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

特許第4904920号
(P4904920)

(45)発行日　平成24年3月28日(2012.3.28)　(24)登録日　平成24年1月20日(2012.1.20)

(51)Int.Cl.　　F I
G06F 17/21 (2006.01)　　G06F 17/21 538M
G06F 17/30 (2006.01)　　G06F 17/30 220C
　　G06F 17/30 170A

請求項の数 4　(全 20 頁)

(21)出願番号　特願2006-143005 (P2006-143005)
(22)出願日　平成18年5月23日 (2006.5.23)
(65)公開番号　特開2007-316743 (P2007-316743A)
(43)公開日　平成19年12月6日 (2007.12.6)
審査請求日　平成21年1月22日 (2009.1.22)

(73)特許権者　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(74)代理人　100089118
弁理士　酒井　宏明
(72)発明者　遠藤　進
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(72)発明者　馬場　孝之
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(72)発明者　椎谷　秀一
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

最終頁に続く

(54)【発明の名称】雛形文書作成プログラム、雛形文書作成方法および雛形文書作成装置

(57)【特許請求の範囲】
【請求項１】
コンピュータに、
複数の文書に対して、各文書の中からページの構成画面が類似している連続したページを部分文書として抽出し、
抽出した部分文書を該部分文書の抽出元の文書と対応付けて記憶装置に保存しておき、
入力された複数の検索条件について、前記記憶装置から当該検索条件を満たす部分文書をそれぞれ検索し、
前記各検索条件について、当該検索条件で検索された前記部分文書の中から、前記複数の検索条件のうち当該検索条件とは異なる他の検索条件を満たす部分文書と所定の組み合わせ条件を満たす部分文書を選択し、
前記各検索条件について選択された前記部分文書をつなぎ合わせて雛形文書を作成する、
処理を実行させるための雛形文書作成プログラム。
【請求項２】
前記所定の組み合わせ条件は、
前記各検索条件を満たす部分文書と前記他の検索条件を満たす部分文書との抽出元の文書が同一であること、
前記各検索条件を満たす部分文書と前記他の検索条件を満たす部分文書とに用いられている単語もしくは語尾の用法の少なくとも一方が使われていること、または、

前記各検索条件を満たす部分文書と前記他の検索条件を満たす部分文書との画面構成が類似していること、
の少なくとも１つを含むことを特徴とする請求項１に記載の雛形文書作成プログラム。
【請求項３】
複数の文書に対して、各文書の中からページの構成画面が類似している連続したページを部分文書として抽出する部分文書抽出部と、
前記部分文書抽出部によって抽出された部分文書を該部分文書の抽出元の文書と対応付けて記憶装置に保存する部分文書情報保存部と、
入力された複数の検索条件について、前記記憶装置から当該検索条件を満たす部分文書を検索する部分文書検索部と、を備え、
前記各検索条件について、当該検索条件で検索された前記部分文書の中から、前記複数の検索条件のうち当該検索条件とは異なる他の検索条件を満たす部分文書と所定の組み合わせ条件を満たす部分文書を選択し、前記各検索条件について選択した前記部分文書をつなぎ合わせて雛形文書を作成する雛形文書作成装置。
【請求項４】
雛形文書作成装置を用いて、
複数の文書に対して、各文書の中からページの構成画面が類似している連続したページを部分文書として抽出し、
抽出した部分文書を該部分文書の抽出元の文書と対応付けて記憶装置に保存しておき、
入力された複数の検索条件について、前記記憶装置から当該検索条件を満たす部分文書をそれぞれ検索し、
前記各検索条件について、当該検索条件で検索された前記部分文書の中から、前記複数の検索条件のうち当該検索条件とは異なる他の検索条件を満たす部分文書と所定の組み合わせ条件を満たす部分文書を選択し、
前記各検索条件について選択された前記部分文書をつなぎ合わせて雛形文書を作成する、
雛形文書作成方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、作成され蓄積されている文書の中から、新規文書の作成に利用できそうな部分を探し出し、容易に再利用可能とする部分文書検索装置に関する。
【背景技術】
【０００２】
従来、オフィス等で多くの文書が作成されており、蓄積されている。新規文書を作成する際には、既存の文書を一部修正したり、既存の文書から一部を抜き出して利用したりといった再利用が行われている。
【０００３】
これにより、新規文書の作成の負荷が軽減されると同時に、質の良い文書を元に文書を作成することで文書の質の向上も望める。
【０００４】
蓄積されている既存の文書から目的の文書を探し出すことを容易にするため、蓄積されている文書から属性情報や検索式などに基づいて抽出した部分集合の文書から、単語を抽出してその出現頻度を求め、この出現頻度に基づいて単語をランキング付けして、その一部を関連キーワードとして検索条件として検索を行うことが提案されている（例えば、特許文献１参照）。
【０００５】
また、文書内の図や表などのオブジェクトのサムネイル画像や文書内での位置情報を保存しておき、ユーザの指定する表や図形といったオブジェクトの種類に応じたオブジェクトのサムネイル画像の一覧を表示し、この一覧から選択されたオブジェクトを表示して再

