(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2010-205060
(P2010-205060A)
(43)公開日　平成22年9月16日(2010.9.16)

| (51)Int.Cl. | | F I | | | テーマコード（参考） |
|---|---|---|---|---|---|
| G06F 17/30 | (2006.01) | G06F | 17/30 | 210A | 5B075 |
| | | G06F | 17/30 | 170B | |

審査請求 未請求　請求項の数 10　ＯＬ　（全 16 頁）

(21)出願番号　特願2009-50950(P2009-50950)
(22)出願日　平成21年3月4日(2009.3.4)

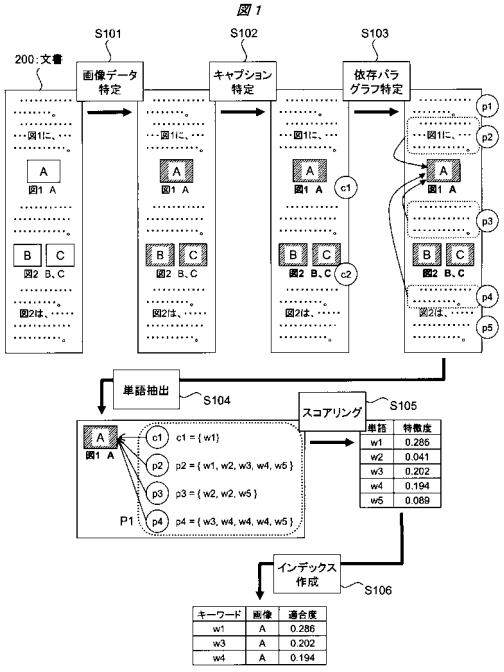
(71)出願人　000155469
株式会社野村総合研究所
東京都千代田区丸の内一丁目６番５号
(74)代理人　100080001
弁理士　筒井　大和
(74)代理人　100093023
弁理士　小塚　善高
(74)代理人　100117008
弁理士　筒井　章子
(72)発明者　竹原　一彰
東京都千代田区丸の内一丁目６番５号　株式会社野村総合研究所内

Ｆターム(参考)　5B075 ND08　NK02　NK31　PQ74

(54)【発明の名称】文書内画像検索方法および文書内画像検索システム

(57)【要約】
【課題】文書から抽出した画像に対して、キーワードを文脈情報を考慮して効率的に精度良く付与することにより、キーワードを利用した高精度で効率の良い画像の検索を可能とする文書内画像検索方法を提供する。
【解決手段】文書の中から画像の位置を特定し、画像のデータを抽出する処理（Ｓ１０１）と、画像のキャプション領域を特定し、画像名とキャプションとを抽出する処理（Ｓ１０２）と、画像について記述している依存パラグラフを特定する処理（Ｓ１０３）と、依存パラグラフから単語を抽出する処理（Ｓ１０４）と、各単語について依存パラグラフ内での特徴度をスコアリングする処理（Ｓ１０５）と、特徴度が上位の単語をキーワードとして抽出し、インデックステーブルに格納する処理（Ｓ１０６）とを実行し、指定された検索語に基づいてインデックステーブル内のキーワードを検索し、一致するキーワードが付与された画像を出力する処理を実行する。
【選択図】図１

図1



【特許請求の範囲】
【請求項１】
コンピュータシステムにより、画像と文字列が混在する文書から抽出した画像に対してキーワードを付与し、ユーザから指定された検索語に基づいて前記キーワードを検索し、一致する前記キーワードが付与された画像を出力する文書内画像検索方法であって、
前記コンピュータシステムは、
前記文書を解析して前記文書の中から画像の位置を特定し、前記画像のデータを抽出して格納する画像データ特定処理と、
前記画像データ特定処理で特定した前記画像について、前記文書を解析して前記画像のキャプション領域を特定し、前記キャプション領域から画像名とキャプションとを抽出して前記画像と対応付けて格納するキャプション特定処理と、
前記画像データ特定処理で特定した前記画像について、前記文書を解析して前記文書中で前記画像について記述しているパラグラフである依存パラグラフを特定する依存パラグラフ特定処理と、
前記依存パラグラフ特定処理で特定した前記依存パラグラフから単語を抽出する単語抽出処理と、
前記単語抽出処理で抽出した前記各単語について、前記依存パラグラフ内での特徴度を所定の方法によりスコアリングするスコアリング処理と、
前記スコアリング処理で算出した前記各単語の特徴度が上位の所定の前記単語を前記キーワードとして抽出し、抽出した前記キーワードを対象の前記画像のインデックスとし、その前記特徴度を対象の前記画像に対する適合度として、インデックステーブルに格納するインデックス作成処理とを実行し、
前記ユーザによって指定された前記検索語に基づいて、前記インデックステーブル内の前記キーワードを検索し、一致する前記キーワードが付与された前記画像を出力する画像検索処理を実行することを特徴とする文書内画像検索方法。
【請求項２】
請求項１に記載の文書内画像検索方法において、
前記依存パラグラフ特定処理では、
前記画像データ特定処理で特定した前記画像の前記画像名によって前記文書をサーチし、前記画像名の文字列が最初に出現したパラグラフから、次の画像の画像名の文字列が出現するパラグラフの直前のパラグラフまでを、対象の前記画像についての前記依存パラグラフとして特定することを特徴とする文書内画像検索方法。
【請求項３】
請求項１または２に記載の文書内画像検索方法において、
前記スコアリング処理では、
前記単語抽出処理で抽出した前記各単語について、対象の前記画像の前記依存パラグラフ内での前記単語の出現頻度と、前記文書内の全ての画像の前記依存パラグラフの中での前記単語が出現する前記依存パラグラフの数とに基づいて、前記特徴度をスコアリングすることを特徴とする文書内画像検索方法。
【請求項４】
請求項３に記載の文書内画像検索方法において、
前記スコアリング処理では、
前記単語抽出処理で抽出した前記各単語について、対象の前記画像の前記依存パラグラフ内での前記単語の出現頻度を、対象の前記画像の前記依存パラグラフ内での前記単語の出現位置に基づいて重み付けして算出することを特徴とする文書内画像検索方法。
【請求項５】
請求項１～４のいずれか１項に記載の文書内画像検索方法において、
前記画像検索処理では、
前記検索語に基づいて前記画像を出力する際に、前記インデックステーブルから、前記検索語に一致する前記キーワードと対応する前記画像との前記適合度を取得し、前記適合

