 17/30	330C	

審査請求 未請求　請求項の数 14　ＯＬ　（全 20 頁）

(21)出願番号　特願2005-179703 (P2005-179703)
(22)出願日　平成17年6月20日 (2005.6.20)

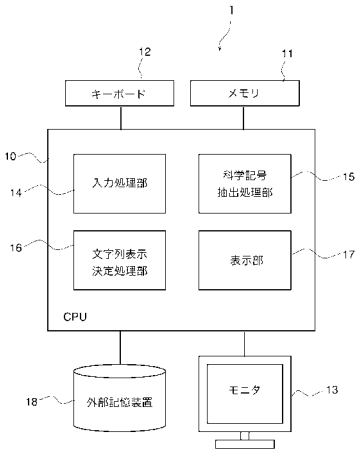
(71)出願人　301022471
独立行政法人情報通信研究機構
東京都小金井市貫井北町４－２－１
(74)代理人　100130111
弁理士　新保　斉
(72)発明者　村田　真樹
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　白土　保
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　井佐原　均
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内

Ｆターム(参考)　5B009 NB07　QA03　RB32　VA02

最終頁に続く

(54)【発明の名称】データ表示装置及びデータ表示方法

(57)【要約】
【課題】　第１データから第２データへのデータ変換及び、その逆方向のデータ変換が可能なデータ変換時に、変換適性を自動的に評価する技術を提供すること。
【解決手段】　科学記号を含むテキストデータを表示するデータ表示装置１において、テキストデータの入力処理部１４と、科学記号データベース１８と、データベースを参照してテキストデータから科学記号候補文字列を抽出する科学記号抽出処理部１５と、発現条件と照合して科学記号候補文字列の表示態様を決定する文字列表示決定処理部１６と、決定された表示態様の科学記号候補文字列を含むテキストデータを表示する表示部１７・１３とを備える。
【選択図】　図１



【特許請求の範囲】
【請求項１】
少なくとも自然科学で用いられる記号又は式（以下、科学記号と呼ぶ）を含むテキストデータを表示するデータ表示装置であって、
テキストデータを入力する入力処理部と、
科学記号として用いる１文字以上の文字列と当該文字列が発現する発現条件とを予め格納した科学記号データベースと、
該科学記号データベースを参照して該テキストデータから科学記号候補文字列を抽出する科学記号抽出処理部と、
該発現条件と照合して該科学記号候補文字列の表示態様を決定する文字列表示決定処理部と、
決定された表示態様の科学記号候補文字列を含むテキストデータを表示する表示部と
を備えることを特徴とするデータ表示装置。
【請求項２】
前記文字列表示決定処理部において、
テキストデータの基本文字色を予め設定すると共に、
該科学記号候補文字列の表示色を該テキストデータの表示に用いる基本文字色から変化させる処理を行う
請求項１に記載のデータ表示装置。
【請求項３】
前記文字列表示決定処理部において、
前記科学記号候補文字列が科学記号である確度を算出する科学記号確度算出部を備え、
該算出結果に基づいて確度が閾値よりも高い科学記号候補文字列については基本文字色と色相の異なる表示色を設定すると共に、確度が閾値よりも低い科学記号候補文字列については基本文字色と彩度又は明度が異なる表示色を設定する
請求項２に記載のデータ表示装置。
【請求項４】
前記データ表示装置に形態素解析処理部を備え、
前記テキストデータを形態素解析処理すると共に、
科学記号確度算出部において、
前記科学記号候補文字列の前後所定個数の形態素の少なくとも文字列情報又は文法情報のいずれかを用い、前記科学記号データベースに備えた少なくとも文字列情報又は文法情報のいずれかを参照して確度を算出する
請求項３に記載のデータ表示装置。
【請求項５】
前記科学記号確度算出部において、
前記科学記号候補文字列の前後に同一又は異なる科学記号候補文字列が連続して出現した場合に当該科学記号候補文字列の確度を所定値だけ高める処理を含む
請求項３又は４に記載のデータ表示装置。
【請求項６】
前記科学記号抽出処理部が、
予め定めた科学記号を構成する特定表現を抽出し、
前記科学記号確度算出部において、
該特定表現が前後所定個数の形態素内、又は同一文、又は同一テキストデータ中に出現した場合に、当該科学記号候補文字列の確度を所定値だけ高める処理を含む
請求項３ないし５のいずれかに記載のデータ表示装置。
【請求項７】
前記特定表現を、特定表現データベースに格納する構成において、
着目している科学記号候補文字列と共に、テキストデータの同一文又は所定個数の形態素内に特定表現候補が出現する回数Ｎ１を計数する一方、該特定表現候補がその他の文に

おいて単独に出現する回数Ｎ２を計数し、Ｎ１／Ｎ２（Ｎ２≠０）又はＮ１／（Ｎ１＋Ｎ２）（Ｎ２が０のときも含む）の少なくともいずれかの値が閾値以上の場合に、該特定表現データベースに格納する処理を含む
請求項３ないし６のいずれかに記載のデータ表示装置。
【請求項８】
少なくとも自然科学で用いられる記号又は式（以下、科学記号と呼ぶ）を含むテキストデータを表示するデータ表示装置におけるデータ表示方法であって、
入力処理部がテキストデータを入力する入力ステップ、
科学記号として用いる１文字以上の文字列と当該文字列が発現する発現条件とを予め格納した科学記号データベースを参照し、科学記号抽出処理部が該テキストデータから科学記号候補文字列を抽出する科学記号抽出ステップ、
該発現条件と照合して文字列表示決定処理部が該科学記号候補文字列の表示態様を決定する文字列表示決定ステップ、
表示部が決定された表示態様の科学記号候補文字列を含むテキストデータを表示する表示ステップ
を含むことを特徴とするデータ表示方法。
【請求項９】
前記文字列表示決定ステップにおいて、
テキストデータの基本文字色を予め設定すると共に、
該科学記号候補文字列の表示色を該テキストデータの表示に用いる基本文字色から変化させる処理を行う
請求項８に記載のデータ表示方法。
【請求項１０】
前記文字列表示決定ステップにおいて、
科学記号確度算出部が、科学記号候補文字列が科学記号である確度を算出する科学記号確度算出ステップを行った後に、
該算出結果に基づいて確度が閾値よりも高い科学記号候補文字列については基本文字色と色相の異なる表示色を設定すると共に、確度が閾値よりも低い科学記号候補文字列については基本文字色と彩度又は明度が異なる表示色を設定する
請求項９に記載のデータ表示方法。
【請求項１１】
前記データ表示方法において、
形態素解析処理部が前記テキストデータを形態素解析する形態素解析ステップを前記科学記号抽出ステップの前に実行し、
科学記号確度算出ステップにおいて、該科学記号候補文字列の前後所定個数の形態素の少なくとも文字列情報又は文法情報のいずれかを用い、前記科学記号データベースに備えた少なくとも文字列情報又は文法情報のいずれかを参照して確度を算出する
請求項１０に記載のデータ表示方法。
【請求項１２】
前記科学記号確度算出ステップにおいて、
前記科学記号候補文字列の前後に同一又は異なる科学記号候補文字列が連続して出現した場合に当該科学記号候補文字列の確度を所定値だけ高める処理を含む
請求項１０又は１１に記載のデータ表示方法。
【請求項１３】
前記科学記号抽出ステップにおいて、
予め定めた科学記号を構成する特定表現を抽出し、
前記科学記号確度算出ステップにおいて、
該特定表現が前後所定個数の形態素内、又は同一文、又は同一テキストデータ中に出現した場合に、当該科学記号候補文字列の確度を所定値だけ高める処理を含む
請求項１０ないし１２のいずれかに記載のデータ表示方法。