利用させることが提案されている（例えば、特許文献２参照）。
【特許文献１】特開平１１－２５１０８号公報
【特許文献２】特開２００１－２７３３１４号公報
【発明の開示】
【発明が解決しようとする課題】
【０００６】
文書を再利用する場合には、文書全体が必要なわけではなく、その一部だけが必要になることが多い。しかしながら、特許文献１に記載のようなものでは、検索された文書からもう一度検索して必要箇所を探し、さらにその部分をコピーして新しい文書に貼り付けるなどしなければならず、手間がかかっていた。
【０００７】
また、特許文献２に記載のようなものでは、テキストは対象となっておらず、テキストの部分の再利用をするには、特許文献１のようにしなければならない。
【０００８】
また、再利用したいものが、ある事柄を数行程度で表した文章とか、ある事柄を３ページ程度で表したプレゼンテーション資料のように長さを限定したい場合、文書サイズや部分要約では目的の部分の長さまでは分からないので、検索された文書の対象の部分を確認しなければならなかった。
【０００９】
そこで、本発明は、文書を容易に再利用することができる文書検索装置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】
上記課題を解決する発明は、文書の中から関連する記載の範囲を抽出して部分文書とし、抽出した部分文書の情報を保存しておき、該保存された情報から入力された検索条件に一致する前記部分文書を検索して出力するものである。
【００１１】
この発明では、文書から該文書の一部分である部分文書が抽出されて保存され、この部分文書から検索条件に一致するものが検索され出力される。したがって、検索された部分文書をそのまま利用することができる。
【００１２】
ここで、単語の出現頻度が類似する連続した段落またはページを前記関連する記載の範囲とすることとした。
【００１３】
このようにすれば、単語の出現頻度が類似する連続した段落またはページが部分文書とされ、関連する記載の範囲を良好に抽出することができる。
【００１４】
また、言葉を説明する形式になっている段落またはページを前記関連する記載の範囲とすることとした。
【００１５】
このようにすれば、言葉を説明する形式になっている段落またはページが部分文書とされ、関連する記載の範囲を良好に抽出することができる。
【００１６】
また、再利用された回数が設定された回数より多い範囲を前記関連する記載の範囲とすることとした。
【００１７】
このようにすれば、再利用された回数が設定された回数より多い範囲が部分文書とされ、関連する記載の範囲を良好に抽出することができる。
【００１８】
また、各ページにタイトルが付けられている文書において、前記タイトルの内容に関連

性がある連続したページを前記関連する記載の範囲とすることとした。
【００１９】
このようにすれば、タイトルの内容に関連性がある連続したページが部分文書とされ、関連する記載の範囲を良好に抽出することができる。
【００２０】
また、ページの画面構成が類似している連続したページを前記関連する記載の範囲とすることとした。
【００２１】
このようにすれば、ページの画面構成が類似している連続したページが部分文書とされ、関連する記載の範囲を良好に抽出することができる。
【００２２】
また、ページの色構成が類似している連続したページを前記関連する記載の範囲とすることとした。
【００２３】
このようにすれば、ページの色構成が類似している連続したページが部分文書とされ、関連する記載の範囲を良好に抽出することができる。
【００２４】
また、前記部分文書の情報として部分文書の長さを保存し、前記検索条件として部分文書の長さを入力させることとした。
【００２５】
このようにすれば、部分文書の長さも含めた検索条件で部分文書が検索され、必要な長さの部分文書を検索することができる。
【発明の効果】
【００２６】
本発明によれば、文書から該文書の一部分である部分文書を抽出して保存し、この部分文書から検索条件に一致するものを検索して出力する。したがって、検索された部分文書をそのまま利用することができ、文書を容易に再利用することができる。
【００２７】
また、関連する記載の範囲として、単語の出現頻度が類似する連続した段落またはページや、言葉を説明する形式になっている段落またはページや、再利用された回数が設定された回数より多い範囲や、各ページにタイトルが付けられている文書において前記タイトルの内容に関連性がある連続したページや、ページの画面構成が類似する連続したページや、ページの色構成が類似する連続したページを部分文書とすれば、関連する記載の範囲を良好に抽出することができ、文書を容易に再利用することができる。
【００２８】
また、部分文書の情報として部分文書の長さを保存し、この長さにより検索可能としているので、必要な長さの部分文書を検索することができ、文書を容易に再利用することができる。
【発明を実施するための最良の形態】
【００２９】
以下、本発明を図面を参照して説明する。
図１は本発明の一実施形態の部分文書検索装置を示す図である。
【００３０】
図１において、本実施形態の部分文書検索装置は、再利用のための検索対象となる文書を取得する文書取得部１と、取得した文書から一部分を部分文書として抽出する部分文書抽出部２と、抽出した部分文書の情報を記憶装置７に保存する部分文書情報保存部３と、入力された検索条件に従って部分文書を検索する部分文書検索部４と、部分文書を検索する条件を入力する検索条件入力部５と、検索された部分文書の情報を出力する検索結果出力部６とを備えている。
【００３１】

このような部分文書検索装置において、利用者からの指示により、文書取得部１は、再利用のための検索対象となる文書を取得し、部分文書抽出部２により部分文書を抽出し、部分文書情報保存部３で部分文書の情報を記憶装置７に保存する。

【００３２】

具体的には、図２のフローチャートに示すように、文書取得部１は、利用者から指示された、コンピュータに蓄積されている全ての文書、あるフォルダ以下の文書、または文書管理システムに登録されている文書などの文書から一つの文書を取得し、部分文書抽出部２に入力する（Ｓ１１）。

【００３３】

部分文書抽出部２は、文書を入力されると、内容の類似性による部分文書の抽出を行う（Ｓ１２）。

【００３４】

これは、文書が複数のトピックで成り立っている場合に、その各トピックに関連して内容が類似している範囲をそれぞれ部分文書として抽出しようとするものである。トピックに関連して内容が類似する範囲を判別する方法として、単語の出現頻度による方法を使う。

【００３５】

予めトピックとして抽出したいジャンルに関連する単語をジャンルごとに分類した辞書を作成し、各文書の各段落（プレゼンテーション用文書の場合は各ページ）から作成した辞書にある単語を抽出し、各段落（あるいは各ページ）がどのジャンルの単語から構成されているかを取得する。