度に応じて前記画像の出力方法を制御することを特徴とする文書内画像検索方法。
【請求項６】
画像検索サーバおよび前記画像検索サーバに接続されたクライアント端末を有し、画像と文字列が混在する文書から抽出した画像に対してキーワードを付与し、ユーザから指定された検索語に基づいて前記キーワードを検索し、一致する前記キーワードが付与された画像を出力する文書内画像検索システムであって、
前記画像検索サーバは、
前記文書の中から画像の位置を特定し、前記画像のデータを抽出して格納する画像データ特定部と、
前記画像データ特定部で特定された前記画像について、前記画像のキャプション領域を特定し、前記キャプション領域から画像名とキャプションとを抽出して前記画像と対応付けて格納するキャプション特定部と、
前記画像データ特定部で特定された前記画像について、前記文書中で前記画像について記述しているパラグラフである依存パラグラフを特定する依存パラグラフ特定部と、
前記依存パラグラフ特定部で特定された前記依存パラグラフから単語を抽出する単語抽出部と、
前記単語抽出部で抽出された前記各単語について、前記依存パラグラフ内での特徴度を所定の方法によりスコアリングするスコアリング部と、
前記スコアリング部で算出された前記各単語の特徴度が上位の所定の前記単語を前記キーワードとして抽出し、抽出した前記キーワードを対象の前記画像のインデックスとし、その前記特徴度を対象の前記画像に対する適合度として、インデックステーブルに格納するインデックス作成部と、
前記クライアント端末を利用して前記ユーザによって指定された前記検索語に対して、前記インデックステーブルから前記検索語と一致する前記キーワードに対応する前記画像を取得する検索処理部と、
前記クライアント端末上に表示させる、前記検索語の入力や検索結果の出力のための画面を生成するユーザインタフェース部とを有することを特徴とする文書内画像検索システム。
【請求項７】
請求項６に記載の文書内画像検索システムにおいて、
前記依存パラグラフ特定部は、
前記画像データ特定部で特定された前記画像の前記画像名によって前記文書をサーチし、前記画像名の文字列が最初に出現したパラグラフから、次の画像の画像名の文字列が出現するパラグラフの直前のパラグラフまでを、対象の前記画像についての前記依存パラグラフとして特定することを特徴とする文書内画像検索システム。
【請求項８】
請求項６または７に記載の文書内画像検索システムにおいて、
前記スコアリング部は、
前記単語抽出部で抽出された前記各単語について、対象の前記画像の前記依存パラグラフ内での前記単語の出現頻度と、前記文書内の全ての画像の前記依存パラグラフの中での前記単語が出現する前記依存パラグラフの数とに基づいて、前記特徴度をスコアリングすることを特徴とする文書内画像検索システム。
【請求項９】
請求項８に記載の文書内画像検索システムにおいて、
前記スコアリング部は、
前記単語抽出部で抽出された前記各単語について、対象の前記画像の前記依存パラグラフ内での前記単語の出現頻度を、対象の前記画像の前記依存パラグラフ内での前記単語の出現位置に基づいて重み付けして算出することを特徴とする文書内画像検索システム。
【請求項１０】
請求項６～９のいずれか１項に記載の文書内画像検索システムにおいて、

前記検索処理部は、
前記ユーザによって指定された前記検索語に対して、前記インデックステーブルから前記検索語と一致する前記キーワードに対応する前記画像を前記適合度と合わせて取得し、
前記ユーザインタフェース部は、
前記検索結果の出力のための画面を生成する際に、前記適合度に応じて検索結果の前記画像の出力方法を制御することを特徴とする文書内画像検索システム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、画像と文字列が混在する文書内の画像を検索する技術に関し、特に、画像の内容を表すキーワードを指定することにより画像を検索する文書内画像検索方法および文書内画像検索システムに適用して有効な技術に関するものである。
【背景技術】
【０００２】
近年、ＩＴ技術の進展により、従来は紙などの物理的な媒体によって保存されていた文書等を含む大量の情報が電子化されて保存されるようになってきている。さらに、これらの情報に対してコンピュータを利用して、例えば検索エンジンや検索システム等によって検索して所望の情報を取得し、情報を有効活用するということが行われている。
【０００３】
電子化された文書群に対して、検索語を指定し、文書内のテキスト（もしくは文書の内容を表すキーワードやタグ等）に検索語と一致する文字列を含む文書や、その文書内における位置などを検索することは広く一般的に行われている。一方、電子化された文書には、テキスト情報以外に図や表、写真などの画像も含まれる。この文書内に含まれる画像についてもテキストと同様に検索語を指定することにより検索したいという要望がある。この場合、画像データ自体は文字情報を含まないため、画像に対してその内容を表す文字情報を何らかの手段で付与する必要がある。
【０００４】
これに対して、例えば、特開平８－２０２７３１号公報（特許文献１）には、スキャナにより入力した文書を、画像分離手段により文字領域と画像領域とに分離し、文字領域から文字認識手段により文字列を認識し、認識された文字列から単語分離手段により単語を抽出し、抽出した単語の文書内での出現頻度に基づいて入力画像に付加するキーワードを判定し、キーワードと入力画像をデータベースに登録することにより、オペレータの手を介さずに入力画像にキーワードを付加する技術が開示されている。
【０００５】
また、例えば、特開平１１－２５１１３号公報（特許文献２）には、画像および文字列が混在した文書から画像を抽出して画像ＤＢに格納する際に、文書中から画像について記述した文字列（キーテキスト）を自動的に抽出して画像に関連付けて格納し、入力された検索語に基づいてキーテキストを検索することによって該当する画像を得る技術が開示されている。
【先行技術文献】
【特許文献】
【０００６】
【特許文献１】特開平８－２０２７３１号公報
【特許文献２】特開平１１－２５１１３号公報
【発明の概要】
【発明が解決しようとする課題】
【０００７】
特許文献１に記載されたような画像に対するキーワードの付与方法では、キーワードの抽出に際して、文書中での対象の画像に対する言及などの文脈情報を考慮したものとなっていないため、特に、文書中の画像が複数になった場合には、画像の内容を表した適切な

キーワードを付与することができず、画像検索の際の適合率が低くなるという問題が生じる。

【０００８】

一方、特許文献２に記載されたようなキーテキストの付与方法では、画像について記述した文字列をキーテキストとするため、文脈情報を考慮したキーテキストを画像に付与することができる。また、キーキャプションを使用すればノイズが少ない検索が可能であり、また、キーページを使用すれば広範囲の検索を行うことが可能である。

【０００９】

しかし、逆に、キーキャプションを使用すれば漏れが大きくなり、また、キーページを使用すればノイズが大きくなるため適合率が低くなるという問題が生じる。さらに、画像と関連付けて格納するキーテキストの情報（特にキーページ）がキーワードの場合と比べて格段に大きくなるという問題や、検索時に検索対象のキーテキストの種別を多く指定するほど検索処理に時間を要するという問題を生じる。