【請求項１４】
前記特定表現を、特定表現データベースに格納する構成において、
着目している科学記号候補文字列と共に、テキストデータの同一文又は所定個数の形態素内に特定表現候補が出現する回数Ｎ１を計数する一方、該特定表現候補がその他の文において単独に出現する回数Ｎ２を計数し、Ｎ１／Ｎ２（Ｎ２≠０）又はＮ１／（Ｎ１＋Ｎ２）（Ｎ２が０のときも含む）の少なくともいずれかの値が閾値以上の場合に、該特定表現データベースに格納する処理を含む
請求項１０ないし１３のいずれかに記載のデータ表示方法。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明はコンピュータにおけるテキストデータの表示装置及び方法に関し、特に科学記号を視認しやすく表示する技術に係るものである。
【背景技術】
【０００２】
自然科学論文には多くの数式や記号が記述されており、それらが論文の内容を端的に表現していることが多い。従って研究者は論文集など多数の論文から所望のトピックの論文を抽出する際に、数式や記号などを概観して選び出す作業を行うことがある。
近年では学会において発行される論文誌は従来の紙媒体からＣＤ－ＲＯＭやインターネットに変わりつつあり、論文をパーソナルコンピュータ上で閲覧する機会も多い。
【０００３】
このような時に、論文を構成するテキストから数式や記号を迅速に識別することができれば効率がよい。しかし、特に英語などのラテン文字を用いる論文では、同じくラテン文字で記載されることの多い数式や記号がテキスト中に埋没してしまい、詳細に閲読しなければならなかったり、肝心の数式や記号を見落とす恐れがあった。
【０００４】
従来からワードプロセッサにおいて文字種別に応じて表示色を変えることは行われている。例えばひらがな及び漢字は黒色、カタカナは緑色、半角英数字は茶色などのように区別して表示する製品が知られている。これは特に日本語と英語等では半角と全角の区別や長音とハイフンの区別が不明確になりやすいために、入力者に分かりやすく表示するものである。
【０００５】
この方法は日本語論文中に半角の英数字が含まれている場合には、数式や記号をある程度見やすくすることには寄与するが、上述したようにすべてラテン文字で記述された論文の場合には全て同色となってしまうため判別しやすくならない。
【０００６】
また、特許文献１には化学式の中から任意の化学物質について、その化学物質が有する様々な特徴を容易に表示する技術が開示されている。すなわち、元素記号によって色を変えると共に、固体や気体などの場合には斜体や太字にするなどの書式を変化させることが記載されている。
【０００７】
【特許文献１】特開平10-240748号公報
【０００８】
本技術ではテーブルデータに単に元素記号の文字列を備えて一致した文字列の色を変化させるだけであるため、偶然に元素記号等と一致した文字列がテキスト中に存在すれば誤って色を変化させることになり、誤解を生じさせたり、かえって読みにくくなる結果を招きやすい。特に、ラテン文字を用いたテキスト中ではその誤りが頻出する問題がある。
【発明の開示】
【発明が解決しようとする課題】

【０００９】
本発明は、上記従来技術の有する問題点に鑑みて創出されたものであり、文献中の数式や記号を読者が識別容易に表示すると共に、特に読者が誤解を生じることなく必要な数式や記号を読み取ることのできる表示装置及び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】
【００１０】
本発明は、上記の課題を解決するために、次のようなデータ表示装置を提供する。
すなわち本発明は、少なくとも自然科学で用いられる記号又は式（以下、科学記号と呼ぶ）を含むテキストデータを表示するデータ表示装置を提供する。
請求項１に記載の発明によれば、本データ表示装置に、テキストデータを入力する入力処理部と、科学記号として用いる１文字以上の文字列と当該文字列が発現する発現条件とを予め格納した科学記号データベースとを備える。発現条件の採録により従来技術のような単純な表示ではなく、高精度に表示態様を変化させることができる。
【００１１】
本発明の装置にはさらに、科学記号データベースを参照して該テキストデータから科学記号候補文字列を抽出する科学記号抽出処理部と、発現条件と照合して該科学記号候補文字列の表示態様を決定する文字列表示決定処理部と、決定された表示態様の科学記号候補文字列を含むテキストデータを表示する表示部とを備える。
【００１２】
請求項２に記載の発明では、文字列表示決定処理部において、テキストデータの基本文字色を予め設定すると共に、該科学記号候補文字列の表示色を該テキストデータの表示に用いる基本文字色から変化させる処理を行うことを特徴とする。
【００１３】
請求項３に記載の発明は上記の文字列表示決定処理部において、科学記号候補文字列が科学記号である確度を算出する科学記号確度算出部を備える。
そして、該算出結果に基づいて確度が閾値よりも高い科学記号候補文字列については基本文字色と色相の異なる表示色を設定すると共に、確度が閾値よりも低い科学記号候補文字列については基本文字色と彩度又は明度が異なる表示色を設定することを特徴とする。
【００１４】
請求項４に記載の発明はデータ表示装置に形態素解析処理部を備え、テキストデータを形態素解析処理する構成である。そして科学記号確度算出部において、科学記号候補文字列の前後所定個数の形態素の少なくとも文字列情報又は文法情報のいずれかを用い、科学記号データベースに備えた少なくとも文字列情報又は文法情報のいずれかを参照して確度を算出するものである。
【００１５】
請求項５に記載の発明は、科学記号確度算出部において、科学記号候補文字列の前後に同一又は異なる科学記号候補文字列が連続して出現した場合に当該科学記号候補文字列の確度を所定値だけ高める処理を含むことを特徴とする。
【００１６】
請求項６に記載の発明は、上記の科学記号抽出処理部が、予め定めた科学記号を構成する特定表現を抽出し、科学記号確度算出部において、該特定表現が前後所定個数の形態素内、又は同一文、又は同一テキストデータ中に出現した場合に、当該科学記号候補文字列の確度を所定値だけ高める処理を含むことを特徴とする。
【００１７】
請求項７に記載の発明は、前記特定表現を、特定表現データベースに格納する構成において、着目している科学記号候補文字列と共に、テキストデータの同一文又は所定個数の形態素内に特定表現候補が出現する回数Ｎ１を計数する一方、該特定表現候補がその他の文において単独に出現する回数Ｎ２を計数し、Ｎ１／Ｎ２（Ｎ２≠０）又はＮ１／（Ｎ１＋Ｎ２）（Ｎ２が０のときも含む）の少なくともいずれかの値が閾値以上の場合に、該特定表現データベースに格納する処理を含むものである。