【００３６】

次に、前後の段落でのジャンルの構成を比較し、類似している場合は同じトピックについて記載されていると判定し、一つの部分文書とする。

【００３７】

類似の判定方法としては、各ジャンルの単語の出現頻度はベクトルとみなすことができるため、各段落での各ジャンルの単語の出現頻度をベクトルとし、ジャンルごとにそのベクトル間の距離を計算し、予め設定された閾値以下の場合、類似しているとみなす。

【００３８】

ベクトル間の距離の計算にはユークリッド距離を使用することができる。次元数ｎの二つのベクトルｕ，ｖ間のユークリッド距離は以下の式で計算可能である。

【００３９】

【数１】

$$dist(u,v)=\sqrt{(v_1-u_1)^2+(v_2-u_2)^2+\ldots+(v_n-u_n)^2}$$

【００４０】

また、類似度の判定にＬＳＩ（Latent Semantic Indexing）法を用いることもできる。

【００４１】

ＬＳＩ法は、特異値分解を利用した検索手法である。単語数ｔ、文書数ｄとした場合に、全文書の単語の出現頻度を行列として表したものをH_{td}とした場合、特異値分解により

$H_{td}=U_{tr}D_{rr}V_{rd}$

という三つの行列の積として分解する。ここで、D_{rr}は対角行列であり、次元数ｒは行列の階数になる。

【００４２】

ここで、V_{rd}は、階数×文書数の行列になる。この各列のベクトルが各文書の特徴になる。このベクトル間の距離が小さいほど文書が類似しているとみなすことができる。

【００４３】

すなわち、各ジャンルの単語について各段落（または各ページ）の単語の出現頻度によりベクトルを求め、ベクトル間の距離が予め設定された閾値以下の場合、類似していると

みなす。ベクトル間の距離は、上述のユークリッド距離を用いることができる。

【００４４】

また、予め各段落（または各ページ）がどのジャンルに属するかを判定し、同じジャンルの段落（またはページ）が続いていれば、それらをまとめて部分文書としてもよい。

【００４５】

この場合、各段落（または各ページ）から、ジャンルごとに分類した辞書にある単語の出現頻度を取得し、ジャンルごとの単語の出現頻度が予め設定された閾値以上である場合、その段落（またはページ）をそのジャンルに属するものとする。どのジャンルの単語の出現頻度も閾値を超えない場合は、部分文書としない。

【００４６】

例として、以下のような文書を内容の類似性で部分文書を抽出する場合を説明する。基本的な動作は文字だけの文書だけでなく、文字と画像が混ざった文書でも同様に可能である。

【００４７】

「当研究所では、大量の画像や映像、音声等のマルチメディア情報の中から有用な情報を抽出するための手法としてマルチメディア検索技術を開発しています。マルチメディア検索では仮想３次元空間に情報を配置し、その空間を概観しウォークスルーしながら欲しい情報を検索することができます。

また、家庭やオフィスで人をサポートするロボット技術の開発も行っています。各種センサから得られる情報を元に現実空間を認識し、高度な姿勢制御アルゴリズムにより安定した二足歩行を行うことができます。」

【００４８】

この文書では、マルチメディア検索とロボットという二つの異なるトピックが含まれている。

【００４９】

最初の段落では、「画像」、「映像」、「音声」、「マルチメディア情報」などの「マルチメディア情報」のジャンルに登録された単語が現れ、この段落はマルチメディア情報関連のトピックと判定される。

【００５０】

次の段落では、「ロボット」、「姿勢制御」、「二足歩行」などの「ロボット」のジャンルに登録された単語が現れ、この段落はロボット関連のトピックと判定され、この二つの段落では、それぞれの段落のトピックが異なるため、別々の部分文書とされる。

【００５１】

次に、部分文書抽出部２は、説明文による部分文書の抽出を行う（Ｓ１３）。

【００５２】

段落（プレゼンテーション用文書の場合はそのページ）が、「～とは、」などの特定のキーワード（事柄）を解説する形式になっている場合、その段落を一つの部分文書とする。

【００５３】

また、キーワードをより詳細に説明する場合など、一つの段落で説明が完了しない場合も考えられる。このため、以下の方法で連続した段落に説明が続いているかを判定し、説明が続いていると判定されたら、それらを一つの部分文書とする。

【００５４】

後続の段落（またはページ）と前の段落（またはページ）で出現する単語が類似している場合、説明が続いていると判定する。

【００５５】

類似性の判定としては、各段落の単語の出現頻度をベクトルとし、各ベクトルのユークリッド距離が閾値以下の場合に類似しているとしたり、単語の出現頻度のベクトルの次元数を上述のＬＳＩ法により圧縮してそのユークリッド距離が閾値以下の場合に類似しているとしたりする。