【００１０】

そこで本発明の目的は、文書から抽出した画像に対して、画像の内容を表すキーワードを文脈情報を考慮して効率的に精度良く付与することにより、キーワードを利用した高精度で効率の良い画像の検索を可能とする文書内画像検索方法および文書内画像検索システムを提供することにある。本発明の前記ならびにその他の目的と新規な特徴は、本明細書の記述および添付図面から明らかになるであろう。

【課題を解決するための手段】

【００１１】

本願において開示される発明のうち、代表的なものの概要を簡単に説明すれば、以下のとおりである。

【００１２】

本発明の代表的な実施の形態による文書内画像検索方法は、文書を解析して前記文書の中から画像の位置を特定し、前記画像のデータを抽出して格納する画像データ特定処理と、前記画像データ特定処理で特定した前記画像について、前記文書を解析して前記画像のキャプション領域を特定し、前記キャプション領域から画像名とキャプションとを抽出して前記画像と対応付けて格納するキャプション特定処理と、前記画像データ特定処理で特定した前記画像について、前記文書を解析して前記文書中で前記画像について記述しているパラグラフである依存パラグラフを特定する依存パラグラフ特定処理と、前記依存パラグラフ特定処理で特定した前記依存パラグラフから単語を抽出する単語抽出処理と、前記単語抽出処理で抽出した前記各単語について、前記依存パラグラフ内での特徴度を所定の方法によりスコアリングするスコアリング処理と、前記スコアリング処理で算出した前記各単語の特徴度が上位の所定の前記単語を前記キーワードとして抽出し、抽出した前記キーワードを対象の前記画像のインデックスとし、その前記特徴度を対象の前記画像に対する適合度として、インデックステーブルに格納するインデックス作成処理とを実行し、ユーザによって指定された前記検索語に基づいて、前記インデックステーブル内の前記キーワードを検索し、一致する前記キーワードが付与された前記画像を出力する画像検索処理を実行することを特徴とするものである。

【発明の効果】

【００１３】

本願において開示される発明のうち、代表的なものによって得られる効果を簡単に説明すれば以下のとおりである。

【００１４】

本発明の代表的な実施の形態によれば、文書から抽出した画像に対して、画像に関連するキーワードを文脈情報を考慮して効率的に精度良く付与することが可能となり、キーワードを利用した高精度で効率の良い画像の検索が可能となる。

【図面の簡単な説明】

【００１５】

【図１】本発明の実施の形態１における、画像抽出部でのインデックス作成処理の例について説明する図である。
【図２】本発明の実施の形態１である文書内画像検索システムの構成例の概要を示す図である。
【図３】本発明の実施の形態１における、画像情報およびインデックステーブルのデータ構成の例を示した図である。
【図４】本発明の実施の形態１における、依存パラグラフに含まれる各単語のＴＦ×ＩＰＦ値を算出した例を示した図である。
【図５】本発明の実施の形態１における、画像を検索する際にクライアント端末に表示されるユーザインタフェースの例を示した図である。
【図６】本発明の実施の形態２における、依存パラグラフに含まれる各単語のＴＦ×ＩＰＦ値を、単語の出現位置に応じて重み付けして算出した例を示した図である。
【発明を実施するための形態】
【００１６】
以下、本発明の実施の形態を図面に基づいて詳細に説明する。なお、実施の形態を説明するための全図において、同一部には原則として同一の符号を付し、その繰り返しの説明は省略する。
【００１７】
＜実施の形態１＞
本発明の実施の形態１である文書内画像検索システムは、画像と文字列が混在した文書から画像を抽出し、文書中で当該画像について記述しているパラグラフ（依存パラグラフ）内の単語から、特徴度のスコアリングにより上位のものをキーワードとして抽出し、当該キーワードを当該画像に関連するキーワードとして付与してインデックスを作成する。このとき、当該キーワードの特徴度を当該画像に対する適合度とする。また、ユーザにより画像を検索するための検索語が入力されると、インデックスに基づいて検索語と一致するキーワードに対応する画像を取得して画面表示により出力する。このとき、対応する画像が複数ある場合は、適合度に応じて優先付けして画面表示する。
【００１８】
［システム構成］
図２は、本発明の実施の形態１である文書内画像検索システムの構成例の概要を示す図である。文書内画像検索システム１は、例えば、コンピュータシステムによる画像検索サーバ１００とデータベース、および、インターネットや社内ＬＡＮ等のネットワーク５００を介して画像検索サーバ１００に接続された、ＰＣ等のクライアント端末４００から構成される。また、画像検索サーバ１００は、データベースもしくはファイル等により、画像とテキストが混在する複数の文書２００を保持している。
【００１９】
画像検索サーバ１００は、例えば、画像抽出部１１０および画像検索部１２０を有する。また、データベースとして、画像情報３１０およびインデックステーブル３２０を有する。これらのデータベースは、画像検索サーバ１００が直接保持してもよいし、アクセス可能な他のデータベースサーバに保持する構成としてもよい。画像抽出部１１０は、文書２００内の画像を抽出し、キーワードを付与してインデックスを作成する処理を行い、例えば、画像データ特定部１１１、キャプション特定部１１２、依存パラグラフ特定部１１３、単語抽出部１１４、スコアリング部１１５、およびインデックス作成部１１６を有する。
【００２０】
画像データ特定部１１１は、各文書２００を解析して文書２００中の画像の位置を特定し、当該画像のデータを抽出して、画像情報３１０に格納する。キャプション特定部１１２は、画像データ特定部１１１によって特定された各画像について、文書２００を解析して対象の画像のキャプション領域を特定し、当該キャプション領域から画像名とキャプションとを抽出して、画像情報３１０の対象の画像のエントリに格納する。依存パラグラフ

特定部１１３は、画像データ特定部１１１によって特定された各画像について、文書２００を解析して後述する依存パラグラフを特定し、画像情報３１０の対象の画像のエントリに格納する。

【００２１】

単語抽出部１１４は、依存パラグラフ特定部１１３によって特定された依存パラグラフから自然言語処理により単語（複合名詞）を抽出する。スコアリング部１１５は、単語抽出部１１４によって抽出された各単語について、依存パラグラフ内での特徴度を後述する方法によりスコアリングする。インデックス作成部１１６は、スコアリング部１１５によって算出された各単語の特徴度が上位の所定の単語をキーワードとして抽出し、抽出したキーワードを対象の画像のインデックスとし、その特徴度を対象の画像に対する適合度として、インデックステーブル３２０に格納する。