【００１８】
本発明は上記のようなデータ表示装置を提供することができる。さらに、該データ表示装置などにおいて用いるデータ表示方法として提供することもできる。
すなわち請求項８に記載の発明は、少なくとも自然科学で用いられる記号又は式（以下、科学記号と呼ぶ）を含むテキストデータを表示するデータ表示装置におけるデータ表示方法である。
【００１９】
該方法は、入力処理部がテキストデータを入力する入力ステップ、科学記号として用いる１文字以上の文字列と当該文字列が発現する発現条件とを予め格納した科学記号データベースを参照し、科学記号抽出処理部が該テキストデータから科学記号候補文字列を抽出する科学記号抽出ステップ、該発現条件と照合して文字列表示決定処理部が該科学記号候補文字列の表示態様を決定する文字列表示決定ステップ、表示部が決定された表示態様の科学記号候補文字列を含むテキストデータを表示する表示ステップを含むことを特徴とするものである。
【００２０】
請求項９に記載の発明は、上記の文字列表示決定ステップにおいて、テキストデータの基本文字色を予め設定すると共に、該科学記号候補文字列の表示色を該テキストデータの表示に用いる基本文字色から変化させる処理を行うデータ表示方法を提供する。
【００２１】
請求項１０に記載の発明は、上記の文字列表示決定ステップにおいて、科学記号確度算出部が、科学記号候補文字列が科学記号である確度を算出する科学記号確度算出ステップを行う方法を提供する。
該方法において、算出結果に基づいて確度が閾値よりも高い科学記号候補文字列については基本文字色と色相の異なる表示色を設定すると共に、確度が閾値よりも低い科学記号候補文字列については基本文字色と彩度又は明度が異なる表示色を設定するものである。
【００２２】
請求項１１に記載の発明は、形態素解析処理部が前記テキストデータを形態素解析する形態素解析ステップを前記科学記号抽出ステップの前に実行する。そして科学記号確度算出ステップにおいて、該科学記号候補文字列の前後所定個数の形態素の少なくとも文字列情報又は文法情報のいずれかを用い、前記科学記号データベースに備えた少なくとも文字列情報又は文法情報のいずれかを参照して確度を算出することを特徴とする。
【００２３】
さらに請求項１２に記載の発明は、科学記号確度算出ステップにおいて、科学記号候補文字列の前後に同一又は異なる科学記号候補文字列が連続して出現した場合に当該科学記号候補文字列の確度を所定値だけ高める処理を含む。
【００２４】
また、請求項１３に記載の発明は、科学記号抽出ステップにおいて、予め定めた科学記号を構成する特定表現を抽出すると共に、科学記号確度算出ステップにおいて、該特定表現が前後所定個数の形態素内、又は同一文、又は同一テキストデータ中に出現した場合に、当該科学記号候補文字列の確度を所定値だけ高める処理を含むことを特徴とする。
【００２５】
請求項１４に記載の発明は、特定表現を、特定表現データベースに格納する構成において、着目している科学記号候補文字列と共に、テキストデータの同一文又は所定個数の形態素内に特定表現候補が出現する回数Ｎ１を計数する一方、該特定表現候補がその他の文において単独に出現する回数Ｎ２を計数し、Ｎ１／Ｎ２（Ｎ２≠０）又はＮ１／（Ｎ１＋Ｎ２）（Ｎ２が０のときも含む）の少なくともいずれかの値が閾値以上の場合に、該特定表現データベースに格納する処理を含むものである。
【発明の効果】
【００２６】
本発明は、上記構成を備えることにより、高精度に科学記号の表示態様を変化させ、読

者が科学記号を識別しやすい表示装置及び方法を提供することができる。とくにラテン文字により記述されたテキスト中であっても科学記号を適切に表示できるため必要な情報を容易に読み取ることができるようになる。

また、科学記号であるか否か、確度により表示態様を区別することで確度の低い科学記号候補は読者が気にならない程度の表示を行う一方、確度の高い科学記号候補は明確に色分けすることができる。本方式を採用することで、過剰な言語処理技術を用いて処理速度の遅延やデータベースの増大を引き起こすことなく簡便な装置に寄与する。

【発明を実施するための最良の形態】

【００２７】

以下、本発明の実施形態を、図面に示す実施例を基に説明する。なお、実施形態は下記に限定されるものではない。

図１は本発明に係るデータ表示装置(１)の全体構成図である。本発明は公知のパーソナルコンピュータにより容易に実現することが可能であり、演算処理やテキスト処理などを司るＣＰＵ（１０）によって本発明の各ステップを実行処理する。ＣＰＵ（１０）は周知のようにメモリ（１１）と協働して動作し、キーボード（１２）やマウスなどの入力手段の他、出力結果を表示するモニタ（１３）、ハードディスク等の外部記憶装置（１８）などを備えている。

【００２８】

図２に示すように、本装置（１）に対して論文などのテキストデータ（２０）を入力処理部（１４）の作用によって装置に取得（２１）する。テキストデータ（２０）としては英語等のラテン文字を用いた科学論文の場合に本発明は最も有効に作用する。

そして、該テキストデータ（２０）から化学記号や物理記号、特に元素記号、電子配置、分光記号などの自然科学で用いる記号及び化学式、数式などの式を抽出表示する。本発明ではこれらを総称して科学記号と呼ぶ。

【００２９】

入力されたテキストデータから科学記号抽出処理部（１５）において予め科学記号とその発現条件を格納したデータベース（２３）を参照して科学記号の抽出処理（２２）を行う。該データベース（２３）は外部記憶装置（１８）内に格納される。

図３に示すような元素記号が含まれた論文を入力すると、文頭から各文字列を順に読み出し、データベース（２３）に含まれる科学記号情報と照合する。合致する文字列があるとその文字列を抽出し、どのような表示態様で表示をおこなうか決する文字列表示決定処理部（１６）にて処理を行う。

【００３０】

ここでデータベース（２３）の内容例を図４に示す。データベースには各元素記号等（３０）に対応して、それが単体でテキスト上に発現したときの科学記号である確度（３１）が定義されている。例えば水素（Ｈ）に対しては０．１、ヘリウム（Ｈｅ）に対しては、０．２、リチウム（Ｌｉ）に対しては０．５というように定義している。

【００３１】

このように各元素に対して確度が異なるのは、元素記号が英単語と一致することがあり、その一致の可能性の大小によって定義しているからである。すなわち、Ｈｅの場合、英単語の彼を表すＨｅと一致しているため、文頭に単独で発現した場合には「Ｈｅ（彼）」か「Ｈｅ（ヘリウム）」かの判断が難しい。そのため確度は０．１となる。一方、ネオン「Ｎｅ」の場合、英語で文頭にＮｅが書かれる場合は極めて希であるから、確度は０．７としている。

【００３２】

このように確度は対象とする言語によっても異なるため、テキストの言語に応じてそれぞれ定義されることが望ましい。例えば日本語論文の中でＨｅが発現するのは通常は多くないため、より高い確度を定義してもよいと考えられる。

【００３３】

本発明の構成では、以上の確度を取得することにより、文字列表示決定処理部（１６）

で確度に応じた文字色を決定し、表示部（１７）の処理によってモニタ（１３）上にテキストを表示する。

各確度に対する表示色は予め装置（１）上に設定する。文字色としては次のような実施形態が挙げられる。

【００３４】

すなわち、テキストの全文又は一領域が黒色である場合、確度が閾値以上の場合にそれを赤色で表示する一方、閾値よりも低い場合には色を変化させないことができる。この場合、例えば閾値を０．２とするとＨは黒色のまま、Ｈｅは赤色で表示されることとなる。データベース（２３）の通り、元素名を表す英語名称（hydrogenなど）は確度がいずれも１であるから、すべて赤色で表示される。