【００５６】
なお、後続の段落（またはページ）に最初のキーワードが含まれている場合、閾値の値を上げるようにしてもよい。
【００５７】
例として、以下のような文書を説明文により部分文書を抽出する場合を説明する。基本的な動作は文字だけの文書だけでなく、文字と画像が混ざった文書でも同様に可能である。
【００５８】
「マルチメディア検索では、画像の配置に自己組織化マップという手法を使用する。
自己組織化マップとは、ニューラルネットワークをベースとした教師なし学習方法であり、元の多次元ベクトル空間の分布をなるべく保ったまま、二次元空間に配置することができる。学習を繰り返すことで、元のベクトル空間で密度が高い部分は、二次元空間の広い範囲に配置され、ベクトル空間で密度が低い部分は、二次元空間の狭い範囲に配置される。
自己組織化マップを使用することで、元の多次元ベクトル空間の大雑把な密度分布を二次元空間上で把握することが可能になる。また、類似している物同士をなるべく近くに配置することで探しやすくすることも可能である。
次に、マルチメディア検索では、画像が表示された仮想空間上をウォークスルーして、欲しい情報を探し出す。単にウォークスルーするだけでなく、キーワード等を利用してキーワードに合致する画像を手前に目立つように表示することも可能である。」
【００５９】
この文書では、二段落目で「自己組織化マップ」というキーワードを説明している。この段落では、「配置」、「ベクトル空間」、「二次元空間」、「分布」、「密度」などの単語が出現している。続く段落では、説明の元となっている「自己組織化マップ」というキーワードが出現しているだけでなく、「配置」、「ベクトル空間」、「二次元空間」、「分布」、「密度」などの単語が同様に出現している。これにより、この二つの段落は連続していて、「自己組織化マップ」を説明していると判定される。
【００６０】
その次の段落では、一転して上述の単語は出現せず、「表示」、「ウォークスルー」、「キーワード」等の単語が出現するため、この段落は連続していないと判定され、「自己組織化マップとは、」から「配置することで探しやすくすることも可能である。」までを一つの部分文書として抽出する。
【００６１】
次に、部分文書抽出部２は、再利用の頻度による部分文書の抽出を行う（Ｓ１４）。
【００６２】
利用者がある程度の期間文書作成を行っている場合、または複数の利用者が同じテーマで文書作成を行っている場合、利用者が同じ範囲を何度も再利用することが考えられる。このように再利用された回数が多い範囲を部分文書とする。
【００６３】
この場合、利用者が行ったキーボード入力やマウス操作を監視し、文書編集中にコピー＆ペーストやファイルの挿入などで各文書のどの部分が新しい文書に挿入されたかを記録する。この結果、予め設定された閾値以上新しい文書に挿入された範囲を部分文書とする。
【００６４】
次に、部分文書抽出部２は、ページ間の関連性による部分文書の抽出を行う（Ｓ１５）。
【００６５】
複数ページの文書で、各ページのタイトルが連番である等の関連性がある場合、関連した範囲を部分文書とする。
【００６６】

例えば、タイトルが「検索手法（１）」、「検索手法（２）」、「検索手法（３）」などとなっている場合、この三つのページは関連しているとみなし、一つの部分文書とする。

【００６７】

関連性を判断する方法は、各ページのタイトルに使用されている文章を単語に分割し、前後のページで同じ単語が使われている割合を算出し、予め設定された閾値以上の場合にタイトルの文章が類似しており、関連していると判定する。

【００６８】

簡単な計算方法としては、二つのタイトル文章の単語数をそれぞれt_1、t_2とし、共通して使われている単語の数をｃとした場合、類似度ｓを

$$s=((c/t_1)+(c/t_2))/2$$

とする。ｓは０から１の間の値を持ち、値が大きいほど類似しているとみなすことができる。

【００６９】

また、タイトルの文章中の括弧内の数字のみを抽出し、その数値が一つずつ増加している範囲を関連しているとみなしてもよい。

【００７０】

また、上述のＬＳＩ法を用い、ページのタイトルの文章に含まれる単語の出現頻度を特徴ベクトル化し、ベクトル間のユークリッド距離を計算し、予め設定された閾値以下である場合に、タイトルの文章が類似しており、関連しているとみなしてもよい。

【００７１】

また、ページ間の見た目により関連性を判定してもよい。一つの図を何ページにも渡って説明するような場合、同じような図が何度も出現する場合がある。このような場合、ページから色や形などの特徴を抽出して、その類似性により関連しているかを判定する。

【００７２】

色の類似性には色ヒストグラムという手法を利用することができる。色ヒストグラムは、画像中の各画素の色をいくつかの色区分のいずれかに分類し、色区分された各色に分類された画素の割合により類似性を判定する。

【００７３】

レイアウトの類似性には色レイアウトという手法を利用することができる。色レイアウトは、画像をいくつかの部分領域に分割し、それぞれの領域の平均色を算出し、各領域での平均色の類似性により類似性を判定する。

【００７４】

また、ページの構造を解析して、各ページのどの位置に図形が配置されていて、どの位置にテキストが配置されているかという特徴を利用することもできる。この形の類似性の判定にはウェーブレット（Wavelet）変換が利用できる。ウェーブレット変換では、画像を解析し、画面上の位置とその位置における縦横斜め方向の周波数（細かい変化があるか、大きな変化があるかといった情報）を出力する。

【００７５】

これらの見た目の特徴を抽出する方法の出力はベクトルとして表現できるため、類似度の算出には上述のユークリッド距離を使うことができる。

【００７６】

これらの部分文書抽出のための処理は、上述の方法全てを実行してもかまわないし、複数の方法を組み合わせて実行してもかまわないし、どれか一つの方法だけにより部分文書を抽出してもよい。また、一つの文書から複数の部分文書を抽出しても、各部分文書の範囲に重なっている部分があってもかまわない。

【００７７】

複数の部分文書抽出方法で同一の範囲の部分文書が抽出された場合は、それらをまとめて一つの部分文書としてもかまわない。

【００７８】

部分文書抽出部２は、このようにして抽出した部分文書の情報を部分文書情報保存部３に出力して記憶装置７に保存させる（Ｓ１６）。

【００７９】

なお、それぞれの部分文書は、元文書とは独立した新しい文書を作成してもかまわないし、元文書中の位置や長さなどの情報だけを保存するようにしてもよい。

【００８０】

保存する情報としては、各文書情報（ファイル名（ＵＲＬ：Uniform Resource Locater）、作成日付、タイトル、作成者等）と各部分文書の情報（文書のＩＤ、文書中の位置、長さ、部分文書抽出方法、テキスト、画像等）があげられる。