【００２２】

画像検索部１２０は、クライアント端末４００を利用してユーザによって指定された検索語に基づいて、インデックステーブル３２０内のキーワードを検索し、一致するキーワードが付与された画像を出力する画像検索処理を行い、例えば、検索処理部１２１およびユーザインタフェース部１２２を有する。検索処理部１２０は、ユーザによって指定された検索語に対して、インデックステーブル３２０から検索語と一致するキーワードに対応する画像を適合度と合わせて取得する。ユーザインタフェース部１２２は、クライアント端末４００上に表示させる、検索語の入力や検索結果の出力のための画面を生成する。

【００２３】

画像抽出部１１０および画像検索部１２０の各部は、ソフトウェアプログラムとして実現され、例えば、図示しないＷｅｂサーバプログラム上で稼働するアプリケーションプログラムとして実現される。また、ユーザインタフェース部１２２では、例えば、ＨＴＭＬ（HyperText Markup Language）によって画面を生成し、図示しないＷｅｂサーバプログラムを介して、クライアント端末４００上の図示しないＷｅｂブラウザによって表示させる。

【００２４】

文書２００は、例えば、ワードプロセッサ等のアプリケーションプログラムで作成された画像を含むテキスト文書や、ＨＴＭＬ等のタグ文書など、画像抽出部１１０により画像と文字列の認識が可能である電子化された文書であれば取り扱うことが可能である。なお、紙媒体の文書であっても、例えば、特許文献１、２等に記載されているように、スキャナによって紙媒体の文書を読み取り、読み取ったデータに基づいて文字領域と画像領域とを識別し、文字領域についてはＯＣＲ（Optical Character Reader）等により文字認識を行うことによって文書２００として取り込むことが可能である。

【００２５】

［データ構成］

図３は、画像情報３１０およびインデックステーブル３２０のデータ構成の例を示した図である。画像情報３１０は、例えば、画像ＩＤ３１１、画像データ３１２、文書名３１３、位置３１４、画像名３１５、キャプション３１６、および依存パラグラフ３１７の項目を有し、文書２００から抽出された画像に関する情報を保持する。

【００２６】

画像ＩＤ３１１は、文書内画像検索システム１内で対象の画像を一意に特定するために付与されるＩＤである。画像データ３１２は、文書２００から抽出された画像のバイナリデータである。文書名３１３および位置３１４は、対象の画像が含まれる文書２００の文書名および文書２００内の位置（行数）である。画像名３１５およびキャプション３１６は、対象の画像の画像名およびキャプションである。依存パラグラフ３１７は、対象の画像の依存パラグラフの文字列である。

【００２７】

インデックステーブル３２０は、例えば、キーワード３２１、画像ＩＤ３２２、および適合度３２３の項目を有し、ユーザから指定された検索語によって画像検索部１２０にお

いて画像を検索する際に利用するインデックスを保持する。キーワード３２１は、文書２００から抽出した各画像に対して画像抽出部１１０での処理によって付与されたキーワードである。画像ＩＤ３２２は、対象のキーワードが付与された画像のＩＤである。適合度３２３は、対象のキーワードの対象の画像に対する適合度を示すスコアである。なお、インデックステーブル３２０は、データベースに限らずファイル形式であってもよい。また、画像情報３１０およびインデックステーブル３２０の各項目は上記のものに限らず、他の項目を有していてもよい。

【００２８】

［インデックス作成処理］

図１は、本実施の形態の画像抽出部１１０でのインデックス作成処理の例について説明する図である。まず、画像データ特定部１１１により、対象の文書２００を解析して文書２００の中から画像の位置を特定し、当該画像データを抽出して画像情報３１０に格納する、画像データ特定処理を行う（ステップＳ１０１）。画像の位置の特定については、例えば、特許文献１や特許文献２に記載されているような方法をとることができる。図１では、画像Ａ、Ｂ、Ｃの３つの画像を特定した場合の例を示している。

【００２９】

なお、抽出した画像にはＩＤを付与し、画像のバイナリデータと合わせて、画像情報３１０の画像ＩＤ３１１および画像データ３１２にそれぞれ格納する。また、当該画像が含まれる文書２００の文書名および文書２００内の位置（行数）を、画像情報３１０の文書名３１３および位置３１４にそれぞれ格納する。

【００３０】

次に、キャプション特定部１１２により、ステップＳ１０１で特定した画像について画像のキャプション領域を特定し、キャプション領域から画像名とキャプションとを抽出して画像情報３１０の該当の画像のエントリに格納する、キャプション特定処理を行う（ステップＳ１０２）。ここで、キャプション領域とは、図や表などの画像についての短い説明が記載された領域であり、例えば、「図１」や「表２」などの画像名と、「インデックス作成処理の例について説明する図」などの画像に対して付与された文字列であるキャプションから構成される。

【００３１】

キャプション領域の特定については、例えば、特許文献２に記載されているような方法をとることができる。ここで、例えば、学術論文などの文書では、一般的に画像が図である場合にはキャプション領域は画像の下部に配置され、画像が表である場合には画像の上部に配置される。従って、画像の上部および下部の所定の小領域をキャプション領域として特定する。キャプション領域から画像名とキャプションを特定する際には、例えば、キャプション領域内の文字列から画像名に相当する文字列を判定するための正規表現を用いて画像名を特定し、その後に続く１文をキャプションとして特定する方法をとることができる。

【００３２】

画像名に相当する文字列を判定するための正規表現としては、例えば、「図￥ｄ＊」、「表￥ｄ＊」、「図表￥ｄ＊」、「グラフ￥ｄ＊」などを用いることができる。これらの正規表現は、予め定義してファイル等に保持しておく。図１では、画像Ａについては「図１　Ａ」というキャプション領域（画像名「図１」、キャプションｃ１「Ａ」）、画像Ｂ、Ｃについては「図２　Ｂ、Ｃ」というキャプション領域（画像名「図２」、キャプションｃ２「Ｂ、Ｃ」）を特定した場合の例を示している。なお、抽出した画像名およびキャプションは、画像情報３１０の該当の画像のエントリの画像名３１５およびキャプション３１６にそれぞれ格納する。

【００３３】

次に、依存パラグラフ特定部１１３により、ステップＳ１０１で特定した画像について、文書２００中で当該画像について記述しているパラグラフである依存パラグラフを特定する、依存パラグラフ特定処理を行う（ステップＳ１０３）。依存パラグラフの特定につ

いては、例えば、ステップＳ１０２で特定した画像の画像名によって文書２００をサーチし、画像名の文字列が最初に出現したパラグラフから、次の画像の画像名の文字列が出現するパラグラフの直前のパラグラフまでを、対象の画像についての依存パラグラフとして特定する。