【００３５】

この方法は科学記号が特有な場合には簡便であるが、元素記号のケースでは色を変化させられないものや、誤って変化させてしまうものが多く見られる。そのため元素記号などの場合には次のような実施形態をとることが望ましい。

すなわち、確度に応じて表示色を変化させる構成である。この場合に閾値を２個以上備えておき、例えば閾値０．６以上の場合には赤色、０．１以上０．６未満の場合は灰色で表示すると定義しておく。

【００３６】

この場合、Ｎｅや元素名称は赤色、それ以外の元素については灰色で表示される。ここで赤色とはテキストを表示する基本文字色（黒色）と色相が異なる色の例であり、色相が異なることで読者は完全に当該文字列を識別することができる。黒色の基本文字色に対してピンク色、黄色なども好適である。

一方、灰色とは基本文字色と明度が異なる色の例である。基本文字色と明度が異なるだけの場合、読者は強い違和感を覚えることがない。特に意識しない限り閲読を妨害しないので快適に閲読することができる。逆に意識をして読むと、明らかに基本文字色と異なるので明確に視認することができる。

【００３７】

このように本方法によれば、確実に科学記号と判定できるものについては読者に強く提示する一方、不確実なものについては注意を促す程度の表示が可能である。明度と共に彩度を変化させる構成でもよい。

なお、色相、明度、彩度は表示部（１７）で周知の技術により変化させて表示することが可能である。

【００３８】

ここで本発明の特徴として確度を算出する時に発現条件に基づいて行うことが挙げられる。以下にこの点を説述する。

本発明における発現条件とは確度を算出する科学記号がテキストデータ中でどのような条件下で発現しているかを定義したものである。例えば上述した例では各元素が「文頭に単独で発現した」ことを条件としている。すなわち文字列表示決定処理部（１６）ではピリオド、読点、改行コードなどに基づいてその発現位置が文頭であるか否かを判定する。

【００３９】

発現条件を用いた確度Ｙの算出は文字列表示決定処理部（１６）において次式に従って行う。

【００４０】

（数１）　$Y=p(str)+\Sigma a_i(str)\times x_i$

上記において$p(str)$は科学記号候補文字列strの基礎となる確度（３１）、$a_i(str)$は科学記号候補文字列strに対するデータベース（２３）の発現条件iで定められた確度であり、x_iは発現条件iに該当するときに１、該当しないときに０をとる。

【００４１】

従って、strが「Ｈ」であるとき、後述するように文頭（３２）になく（$x_{cap}=1$)、連接（３３）する文字列がなく（$x_{cohere}=0$)、イオン表記（３４）でない（$x_{ion}=0$)場合には

、0.1+0.1*1+0.2*0+1*0=0.2が求める確度となる。

なお、上記のx_{cap}、x_{cohere}、x_{ion}はそれぞれ数１におけるx_iの発現条件として「文頭にない」「連接する文字列がある」「イオン表記である」に対応するパラメータである。

【００４２】

テキストデータ（２０）から科学記号抽出処理部（１５）で抽出された科学記号候補文字列がピリオド等の直後に配置される場合には文頭に発現したものと判定できるので、上記データベース（２３）の文頭位置に対応する各確度を取得する。

【００４３】

しかし、同時にデータベース（２３）には当該文字列が文頭でない場合の確度を格納している。これに係る項目が図４のcapで表示された欄（３２）である。データベース（２３）の２行目は、文頭でない位置に「Ｈｅ」が出現した時にその確度は１を加算することを意味している。従って、この場合確度は１．２となる。実際には本実施例では確度が１を最大と規定しており、１を超えた確度は全て１として処理する。

【００４４】

なお、本発明の実施形態としてデータベース（２３）中に大文字を含む文字列が掲載されている場合には大文字と小文字を掲載されている通りに区別し、小文字だけで表記された文字列については全て小文字の他、全て大文字、それらの混在、いずれも抽出対象としている。

英語の場合には文頭以外に先頭が大文字の文字列が配置されていれば固有名詞等である可能性が高く、このようにすることで１文字目を一般的に大文字で表記する元素記号等を高精度に表示することができる。

【００４５】

本発明の発現条件としては文頭か否かだけでなく、データベース（２３）上に掲載された他の文字と分かち書きを行わずに連接して表記されている場合の確度を定義している。本項目は欄（３３）のcohereに続く数値でありこれに基づいて確度を算出する。例えばデータベース（２３）の８行目にあるＯの場合、単独で文頭にある場合には確度は０．１であるが、仮にＨと連接してＯＨと記載されていた場合、確度は０．２が加算されて０．３となる。

以上の構成によりＯＨのように連接した場合には単体のＯよりも確度が高く評価されるため、正確な表示を行うことができるようになる。

【００４６】

なお、ＯＨのように２個の連接でなく、３個以上の科学記号候補文字列が連接した場合にも確度はそれぞれについて０．２を上限として加算するようにしている。これは、略語など大文字が連続した場合でも必ずしも科学記号とは言えない場合が多いためであり、徒に確度が高まるのを防ぐようにしている。

【００４７】

これと関連して、科学記号と判定されやすい特定の文字列について確度を下げるデータベースを外部記憶装置（１８）に備えてもよい。科学記号除外文字列データベース（図示しない）として設け、科学記号抽出処理（２２）において該データベースと一致した場合には抽出しないようにすることができる。すなわち、Convergent Close-Coupling やSmall Office Home Officeを示すＣＣＣ、ＳＯＨＯ等の文字列の場合、これらを構成する文字列はいずれも科学記号であって、連接することから確度が上昇しやすい。しかし、抽出処理の段階で科学記号除外文字列データベースに一致した文字列については抽出結果から取り除く処理を行う。

もちろん、科学記号除外文字列データベースを用いずに本発明は構成することができる。

【００４８】

あるいは、確度算出処理（２４）において、該科学記号除外文字列データベースと一致する文字列については確度を０になるように算出処理をおこなってもよい。この場合、科

学記号除外文字列データベースを別に設けず、上記科学記号及び発現条件を格納したデータベース（２３）に例えば確度－１０として登録しておいてもよい。計算結果で負となる場合に確度０として処理することで、これらの文字列はいずれも確度０となり、科学記号候補から除外される。

【００４９】

上記構成に加えて、連接する各文字列に対して、連接文字列中で最も確度が高くなる文字列と等しい確度を設定することができる。

上記のＯＨを例にとると、Ｏの確度は文頭であってＨと連接するため確度は０．３、Ｈの確度は文頭でなくＯと連接することから０．８となる。このような場合、Ｏの確度は連接文字列中で最も高い０．８と設定する。

本構成により、一連の連接する文字列間で確度に整合性がとれるだけでなく、文字色を確度によって変化させた場合に視認しやすい表示に寄与する。

【００５０】

科学記号候補文字列の並びについては他にも次のような処理が可能である。

まず、分子構造を表す場合などハイフンを用いて元素を接続することがある。本実施例では上記連接の場合と同様にハイフンで接続された文字列も処理する。このように科学記号を接続するのに用いられる文字記号を予め記憶させておき、該文字記号で接続されている場合には連接しているのと同様の処理を行わせてもよい。