【００８１】

検索処理を高速化するために、予め検索用のインデックスを作成しておいてもかまわない。方法としては、一般的なテキスト検索方法（Ｎグラム法や形態素解析による全文検索方法等）や画像類似検索方法等を利用できる。

【００８２】

また、文書を再利用している場合、部分文書単位では再利用した部分が多数現れる可能性がある。このため、複数の文書間で類似した内容の部分文書が抽出された場合、検索結果としてはどれか一つの部分文書のみを提示することが考えられる。

【００８３】

この場合、上述の部分文書の情報としては代表の一つの部分文書の情報のみを登録し、その部分文書に類似した部分文書の情報は別のテーブルに保存することで検索処理を高速化することができる。なお、部分文書間の類似性を判定するには、上述の単語の出現頻度による類似性の判定やページの見た目による類似性の判定を利用できる。

【００８４】

次に、部分文書抽出部２は、利用者から指示された全ての文書を処理したかを判定し（Ｓ１７）、全て処理していなければ、Ｓ１１に戻り次の文書を取得する。全て処理したら終了する。

【００８５】

このようにして、利用者の指定した文書から部分文書が抽出され、その情報が記憶装置７に保存される。

【００８６】

図３は、一つの文書から各部分文書抽出方法により抽出される部分文書の例を示す図である。

【００８７】

図３の例では、再利用頻度で抽出された部分文書の一部が、内容の類似性によりまたは説明文により部分文書として抽出されており、ページ間の関連性で抽出された部分文書と同一のものが内容の類似性により抽出された部分文書となっている。

【００８８】

図４は、図３の例で抽出された部分文書による部分文書情報のデータ構成例を示す図である。この例では、ページ間の関連性と内容の類似性で抽出された同一の部分文書の情報が一つにまとめられている（５行目）。

【００８９】

図５は、このように保存された部分文書の情報から、部分文書を検索する処理を説明するためのフローチャートである。

【００９０】

検索条件入力部５は、利用者から検索の要求を受け付けると、図６に示すような検索条件入力画面を表示し、キーワードや類似画像や部分文書の長さや部分文書抽出時の方法などの検索条件の入力を要求する（Ｓ２１）。

【００９１】

部分文書の長さは、部分文書抽出処理により抽出された部分文書の長さである。長さが短いもの（簡潔）、長いもの（詳細）等、必要な情報の長さを入力することで、適切な長

さの部分文書を検索することができる。

【００９２】

プレゼンテーション資料のようにページ数がはっきりしているタイプの文書の場合、長さの指定はページ数が好ましい。

【００９３】

部分文書としてページの一部も含む場合は、その領域のサイズ（あるいは、領域サイズの１ページに占める割合）などを指定できるようにしてもよい。

【００９４】

また、文字中心の文書を含む場合は、長さとして行数を指定できるようにしてもよい。

【００９５】

検索時に部分文書の長さを指定することにより、詳しく説明している部分、簡単に説明している部分、３ページ程度で説明している部分のような、必要な部分の長さを指定して検索を行うことができる。

【００９６】

検索条件が入力されると、検索条件入力部５は、入力された検索条件を部分文書検索部４に入力する。

【００９７】

部分文書検索部４は、入力された検索条件に従って記憶装置７に保存されている部分文書情報から部分文書を検索し、検索条件にマッチする部分文書情報を取得し（Ｓ２２）、検索結果として取得した部分文書情報を、検索結果出力部６に入力する。

【００９８】

検索結果出力部６は、入力された部分文書情報に基づいて、例えば図７に示すような検索結果表示画面を出力する（Ｓ２３）。

【００９９】

図７の例では、部分文書の情報として、元文書のファイル名、部分文書の長さ、元文書の長さ、要約テキスト、サムネイル画像等を表示している。

【０１００】

元文書の長さと部分文書の長さの比から、部分文書と元文書の関連性を判定し、再利用のし易さ等の判定を行うことができる。

【０１０１】

例えば、部分文書の長さより元文書の長さが大変長い場合は、元の文書は長いが指定したキーワードに関連している部分は少ないため、キーワードと文書全体との関連性は低いことが分かる。

【０１０２】

逆に、部分文書の長さと元文書の長さがあまり違わない場合には、文書全体が関連していることが分かる。

【０１０３】

また、キーワードに関連した短い部分文書が一つの文書中に多数ある場合は、文書全体はキーワードと関連しているが、引用しにくいことなどが分かる。

【０１０４】

このような画面表示中に、検索結果出力部６は、利用者が画面上の「ページを引用」と表示されたボタンをクリックして部分文書のコピーを選択したかを判定し（Ｓ２４）、選択されていれば、対応する部分文書の内容をクリップボードにコピーする（Ｓ２５）。

【０１０５】

部分文書のコピーが選択されていなければ、検索結果出力部６は、利用者が画面上の部分文書のサムネイル画像をクリックして部分文書の表示を選択したかを判定し（Ｓ２６）、選択されていれば、対応するアプリケーションを起動して部分文書を表示させる（Ｓ２７）。

【０１０６】

部分文書の表示が選択されていなければ、検索結果出力部６は、利用者が画面上の「検

索画面」と表示されたボタンをクリックして再検索を選択したかを判定し（Ｓ２８）、選択されていれば、Ｓ２１に戻って検索条件入力部５により検索画面を表示させる。選択されていなければ、Ｓ２４に戻って部分文書のコピーが選択されたか判定する。