【００３４】

図１では、パラグラフｐ１～ｐ５のうち、画像Ａについての依存パラグラフＰ１として、画像Ａの画像名である「図１」が最初に出現するパラグラフｐ２から、次の画像である画像Ｂ、Ｃの画像名である「図２」が出現するパラグラフｐ５の直前のパラグラフｐ４までを特定した場合の例を示している。なお、抽出した依存パラグラフ内の文字列は、画像情報３１０の該当の画像のエントリの依存パラグラフ３１７に格納する。

【００３５】

ここで、実際は、パラグラフｐ５以降にも画像Ａ（「図１」）についての記述がされているパラグラフが存在する場合も想定される。しかし、これらのパラグラフについては、特定するのに多くの処理を要するのに比して、その記述内容と画像Ａとの直接の関連度はそれほど高くない場合が多く、これらのパラグラフから取得されるキーワードの画像Ａとの適合度は低い場合が多い。また、記述内容と画像Ａとの関連度がある場合であっても、記述内容が依存パラグラフと同じような内容である等により、適合度が高いキーワードを独自に抽出できるケースはそれほど多くない。従って、本実施の形態では、上述したように、次の画像名が出現するまでのパラグラフを依存パラグラフとすることで、効率良く十分な精度のキーワードが抽出できる依存パラグラフの特定を可能とする。

【００３６】

なお、ステップＳ１０２において画像名とキャプションが特定できなかった場合（画像にキャプション領域がない場合や、キャプション領域を有していても正規表現と一致する画像名がない場合など）は、ステップＳ１０３以降の処理は行わず、キーワードを付与しないようにしてもよいし、例えば、特許文献２に記載されているような方法やその他の方法により、依存パラグラフに相当するパラグラフを特定するようにしてもよい。

【００３７】

次に、単語抽出部１１４により、ステップＳ１０３で特定した依存パラグラフから自然言語処理によって単語（複合名詞）を抽出する、単語抽出処理を行う（ステップＳ１０４）。ここでは、例えば、一般的な形態素解析により依存パラグラフから複合名詞を抽出する。なお、ステップＳ１０２で特定したキャプションは、処理の便宜上、例えば、依存パラグラフの０段落目（先頭）に相当するものとして依存パラグラフに含めるものとし、同様に形態素解析を行って複合名詞を抽出する。

【００３８】

図１では、画像Ａ（「図１」）について、依存パラグラフＰ１（キャプションｃ１、およびパラグラフｐ２～ｐ４）からそれぞれ、ｗ１～ｗ５の各単語（複合名詞）を抽出した場合の例を示している。ここで、例えばパラグラフｐ３で単語ｗ２が２つ抽出されているのは、単語ｗ２がパラグラフｐ３で２回出現していることを示している。

【００３９】

次に、スコアリング部１１５により、ステップＳ１０４で抽出した各単語について、依存パラグラフ内での特徴度を所定の方法によりスコアリングする、スコアリング処理を行う（ステップＳ１０５）。ここでは、各単語について、後述するＴＦ×ＩＰＦ値（Term Frequency×Inversed Paragraph Frequency）を算出して特徴度とする。図１では、単語ｗ１～ｗ５について、それぞれＴＦ×ＩＰＦ値を算出して特徴度とした場合の例を示している。

【００４０】

次に、画像抽出部１１０により、ステップＳ１０５で算出した各単語の特徴度が上位の所定の単語をキーワードとして抽出し、抽出したキーワードを対象の画像のインデックスとし、その特徴度を対象の画像に対する適合度として、インデックステーブル３２０に格納する、インデックス作成処理を行う（ステップＳ１０６）。

【００４１】
図１では、単語ｗ１～ｗ５のうち、例えば、特徴度が平均値以上である単語ｗ１、ｗ３、ｗ４の３つをキーワードとして抽出し、画像Ａのインデックスとした場合の例を示している。なお、ここでは平均値以上の特徴度を有する単語をキーワードとして抽出しているが、例えば、特徴度が上位から所定の順位のものまでを抽出するなど他の方法であってもよい。また、キャプションに含まれる単語は、画像に直接的に言及しているということから、特徴度のスコアに関わりなくキーワードとして抽出するようにしてもよい。
【００４２】
抽出したキーワードと対応する画像のＩＤ、およびその適合度は、インデックステーブル３２０のキーワード３２１、画像ＩＤ３２２、および適合度３２３にそれぞれ格納する。以上の処理により、文書２００群から画像を抽出し、抽出した画像に対してキーワードを効率的に精度良く付与して、適合度と合わせてインデックス化したインデックステーブル３２０を生成することができる。
【００４３】
［スコアリングとキーワード抽出］
以下では、スコアリング部１１５におけるスコアリング処理（ステップＳ１０５）、および、インデックス作成部１１６におけるインデックス作成処理（ステップＳ１０６）について説明する。スコアリング処理（ステップＳ１０５）では、図１の単語抽出処理（ステップＳ１０４）にて抽出された依存パラグラフ内の各単語について、特徴度としてＴＦ×ＩＰＦ値を算出してスコアリングする。ＴＦ×ＩＰＦ値とは、ＴＦ（Term Frequency）値とＩＰＦ（Inverse Paragraph Frequency）値の積である。
【００４４】
ＴＦ値およびＩＤＦ（Inversed Document Frequency）値を用いてある文書中の特徴的な単語（重要とみなされる単語）を抽出することは一般的に行われている。本実施の形態のスコアリング処理でもこの手法を適用して特徴度を算出するが、本実施の形態では、ＴＦ値およびＩＤＦ値の算出時における単位である「文書（Document）」を依存パラグラフＰとした、ＴＦ値およびＩＰＦ値を用いて特徴度を算出する。なお、特徴度の算出手法はこれに限るものではなく、単語毎に数値として画像との適合度を評価することが可能な手法であれば利用することができる。
【００４５】
本実施の形態のスコアリング処理において、ＴＦ値は、依存パラグラフＰ内における各単語（複合名詞）の出現頻度であり、この値が大きいほど当該単語は依存パラグラフＰ（すなわち対応する画像）の特徴をよく表しているものと考えられる。ある依存パラグラフ$Ｐ_{j}$における単語$ｗ_{i}$のＴＦ値は、例えば、依存パラグラフ$Ｐ_{j}$内の単語$ｗ_{i}$の出現頻度を、依存パラグラフ$Ｐ_{j}$において出現する延べ単語数で正規化して以下の式で表される。
【００４６】
【数１】

$$tf_{w_i} = \frac{|w_i|}{\sum_n |w_n|}$$

$tf_{w_i}$ ：単語$w_i$のTF(Term Frequency)値

$|w_i|$ ：単語$w_i$の依存パラグラフ$P_j$内での出現頻度

$\sum_n |w_n|$ ：依存パラグラフ$P_j$に出現する延べ単語数

【００４７】
一方、ＴＦ値が大きい単語であっても、他の画像についての依存パラグラフＰにも頻繁

に出現する単語は、特定の依存パラグラフＰの特徴を表す単語ではない一般的な単語である場合が多い。ここで、ＩＰＦ値は、対象の単語が出現する依存パラグラフＰの数の逆数であり、この値が大きいほどこの単語が出現する依存パラグラフＰの数が少ない。すなわち、この単語は特定の依存パラグラフＰの特徴をよく表しているものと考えられる。ある単語ｗ$_{i}$のＩＰＦ値は、例えば、単語ｗ$_{i}$が出現する依存パラグラフＰの数の逆数を、対象の文書２００内の全ての依存パラグラフＰの数で正規化して以下の式で表される。