【００５１】

あるいは上記のように連接した場合に連接文字列中で最も高い確度を各文字列に設定するのではなく、所定の確度以上の文字列と連接する場合に、各確度を上昇させるように構成してもよい。すなわち、データベース（２３）に例えばhighという項目を設けて、閾値０．６以上の科学記号候補文字列と連接した場合に、確度＋０．７又は０．８を定義する。この場合、上記ＯＨの例で言えば、Ｈの確度が０．８で閾値以上であるため、Ｏの確度も例えば０．７加算されて１となる。

【００５２】

さらに、連接の概念をより広めて構成することもできる。すなわち、本発明に言う連続とは、科学記号候補文字列が連接した場合、ハイフンで接続された場合に加えて、当該テキストデータの言語における接続詞等を用いて接続した場合を含めても良い。英語であれば、複数の名詞を並列する場合に、A,B and Cのように、コンマと文字列andで接続される。

このとき、抽出された科学記号候補文字列間にコンマ又はandやorなどを含む場合に、連接しているのと同様（この場合を並列と呼ぶ。）に処理することができる。

【００５３】

並列の場合にも、全ての並列する科学記号候補文字列の確度を並列文字列中で最大確度に合わせてもよいし、データベース（２３）に定めた値を加算するようにしてもよい。後者の場合には、連接の場合とは異なる数値を定めることもできる。

以上のように連接や並列の場合に、他の科学記号候補文字列の確度を互いに影響させることで高精度な表示を行うことができる。

【００５４】

発現条件は対象とする科学記号に合わせて適宜定義することができる。例えば元素記号の場合にはイオンを示すプラス・マイナス記号が付されることが多く、これらが付された場合には極めて高い確度で科学記号と判定できる。

具体的にはテキストデータ（２０）中に、タグなどによって書式指定がされ、In⁺のように、上添字の＋によるイオン表記となる科学記号を検出する。同様に（ｎ＋）や（ｎ－）（ｎは任意）などの所定の書式の場合に、図４における欄（３４）に従って確度を１とする。

【００５５】

同様に例えば分光記号におけるＳＰＤＦなどの文字列や、原子軌道を示すｓ軌道、ｐ軌道の電子配置、遺伝子の塩基配列におけるＡ、Ｇ、Ｔ、Ｃ、Ｕなどの文字列を他の文字と

の組み合わせで確度を算出するようにしてもよい。
これらの科学記号は文字の記載順序など確立されたルールに従って発現するため、本発明のように発現条件を付与可能なデータベース（２３）を用いることで効果的に抽出することができる。
【００５６】
イオン表記や、他の文字との組み合わせで確度が高くなった科学記号について、同一のテキストデータ中で単独で出現した場合にもその確度を上げる処理をおこなってもよい。
すなわち、一度全部のテキストデータについて確度算出（２４）を行ってイオン表記等による確度の確定を行い、同ステップ（２４）内において再び抽出された各科学記号候補文字列について確度の再定義処理を行う。
本処理では、イオン表記など所定の発現条件に合致した文字列について、単独で現れているものを抽出し、その確度に所定値、例えば＋０．７を加算する。あるいは、上記イオン表記等で定義された確度と同一値を与えてもよい。
本処理によれば、イオン表記や他の文字との組み合わせの出現によって単体でも現れる蓋然性の高い文字列について高い確度を定義することができる。
【００５７】
本発明の別実施例として、図５に示すような形態素解析処理部（５０）を備えたデータ表示装置（１’）を提供することができる。
形態素解析については公知の技術であり、日本語の形態素解析技術として例えば茶筌（非特許文献２に開示されている）を用いることができる。
【００５８】
【非特許文献２】chasen.aist-nara.ac.jp
【００５９】
また、分かち書きをする英語などのラテン文字を用いるテキストデータでは形態素への分割は容易であるがＨＭＭなどの統計的手法により同様に解析処理が行える。形態素解析を用いて品詞を見分けることも行われている。
【００６０】
形態素解析処理は図６に示すように前述の実施例における科学記号抽出ステップ（２２）の前に行う。このとき周知のように外部記憶装置（１８）に格納された形態素解析辞書（５２）を用いながら解析する。上記実施例ではデータベース（２３）に掲載された情報と照合することで科学記号を抽出（２２）したが、本実施例では解析の結果得られた形態素と該データベース（２３）の内容とを比較して一致するものを抽出（２２）する。
【００６１】
形態素解析をすると、形態素の区切りがより正確になるためデータベース（２３）との照合も確実に行うことができる。さらに形態素解析で各形態素の品詞を取得することができる。これを利用し、データベース（２３）に文字列と共に品詞情報を付与し、上記と同様にその場合の確度を定義しておくこともできる。
本構成によると、例えばＨｅが名詞であれば元素名である確度を高く定義する一方、代名詞であれば科学記号である可能性は極めて低いため確度を０となるように「-10」と定義することもできる。
以上のような別実施例によりさらに高精度なデータの表示装置を提供することが可能である。
【００６２】
さらに本発明では、ある科学記号は特定の文字列と共にテキストデータ中に現れるときに、科学記号である確度が高いことに着目して次のような処理を行うこともできる。すなわち、特定の文字列を手がかり表現とし、テキストデータ中の同一文あるいは前後所定の形態素数内において科学記号と共起しやすい文字列（手がかり表現）が抽出されるときに、対応する科学記号の確度を高める。本構成は、科学記号抽出処理部（１５）において、科学記号を抽出すると共に、図７に示すように手がかり表現テーブル（５４）を参照して手がかり表現を抽出（５３）する。

【００６３】
手がかり表現テーブルには、例えば元素記号と共起しやすい表現である「-like ion」などと、各元素記号との組み合わせを格納しておく。
そして、共起文字列「-like ion」が抽出された場合には、組み合わせとして定義されている各科学記号候補文字列の確度を確度算出（２４）において上昇させる。上昇値は上記のようにデータベース（２３）中に定義しておくか、共起文字列テーブル（５４）中に共起した場合の確度の値を定義しておく。
【００６４】
上記では手がかり表現テーブルを予め人手によって定義するが、これを自動化して該テーブルを構成することもできる。本処理を図８に示す。
本処理には一般的な例文として科学記号を含むテキストコーパス（５５）を用いる。該コーパスについては公知であり、予めテキスト中の単語列の形態素、品詞等が定義されている。文字列が科学記号か否かも定義されている。
なお、本発明では単語列の形態素、品詞などが定義されていないコーパスを用いても良く、その場合には公知の形態素解析器（図示しない）や辞書データベースを用いてこれらを自動的に付与した後に、次の処理に進んでもよい。
【００６５】
まず、テキストコーパス（５５）からデータベース（２３）を参照して科学記号候補文字列を抽出（５６）する。
そして、該テキストコーパス（５５）中の当該科学記号候補文字列を含む同一文に共起する文字列（手がかり表現候補）を抽出する。テキストコーパス（５５）内の全文について手がかり表現候補が科学記号候補文字列と共起する回数Ｎ１をカウント（５７）する。
【００６６】
次に、当該手がかり表現候補を含む文について、当該科学記号候補文字列が現れない回数Ｎ２をカウント（５８）する。すなわち、科学記号候補文字列と手がかり表現候補が共起せず手がかり表現候補のみが単独で現れる回数である。
さらに、Ｎ２が０でなければＮ１／Ｎ２を算出（５９）することにより、共起する割合が所定の閾値以上であるか否かを確認する。Ｎ２が０の場合には閾値以上のときと同様に処理を行っても良いし、Ｎ１が所定回数、例えば３回以上の場合にだけ同様の処理を行っても良い。
あるいは、Ｎ２がすべての場合に適用しうるように、Ｎ１／Ｎ２の算出（５９）に替えてＮ１／（Ｎ１＋Ｎ２）を算出する構成でもよい。
【００６７】
加えて、上記の回数Ｎ１が回数Ｎ２よりも有意に大きいことを二項検定などの公知の統計的検定の手法に基づいて確認（６０）し、確認が取れた場合に、当該手がかり表現候補と科学記号候補文字列との組み合わせを手がかり表現テーブル（５４）に記録する。
【００６８】
本実施例で二項検定を行う方法を説述する。
初期値として、一回の試行で科学記号候補文字列と手がかり表現候補とが共起する確率及び、科学記号候補文字列と手がかり表現候補とが共起せず後者だけが単独で出現する確率をそれぞれ０．５とする。
そして、Ｎ１＋Ｎ２の総出現のうちＮ２回以下、科学記号候補文字列と手がかり表現候補とが共起せず手がかり表現候補のみが出現した確率を求める。
すなわち、この確率
（数２）