【０１０７】

このように、文書の一部分を部分文書として抽出して保存し、保存されている部分文書から利用者の条件により検索し、検索結果の表示画面からワンクリックで部分文書の内容をクリップボードにコピーしているので、検索された文書の中から目的の箇所を探すなどの文書再利用時の手間を削減させることができる。

【０１０８】

なお、図７の検索結果画面で、「類似文書を表示」と表示されたボタンをクリックすると、上述の複数の文書間で類似した内容の部分文書が抽出され、部分文書の情報としては代表の一つの部分文書の情報のみを登録し、その部分文書に類似した部分文書の情報は別のテーブルに保存している場合の、別のテーブルに保存している部分文書の情報が表示され、下線が付いている元文書のファイル名をクリックすると、対応するアプリケーションが起動され元文書が表示される。

【０１０９】

図８は、検索結果表示画面の他の例を示す図である。図８の例では、検索結果の各部分文書のサムネイル画像が横方向に一列に並べられる。縦位置が異なるものが別の部分文書であり縦方向に検索結果順に並べられる。図では画像しか表示していないが、各画像の上や左に部分文書の情報を表示してもよい。

【０１１０】

この表示では、右上のウィンドウ中の矢印のボタンをクリックすることで視点を上下左右に移動することができ、「Ｚ＋」のボタンをクリックすることでズームアップし、「Ｚ－」のボタンをクリックすることでズームバックすることができ、一部の部分文書付近を拡大して表示させたり、全体を俯瞰して表示させたりすることができる。

【０１１１】

次に、本実施形態の部分文書検索装置の部分文書から雛形文書を作成する処理について説明する。

【０１１２】

例えば、会社の技術紹介など複数のトピックからなる文書を作成する場合、各トピックを説明する部分文書をつなぎ合わせて一つの文書を作成したい場合がある。このような場合、既存の紹介資料を再利用しようとしても同じトピックを紹介している文書が見つかるとは限らないため、部分文書を検索してつなぎ合わせるという作業が必要となる。

【０１１３】

また、同じトピックを紹介している既存文書が見つかったとしても、紹介する相手や説明時間によって適切な説明の長さや説明文の内容が異なる場合がある（概要を短時間で説明したい場合、より詳細な技術内容まで踏み込んで説明する場合など）。

【０１１４】

そこで、必要なトピック項目を列挙することで、トピックに関連した部分文書を検索し、検索された部分文書から雛形文書を作成することができるようになっている。

【０１１５】

図９は、雛形文書作成の流れを示す図である。まず利用者は、作成したい文書に必要なトピック項目を列挙した目次文書を作成する。そして、本実施形態の部分文書検索装置に雛形文書作成の指示を行う。

【０１１６】

検索条件入力部５は、雛形文書作成の指示を受けると、図１０のフローチャートに示すように、図９の２．に示すような画面を表示し、目次文書と各トピックの長さの入力を要求する（Ｓ３１）。

【０１１７】

目次文書と各トピックの長さを入力されると、検索条件入力部５は、目次文書で指定さ

れた項目から検索に使用するキーワードを抽出する（Ｓ３２）。キーワード抽出には、形態素解析法などが利用できる。

【０１１８】

図９の例では、例えば、トピック１から「クロスメディア」、「検索」が、トピック２から「類似」、「画像」、「検索」が、トピック３から「オフィス文書」、「検索」が、トピック４から「映像」、「検索」がキーワードとして抽出される。

【０１１９】

それぞれのトピック項目ごとにキーワードを抽出したら、検索条件入力部５は、抽出したトピック項目ごとのキーワードと各トピックの長さを部分文書検索部４に入力する。

【０１２０】

部分文書検索部４は、トピックごとに入力されたキーワードと長さを使って部分文書を検索する（Ｓ３３）。

【０１２１】

次に、部分文書検索部４は、検索結果として得られたトピックごとの部分文書から、適切な組み合わせの部分文書を選択する（Ｓ３４）。

【０１２２】

トピックごとに検索された部分文書は、長さが同じでも説明のレベルが異なる場合があるため、以下の方法により適切な組み合わせを選択する。

（１）同一の文書で使用されている

部分文書の元文書が同一の文書である場合、これらの部分文書の組み合わせは適切な組み合わせであるとする。上述した部分文書の抽出時に類似した部分文書が抽出されている場合、類似した部分文書の中に元文書が同一の文書であるものがあれば、適切な組み合わせであるとする。

（２）類似した用法、単語が使用されている

トピックが異なる場合でも、同様の言葉で説明されている場合は、適切な組み合わせであるとする。上述の形態素解析を用いて単語を抽出し、共通の単語が使われている場合、適切な組み合わせであるとする。また、同様に形態素解析を用いて文の語尾を抽出し、同様の語尾（ですます調、である調の別など）のものを適切な組み合わせであるとしてもよい。

（３）類似した画面構成、色使いが使用されている

画面構成が類似している、あるいは画面の色使いが似ているものを適切な組み合わせとして選択する。画面構成の類似性や色使いの類似性の判断については、上述した部分文書抽出時のページの見た目の類似性の判断に用いた方法を使うことができる。

【０１２３】

部分文書検索部４は、このようにして適切な組み合わせとして選択した各トピックの部分文書を検索結果出力部６に入力する。

【０１２４】

検索結果出力部６は、入力されたトピックごとの部分文書を、例えば図９の４．の上の図のようにサムネイル画像を横方向に並べて表示する（Ｓ３５）。

【０１２５】

なお、適切な部分文書を複数選択してもよいようにし、例えば図９の４．の下の図のように、縦方向に複数の候補を並べて表示し、利用者に選択させるようにしてもよい。

【０１２６】

そして、検索結果出力部６は、利用者が、複数候補がある場合は複数の候補の中から一つの部分文書を選択し、雛形文書作成の指示を入力すると、利用者が作成した目次文書と選択された部分文書をつなぎ合わせて雛形文書を作成し（Ｓ３６）、対応するアプリケーションを起動して、雛形文書を表示させる。