【００４８】

【数２】

$$ipf_{w_i} = \log \frac{|P|}{|w_i \in P|}$$

$ipf_{w_i}$ ：単語$w_i$のIPF(Inversed Paragraph Frequency)値

$|P|$ ：依存パラグラフ$P$の総数

$|w_i \in P|$ ：単語$w_i$を含む依存パラグラフ$P$の数

【００４９】

上記のＴＦ値とＩＰＦ値の両者の値が大きい単語ｗ$_{i}$が、対象の依存パラグラフＰ（すなわち対応する画像）の特徴を真によく表していると考えられるため、ＴＦ値とＩＰＦ値の積であるＴＦ×ＩＰＦ値を算出して、これを単語ｗ$_{i}$の特徴度のスコアとする。このＴＦ×ＩＰＦ値を、依存パラグラフＰ内の各単語について算出する。ＴＦ×ＩＰＦ値が大きい単語は、対象の画像の内容をよく表しており、キーワードとしての適合度が高いものと考えられる。なお、上記のＴＦ値、ＩＰＦ値の算出式については一例であり、精度や処理時間などに応じて正規化や対数計算の式などを適当なものにすることができる。

【００５０】

図４は、図１の例に示した画像Ａについて、その依存パラグラフＰ１に含まれる各単語のＴＦ×ＩＰＦ値を算出した例を示した図である。画像Ａについての依存パラグラフＰ１（キャプションｃ１およびパラグラフｐ２～ｐ４）に含まれる単語ｗ１～ｗ５について、依存パラグラフＰ１内での出現頻度に基づいて数１により算出したＴＦ値と、出現した依存パラグラフＰの数に基づいて数２により算出したＩＰＦ値、およびＴＦ×ＩＰＦ値のスコア（特徴度）が示されている。この特徴度に基づいて、インデックス作成処理（ステップＳ１０６）では、例えば、各単語の特徴度がその平均値（０．１６２）以上である単語ｗ１、ｗ３、ｗ４の３つをキーワードとして抽出する。これにより、画像Ａの内容をよく表した精度の高いキーワードを抽出することができる。

【００５１】

［画像検索処理］

以下では、ユーザがクライアント端末４００を利用して文書２００群に含まれる画像を検索する際のユーザインタフェースおよび画像検索部１２０での画像検索処理について説明する。図５は、画像を検索する際にクライアント端末４００に表示されるユーザインタフェースの例を示した図である。当該画面は、上述したように、画像検索部１２０のユーザインタフェース部１２２によって、例えば、ＨＴＭＬによって生成され、図示しないＷｅｂサーバプログラムを介して、クライアント端末４００上の図示しないＷｅｂブラウザによって表示される。

【００５２】

図５に示した画面の上部には、例えば、ユーザが画像を検索するための検索語を指定することができるフィールドを有する。当該フィールドにはユーザが検索語を複数指定することも可能である。なお、本実施の形態では、画像検索サーバ１００のインデックステー

ブル３２０にキーワードのリストを有しているため、これを参照することにより、ユーザが検索語を入力している途中であっても、途中まで入力された文字列に一致するキーワードの候補を「単語候補」のフィールドに表示することが可能である。ユーザは、表示された候補の中から所望のキーワードをマウスによるクリック等で選択して、検索語として確定させることができる。

【００５３】

このキーワードの候補を表示する処理は、Ｇｏｏｇｌｅ（登録商標）等のＷｅｂサイトで一般的に行われているように、例えば、当該画面コンテンツにＡｊａｘ（Asynchronous JavaScript（登録商標） + XML）等を利用したモジュールを組み込み、当該モジュールが、ユーザが入力した検索語の文字列を取得して画像検索サーバ１００に非同期で送信し、画像検索サーバ１００では、画像検索部１２０の検索処理部１２１によりインデックステーブル３２０を検索することによって、入力された文字列を先頭に含むキーワード３２１のリストを取得してクライアント端末４００に送信し、クライアント端末４００によって「単語候補」のフィールドに表示することで実現することができる。

【００５４】

図５では、ユーザが検索語として「２０」まで入力した時点で、これに該当するキーワード（検索語の候補）として、「２０１０年」、「２０６０年」、「２０世紀」の単語をそれぞれインデックステーブル３２０から取得して「単語候補」フィールドに表示した場合の例を示している。これにより、ユーザが検索語としてキーワードを指定する際の労力を大幅に低減させることができる。

【００５５】

検索語が確定すると、確定した複数の検索語のＡＮＤ条件で、検索語に一致するキーワードが付与された文書２００群内の画像を「画像一覧」のフィールドに表示する。ここでは、例えば、上述のようなＡｊａｘ等を利用したモジュールや、検索ボタン等の押下に伴う処理によって、確定した複数の検索語を画像検索サーバ１００に送信する。

【００５６】

画像検索サーバ１００は、検索処理部１２１によりインデックステーブル３２０を検索し、受信した検索語のＡＮＤ条件により該当する画像ＩＤ３２２を取得する。さらに、画像情報３１０から、対応する画像ＩＤ３１１のエントリの画像データ３１２や、文書名３１３、位置３１４、画像名３１５、キャプション３１６等の他の情報を取得してクライアント端末４００に送信し、クライアント端末４００によって「画像一覧」フィールドに画像データや他の情報を表示する。

【００５７】

なお、複数の画像を表示する際に、例えば、画像検索サーバ１００からクライアント端末４００に送信する画像の検索結果の情報に、キーワードと画像との適合度の情報をインデックステーブル３２０の適合度３２３から取得して追加することができる。これにより、クライアント端末４００では、例えば、画像のキーワードに対する適合度の値（複数のキーワードに対応する場合はその合計）が大きいものを、表示順序を上位にしたり、視覚的に目立つようにしたりなど優先的に表示し、画像とキーワードとの適合度に応じて出力方法を柔軟に制御することが可能となる。