$$P1 = \Sigma\ C(N1+N2,x) * 0.5^{x} * 0.5^{N1+N2-x}$$

（ただし、Σは、x=0からx=N2の和、C(A,B) は A個の異なったものからB個のものを取り出す場合の数である。）
で表され、この確率の値が十分小さければＮ１とＮ２は等価な確率でない、すなわち、Ｎ１がＮ２に比べて有意に大きいことが判断できる。

そして、５％検定ならば上記Ｐ１が５％よりも小さいこと、１０％検定ならばＰ１が１０％よりも小さいこと、が有意に大きいかどうかの判断基準となる。

【００６９】

上記では同一文としたが、単に同一文ではなく、共起する表現を前方で連接する単語列（前方１単語列に共起する）や共起する表現を後方で連接する単語列（後方１単語列に共起する）手がかり表現候補に限定してもよい。単語列としては形態素や、形態素の集合を用いることができる。

【００７０】

科学記号候補文字列の確度を高精度に算出する別の方法として、次の技術を組み合わせて用いることもできる。

本技術は科学記号候補文字列が、一般的な文章に比して多く出現する場合には当該文字列が科学記号である確度が高いと判定するものである。例えば、leadという文字列を考えたとき、これは科学記号（元素名）である可能性と、「導く」などを意味する英単語である可能性とがある。

【００７１】

後者の意味の英単語は一般的な文章において頻繁に出現することは少ないが、科学論文において鉛を話題にした文章では頻繁に出現する。この場合、科学記号として処理するのが好適である。

そこで、図９に示すように、まずテキストデータ（２０）から科学記号を抽出したとき、抽出された当該科学記号の個数と該テキストデータ（２０）を構成する全単語数との比、すなわち出現率Ｒ１（当該科学記号候補文字列の出現数／全文字列総数）を算出（６２）する。

【００７２】

次に、一般的なテキストコーパス（６３）（例えば新聞記事）を用いて、同様に該テキストコーパス（６３）における当該科学記号候補文字列の出現数／全文字列総数を算出（６４）する。これを出現率Ｒ２とする。

そして、出現率の比Ｒ１／Ｒ２を算出（６５）し、所定の閾値より大きいか否かを判定する。

加えて、上記のＲ１がＲ２よりも有意に大きいことを比の検定、またはカイ二乗検定などの公知の統計的検定の手法に基づいて確認（６０）し、確認が取れた場合（例えばカイ２乗検定で１％水準、又は５％水準等で有意と認められた場合）に、当該手がかり表現候補と科学記号候補文字列との組み合わせを手がかり表現テーブル（５４）に記録する。

【００７３】

上記カイ２乗検定について説述すると、Ｒ１を計算する分母、分子をそれぞれＮ１、Ｆ１とし、Ｒ２を計算する分母、分子をそれぞれＮ２、Ｆ２とする。

Ｎ＝Ｎ１＋Ｎ２として、カイ２乗値は次式により求められる。

（数３）

$$\text{カイ二乗値} = \frac{(N*(F1*(N2-F2)-(N1-F1)*F2)^2)}{((F1+F2)*(N-(F1+F2))*N1*N2)}$$

【００７４】

そして、このカイ二乗値が大きいほどＲ１とＲ２は有意差があると言え、例えばカイ二乗値が3.84よりも大きいとき危険率５％の有意差があると言え、カイ二乗値が6.63よりも大きいとき危険率 1% の有意差があると言える。

【００７５】

次に比の検定を用いる場合を説述する。まず、

（数４）

$$p = (F1+F2)/(N1+N2)$$
$$p1 = R1$$
$$p2 = R2$$

と定義する。

そして、２群の比率の差の検定における検定統計量は、
（数５）
Z = |p1-p2|/√(p*(1-p)*(1/N1+1/N2))
で表される。
このとき、Ｚが大きいほど、Ｒ１とＲ２は有意差があると言え、Ｚが1.96よりも大きいとき危険率５％の有意差があると言え、Ｚが2.58よりも大きいとき危険率１％の有意差があると言える。
【００７６】
これらの実施例において確度をデータベース（２３）に予め定義する構成を説述した。しかし以下のようにテキストデータから確度を自動的に修正する構成を用いることもできる。
図１０に示すように、テキストを入力（２１）した後、科学記号を抽出（２２）する際に、テキストデータ（２０）中の科学記号候補文字列の数をカウント（７０）する。該カウントはＣＰＵにより公知の方法で実行処理することができる。
【００７７】
そして、該カウントが予め定めた閾値（例えば５００ワード中に５回以上などと定義する）である場合（７１）には、データベース（２３）に定義された確度を上昇させる書き換え処理（７２）を行う。
このように書き換えられたデータベース（２３）を用いて確度の算出を行うことで、頻繁に出現する文字列については科学記号であるとの判定が出やすくする。本方法が有効であるのは例えば英語の前置詞と元素記号が同一スペルの場合に、そのスペルの文字列が一定以上多い場合には、そのテキストデータには当該元素記号に係る内容が含まれている可能性が高く、これらをもれなく抽出表示するためである。
【００７８】
また、ＮやＯなどの大文字１文字の場合にも有効であり、文頭以外の場所に頻繁にこれらの文字が発現する場合には、文頭に発現した際にも科学記号であるとの判定が出やすくなる。
【００７９】
なお補足すると、データベース（２３）には確度ではなく表示色を直接定義してもよい。この場合、発現条件毎に表示色を直接定義し、上記同様の効果を奏する。
また、本実施例では表示色を変更する構成を開示したが、色ではなく書式を変化させる構成でもよい。周知のようにテキストデータの表示態様としては文字フォントの変更や下線の付与、網掛け表示、括弧による範囲表示などが知られており、これらを用いて文字色を変化させる代わりに所望の範囲を読者に表示することができる。
【００８０】
以下には、本発明の具体的な実施例として、表示色と各科学記号候補文字列の判定ルールについて説述する。
図１１は、本発明における表示色の定義である。図示のように、ルール１，３，４，５，６，７，８を定め、それぞれにルール１では原子・分子・イオンを表現する場合に桃色で表示すること、ルール３では電子配置の表現に黄色で表示すること、のように定義している。
なお、ルール２は欠番である。
【００８１】
上述した発現条件と関連して、ルール１の判定には電子eや、+/-の上下添字、原子名に上下添字、IVXivxの表現、"like"/"ic"についても同様に桃色で表示することを定義する。
ルール３の判定では、「数字＊」（＊はあってもなくても良いことを示す。以下同じ。）「s/p/d/f/g」「上下添字＊」の一回以上の繰り返しでかつ、数字が少なくとも１回は含まれることを条件とする。
【００８２】