【０１２７】

このようにして、利用者が作成した目次の項目に関連する部分文書がそれぞれ検索され、検索された部分文書の中から適切な組み合わせの部分文書が選択され、選択された部分

文書により雛形文書が作成され、目次の項目を作成するだけで利用者の目的に合った雛形文書を作成することができる。

【０１２８】

なお、本実施形態においては、部分文書の適切な組み合わせを選択して表示するようにしたが、単純にキーワードの類似順、あるいは再利用頻度順など何らかの順番に規定数のみ提示し、利用者に選択させるようにしてもよい。

【０１２９】

このように本実施形態においては、文書の一部分を部分文書として抽出して保存し、保存されている部分文書から利用者の条件により検索し、検索結果の表示画面からワンクリックで部分文書の内容をクリップボードにコピーしているので、検索された文書の中から目的の箇所を探すなどの文書再利用時の手間を削減させることができる。

【図面の簡単な説明】

【０１３０】

【図１】本発明の一実施形態の部分文書検索装置を示す図であり、そのブロック図である。

【図２】本実施形態の部分文書抽出処理を説明するためのフローチャートである。

【図３】本実施形態の部分文書抽出の例を示す図である。

【図４】本実施形態の部分文書情報のデータ構成例を示す図である。

【図５】本実施形態の部分文書検索処理を説明するためのフローチャートである。

【図６】本実施形態の検索条件入力画面の例を示す図である。

【図７】本実施形態の検索結果表示画面の例を示す図である。

【図８】本実施形態の検索結果表示画面の他の例を示す図である。

【図９】本実施形態の雛形文書作成の流れを示す図である。

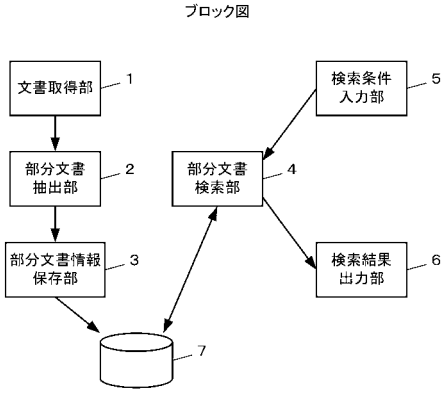
【図１０】本実施形態の雛形文書作成処理を説明するためのフローチャートである。

【符号の説明】

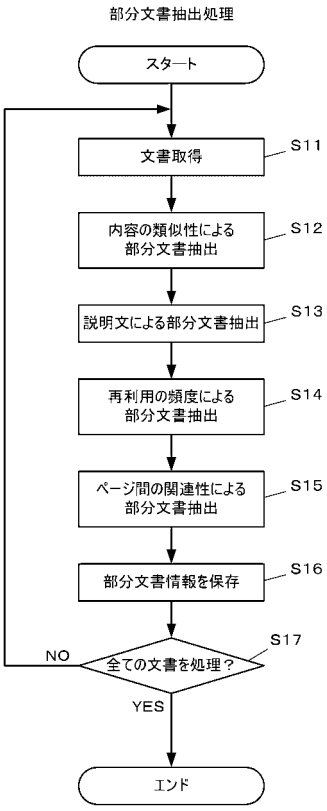
【０１３１】

１　文書取得部
２　部分文書抽出部
３　部分文書情報保存部
４　部分文書検索部
５　検索条件入力部
６　検索結果出力部
７　記憶装置

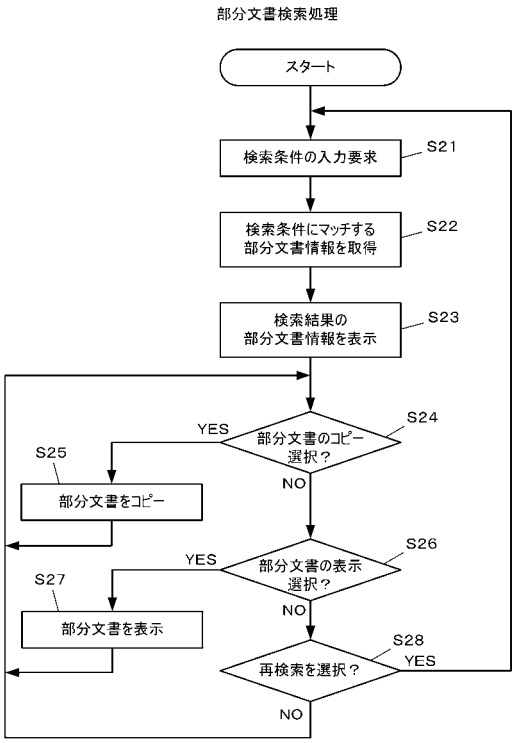
【図１】


ブロック図
文書取得部
1
部分文書
抽出部
2
部分文書情報
保存部
3
部分文書
検索部
4
検索条件
入力部
5
検索結果
出力部
6
7

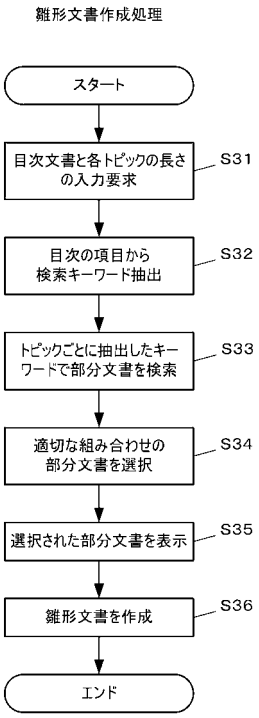
【図２】


部分文書抽出処理
スタート
文書取得
S11
内容の類似性による
部分文書抽出
S12
説明文による部分文書抽出
S13
再利用の頻度による
部分文書抽出
S14
ページ間の関連性による
部分文書抽出
S15
部分文書情報を保存
S16
全ての文書を処理？
S17
NO
YES
エンド