【００５８】

また、表示する画像の適合度の閾値をユーザにより設定できるようにしておき、閾値未満の適合度の画像は表示しない（もしくは画像を検索する際の対象から除外する）ようにしてもよい。例えば、依存パラグラフＰが十分な長さを有しておらず短い場合や、依存パラグラフＰ内のどの単語もあまり特徴的ではなく、各単語のＴＦ×ＩＰＦ値が近似する（ＴＦ×ＩＰＦ値の分散が小さい）場合などは、ＴＦ×ＩＰＦ値が小さくなる傾向が高い。この場合、これらの単語はキーワードとしての精度が低いため、閾値を調整することによって対応する画像が表示されないようにすることができる。

【００５９】

以上のように、本実施の形態の文書内画像検索システム１によれば、文書２００から抽

出した画像に対して、当該画像について記述している依存パラグラフＰを特定することで、画像に関連するキーワードを文脈情報を考慮して効率的に精度良く付与することが可能となり、キーワードを利用した高精度で効率の良い画像の検索が可能となる。また、各キーワードと画像の組合せに対してスコア（適合度）を有するため、適合度に応じて検索結果の画像の表示順序等の出力方法を制御することによってユーザの利便性を高めることが可能となる。

【００６０】

＜実施の形態２＞

本発明の実施の形態２である文書内画像検索システムは、上述した実施の形態１の文書内画像検索システム１において、スコアリング部１１５での特徴度のスコアリング処理（ステップＳ１０５）で、依存パラグラフＰ内の各単語の特徴度をスコアリングする際に、単語の出現位置の情報に基づいて重み付けを行うことによって、抽出するキーワードの精度をより高くすることを可能とするものである。なお、スコアリング部１１５以外の他の構成や処理内容は、実施の形態１で説明したものと同様であるため、再度の説明は省略する。

【００６１】

図６は、図１の例に示した画像Ａについて、その依存パラグラフＰ１に含まれる各単語のＴＦ×ＩＰＦ値を、単語の出現位置に応じて重み付けして算出した例を示した図である。まず、依存パラグラフＰ１内で出現する各単語（ｗ１～ｗ５）を、依存パラグラフＰ１内で出現した行に応じて図６の中段の表に示すように集計する。このとき、例えば、キャプション中の単語は０行目に出現したものとし、依存パラグラフＰ１内の各パラグラフ（ｐ２～ｐ４）を連結して１行目からカウントするものとする。

【００６２】

ここで、行の値をｘ、重み付け値をｙとした重み付け関数ｙ＝ｆ（ｘ）を利用して各行での重み付け値を算出する。図６では、０行目で重み付け値が１であり、２０行目で０となる、傾きマイナス０．０５の一次関数によって重み付け値を算出している。これは、キャプションを始めとして、依存パラグラフＰ内の先頭に近い位置で出現した単語ほど、対応する画像に対して直接的に言及している場合が多いことを考慮した重み付け関数である。

【００６３】

この重み付け関数によれば、キャプションに含まれる単語を無条件に抽出するという処理を行わなくても、これらの単語には自動的に大きい重み付け値を付与することができるため、キーワードとして抽出されるようにすることができる。なお、重み付け関数は、図６に示したものに限らず、例えば、対象の画像の出現行を中心とした正規分布曲線を有する確率密度関数など、単語の出現位置による画像との適合度のモデルに基づいて種々のものを用いることができる。

【００６４】

この重み付け値に基づいて、各単語の出現頻度の値を図６の下段の表に示すように補正する。例えば、単語ｗ１は、０行目（重み付け値１．００）で１回、１行目（重み付け値０．９５）で１回出現しているため、重み補正後の出現頻度は、

１×１．００＋１×０．９５＝１．９５

となる。また、単語ｗ２は、１行目（重み付け値０．９５）で１回、５行目（重み付け値０．７５）で１回、６行目（重み付け値０．７０）で１回出現しているため、重み補正後の出現頻度は、

１×０．９５＋１×０．７５＋１×０．７０＝２．４０

となる。以下、単語ｗ３～ｗ５についても同様に算出する。

【００６５】

以上のように算出された重み補正後の出現頻度に基づいて、上述した数１により重み補正後のＴＦ値を算出し、実施の形態１の場合と同様に数２により算出したＩＰＦ値と乗算することで、重み補正後のＴＦ×ＩＰＦ値を算出する。図６の例では、重み補正後のＴＦ

×ＩＰＦ値の平均値は０．１７４となり、平均値以上の単語をキーワードとして抽出すると、単語ｗ１、ｗ３の２つとなる。実施の形態１の場合と比較して単語ｗ４がキーワードとして抽出されなくなっているが、これは、単語ｗ４は依存パラグラフＰ１の中で後半部分に多く出てきていることから、上述の処理により特徴度（画像との適合度）が相対的に低いものと判断されるためである。

【００６６】

以上のように、本実施の形態の文書内画像検索システム１によれば、依存パラグラフＰ内の各単語の特徴度をスコアリングする際に、単語の出現位置を変数とした重み付け関数を利用して出現頻度に重み付けを行うことによって、単語の出現位置による画像についての言及の程度の違いを考慮して特徴度をスコアリングする。これにより、抽出するキーワードの精度をより高くし、画像検索の際のノイズを低減することが可能となる。

【００６７】

以上、本発明者によってなされた発明を実施の形態に基づき具体的に説明したが、本発明は前記実施の形態に限定されるものではなく、その要旨を逸脱しない範囲で種々変更可能であることはいうまでもない。

【産業上の利用可能性】

【００６８】

本発明は、画像の内容を表すキーワードを指定することにより文書内の画像を検索する文書内画像検索方法および文書内画像検索システムに利用可能である。

【符号の説明】

【００６９】

１…文書内画像検索システム、

１００…画像検索サーバ、１１０…画像抽出部、１１１…画像データ特定部、１１２…キャプション特定部、１１３…依存パラグラフ特定部、１１４…単語抽出部、１１５…スコアリング部、１１６…インデックス作成部、１２０…画像検索部、１２１…検索処理部、１２２…ユーザインタフェース部、