ルール４の判定では、「上下添字＊」「S/P/D/F/G」「上下添字＊」の一回以上の繰り返しでかつ、「上下添字」が少なくとも１回は含まれることを条件とする。
また、上記ルール１と競合した場合は下のより厳密な規則を採用する。
すなわち、「上添字＊」「S/P/D/F/G」「下添字＊」の一回以上の繰り返しでかつ、添字の中身は1から4に限られ、上下添字のいずれかは出現する条件とする。

【００８３】

ルール５の判定では、「n/l」「=/</>」の一回以上の繰り返しや、数字を条件として水色で表示する。
ルール６の判定では、「（ルール３の表現）のゼロ回以上の繰り返し」「数字/n/n-bar l」が出現した場合に、橙色で表示する。

【００８４】

ルール７の判定では、英語アルファベット大文字一文字からなる原子名について、まわりに手がかり表現(-like ion等)などがなければ、原子名でない可能性が高いと判断してルール７に分類する。また、英語アルファベット大文字一文字からなる原子名が連続した表現や"Rev"、の場合にも手がかり表現がなければ同様にルール７に分類する。
As,In,At,Heが文頭に出現した場合、前置詞や代名詞の可能性が高いためルール７に分類する。

【００８５】

さらに以上のような表示色のルールによっていずれの条件にも合致しなかったものの、科学記号候補文字列として抽出されたものをルール８とし、濃い灰色で表示した。
以上のような表示色のルールは、上記確度の算出結果に連動しており、データベース（２３）の構成を適切に設計することによって実現している。

【００８６】

本発明によるシステムの処理により次のような結果が得られた。
まずテキストデータとして単語148個の科学論文要約を入力した。その中でInを調べたところ、出現は24個で、4個が原子名のIn で、20個が前置詞のInであった。本発明のシステムでは、4個の原子名のIn はすべて正しく原子名として抽出し桃色で表示され、20個の前置詞のInは、すべて正しく灰色表示された。
ここで、正しく原子名となった4個の原子名のInは、In⁺のようにInの右肩に"+"がつくイオンの形をしていたため、確度が高くなったものと考えられる。それ以外は文頭の前置詞のInで灰色表示された。

【００８７】

同じテキストデータで大文字「Ｏ」を調べたところ、出現は19個で、18個が原子名のＯ、1個が人名省略のＯ(Ｏ.)だった。本発明のシステムでは、やはりイオン表記を手がかりとして、19個の原子名のＯはすべて正しく原子名として抽出し、1個の人名省略のＯは，正しく灰色表示された。
以上のように、本発明における発現条件を用いた処理は、好適に科学記号の抽出表示が行えることが示された。また、確度の低い文字列については灰色表示されたことで、読む際の障害にならず、好適な視認性を得ることができた。

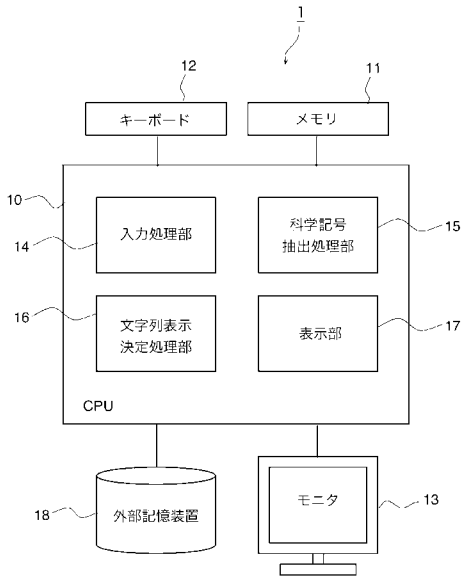
【図面の簡単な説明】

【００８８】

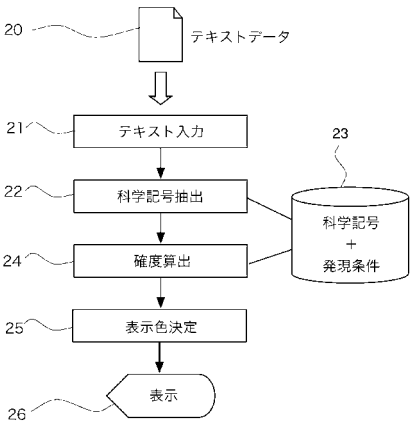
【図１】本発明のデータ表示装置（第１実施例）の全体構成図である。
【図２】本発明のデータ表示方法（第１実施例）の流れ図である。
【図３】本発明で対象とするテキストデータの一例である。
【図４】本発明で用いるデータベースの内容である。
【図５】本発明のデータ表示装置（第２実施例）の全体構成図である。
【図６】本発明のデータ表示方法（第２実施例）の流れ図である。
【図７】本発明のデータ表示方法（第３実施例）の流れ図である。
【図８】本発明のデータ表示方法（第４実施例）の流れ図である。
【図９】本発明のデータ表示方法（第５実施例）の流れ図である。

【図１０】本発明における発現条件を変化させる処理の流れ図である。
【図１１】本発明における表示色の定義である。
【符号の説明】
【００８９】
１　データ表示装置
１０　ＣＰＵ
１１　メモリ
１２　キーボード
１３　モニタ
１４　入力処理部
１５　科学記号抽出処理部
１６　文字列表示決定処理部
１７　表示部
１８　外部記憶装置

【図１】



【図２】



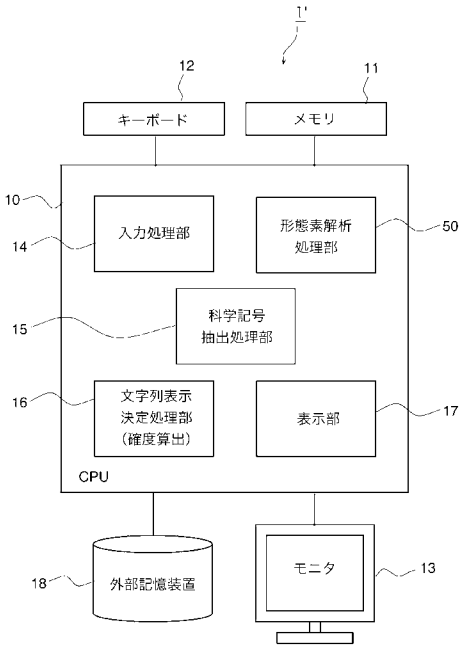
【図３】

InN buffer layers were grown at 300 °C for 2 min directly on 3C-SiC or on c-GaN underlayers. After the growth of buffer layers, the substrates were heated up to 400–550 °C, and then InN films were grown for 1 hour. The crystal structures were monitored during growth by use of reflection high-energy electron diffraction (RHEED). We have studied the crystal structures and morphologies by X-ray diffraction (XRD) and scanning electron microscope (SEM), respectively. We have measured photoluminescence (PL) spectra of the InN films at 5 K using the 633 nm line of a He-Ne laser.