２００…文書、

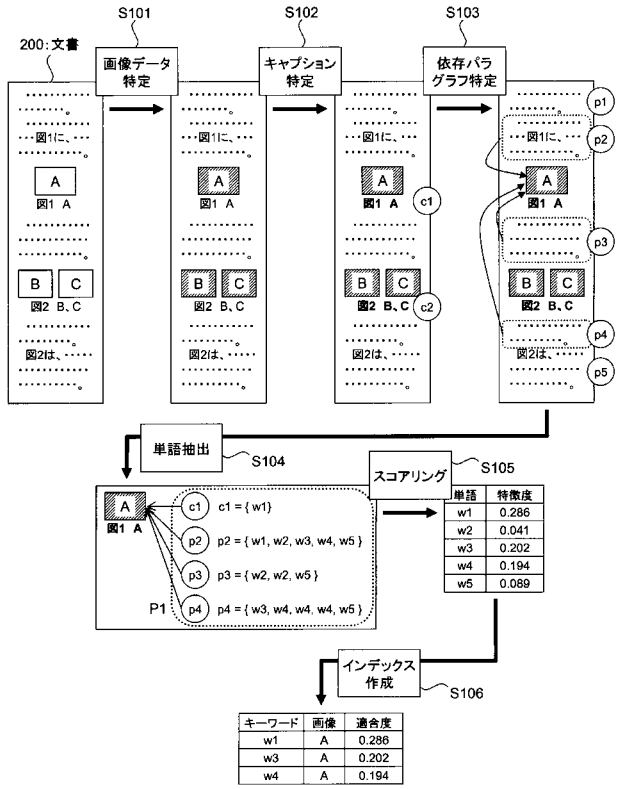
３１０…画像情報、３１１…画像ＩＤ、３１２…画像データ、３１３…文書名、３１４…位置、３１５…画像名、３１６…キャプション、３１７…依存パラグラフ、３２０…インデックステーブル、３２１…キーワード、３２２…画像ＩＤ、３２３…適合度、

４００…クライアント端末、

５００…ネットワーク。

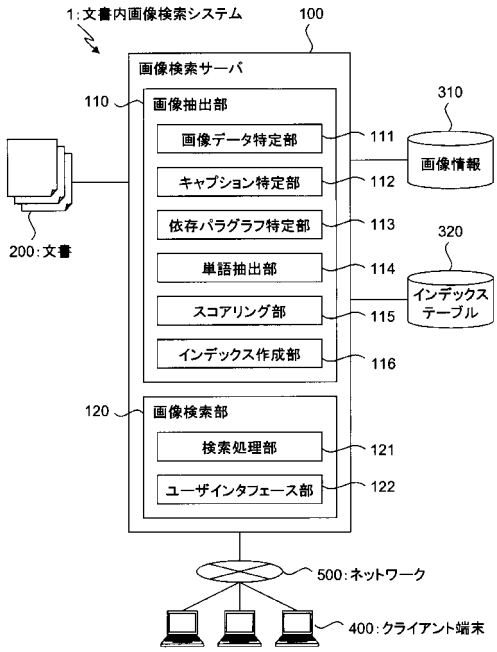
【図１】

図1



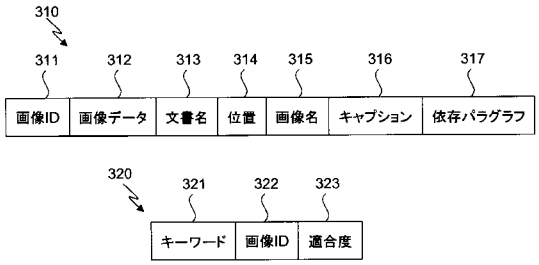

【図２】

図2



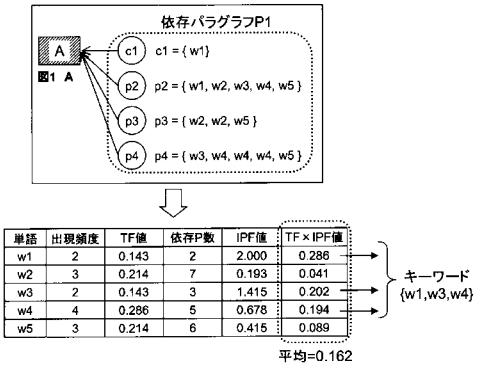

【図３】

図3



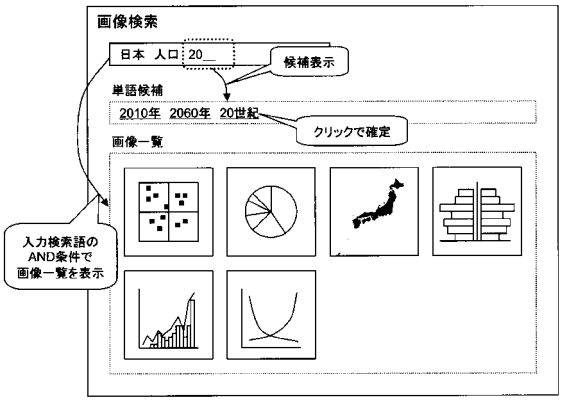

【図４】

図4



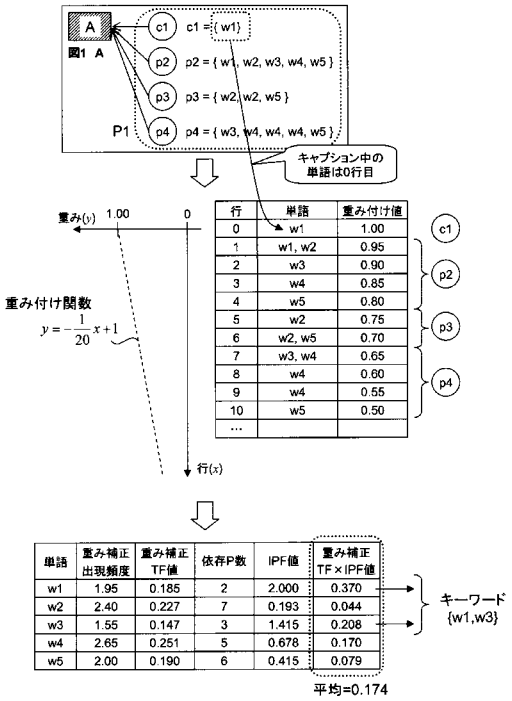

【図５】

図5



【図６】

*図6*




| 単語 | 重み補正出現頻度 | 重み補正TF値 | 依存P数 | IPF値 | 重み補正TF×IPF値 |
|---|---|---|---|---|---|
| w1 | 1.95 | 0.185 | 2 | 2.000 | 0.370 |
| w2 | 2.40 | 0.227 | 7 | 0.193 | 0.044 |
| w3 | 1.55 | 0.147 | 3 | 1.415 | 0.208 |
| w4 | 2.65 | 0.251 | 5 | 0.678 | 0.170 |
| w5 | 2.00 | 0.190 | 6 | 0.415 | 0.079 |

キーワード
{w1,w3}

平均=0.174