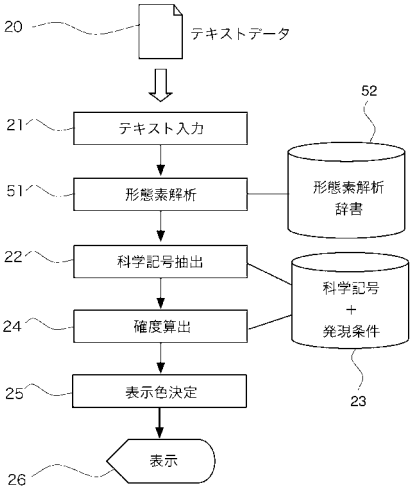
【図４】

30	31	32	33	34
H	0.1	cap +0.1	cohere +0.2	ion +1
He	0.2	cap +1	cohere +1	ion +1
Li	0.5	cap +1	cohere +1	ion +1
Be	0.2	cap +1	cohere +1	ion +1
B	0.1	cap +0.1	cohere 0	ion +1
C	0.1	cap +0.1	cohere 0	ion +1
N	0.1	cap +0.1	cohere +0.2	ion +1
O	0.1	cap +0.1	cohere +0.2	ion +1
F	0.1	cap +0.1	cohere +0.2	ion +1
Ne	0.7	cap +1	cohere +1	ion +1
hydrogen	1			
helium	1			
lithium	1			
beryllium	1			
boron	1			
carbon	1			
nitrogen	1			
oxygen	1			
fluorine	1			
neon	1			

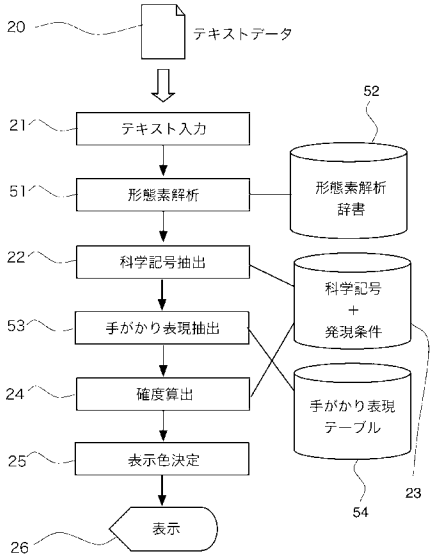
【図５】



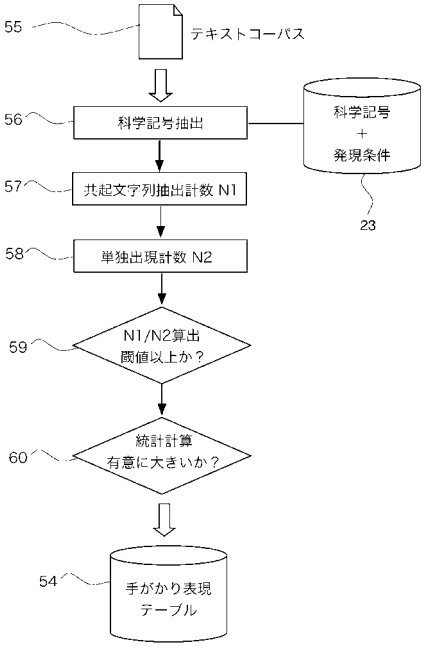
【図６】



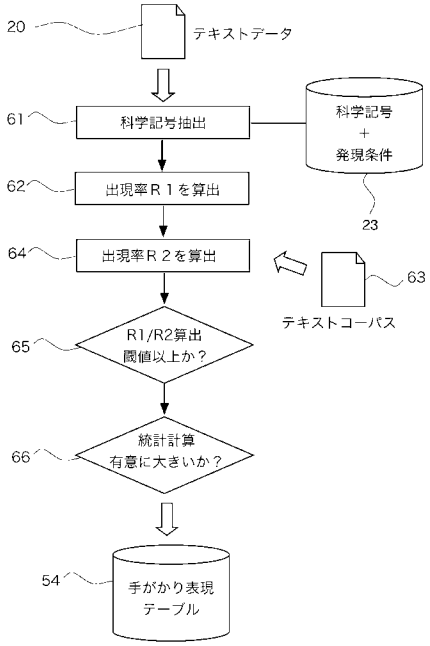
【図７】


20
テキストデータ
21
テキスト入力
51
形態素解析
22
科学記号抽出
53
手がかり表現抽出
24
確度算出
25
表示色決定
26
表示
52
形態素解析
辞書
科学記号
＋
発現条件
手がかり表現
テーブル
54
23

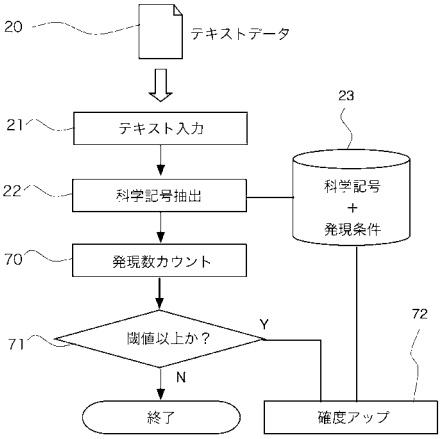
【図８】


55
テキストコーパス
56
科学記号抽出
57
共起文字列抽出計数 N1
58
単独出現計数 N2
59
N1/N2算出
閾値以上か？
60
統計計算
有意に大きいか？
54
手がかり表現
テーブル
科学記号
＋
発現条件
23

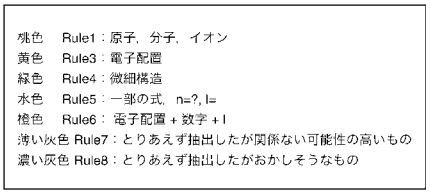
【図９】


20
テキストデータ
61
科学記号抽出
62
出現率Ｒ１を算出
64
出現率Ｒ２を算出
65
R1/R2算出
閾値以上か？
66
統計計算
有意に大きいか？
54
手がかり表現
テーブル
科学記号
＋
発現条件
23
63
テキストコーパス

【図１０】


20
テキストデータ
21
テキスト入力
22
科学記号抽出
70
発現数カウント
71
閾値以上か？
Y
N
終了
確度アップ
72
23
科学記号
＋
発現条件

【図１１】


桃色 Rule1：原子、分子、イオン
黄色 Rule3：電子配置
緑色 Rule4：微細構造
水色 Rule5：一部の式、n=?, l=
橙色 Rule6：電子配置＋数字＋l
薄い灰色 Rule7：とりあえず抽出したが関係ない可能性の高いもの
濃い灰色 Rule8：とりあえず抽出したがおかしそうなもの

フロントページの続き

Ｆターム(参考) 5B075 ND03 QM05
5E501 BA03 BA09 DA16 FA13 FB28 FB44