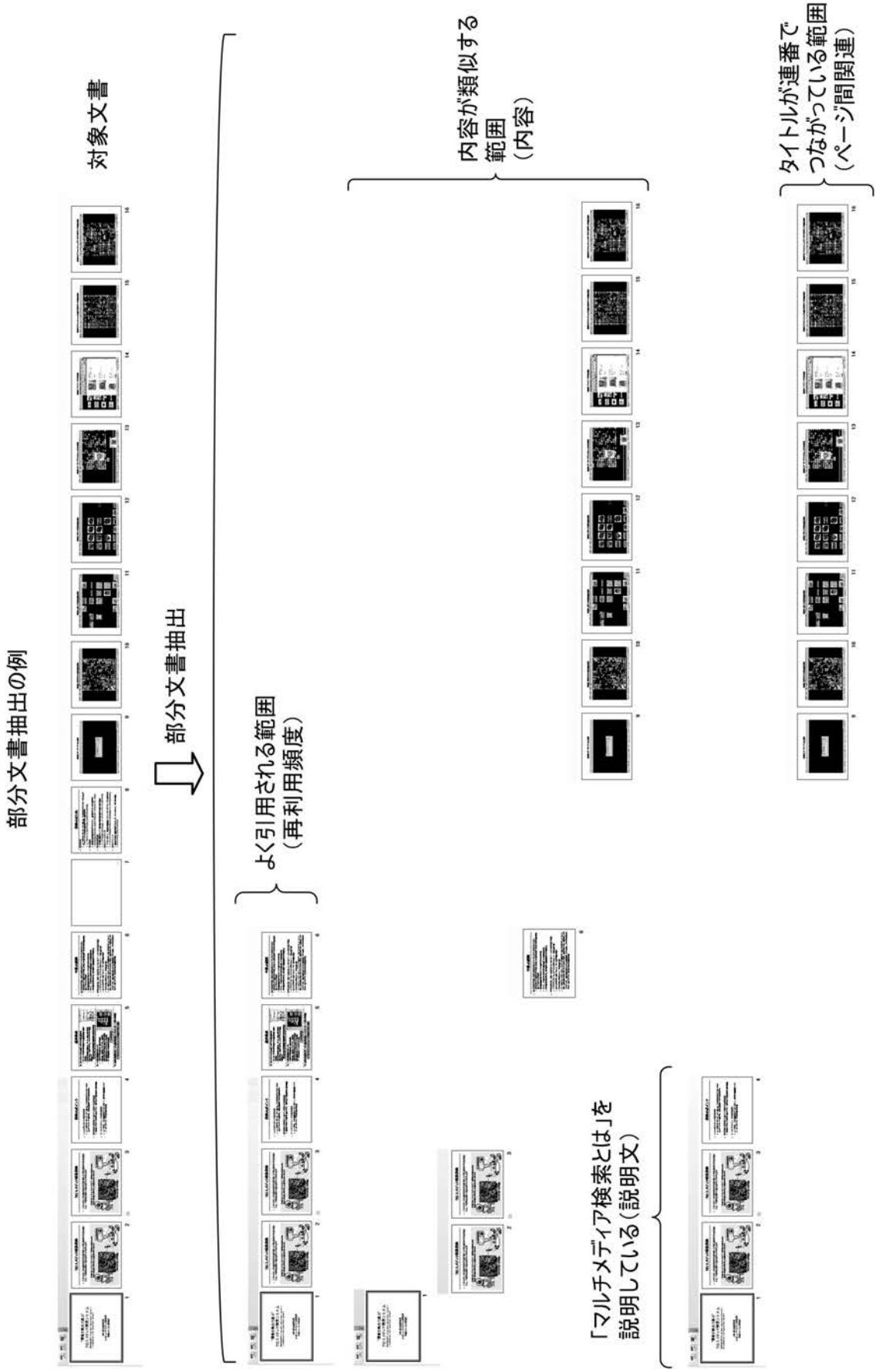
【図５】


部分文書検索処理
スタート
検索条件の入力要求
S21
検索条件にマッチする
部分文書情報を取得
S22
検索結果の
部分文書情報を表示
S23
部分文書のコピー
選択？
S24
YES
NO
S25
部分文書をコピー
部分文書の表示
選択？
S26
YES
NO
S27
部分文書を表示
再検索を選択？
S28
YES
NO

【図１０】

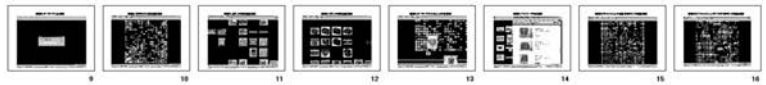

雛形文書作成処理
スタート
目次文書と各トピックの長さ
の入力要求
S31
目次の項目から
検索キーワード抽出
S32
トピックごとに抽出したキー
ワードで部分文書を検索
S33
適切な組み合わせの
部分文書を選択
S34
選択された部分文書を表示
S35
雛形文書を作成
S36
エンド

【図３】


部分文書抽出の例
対象文書
部分文書抽出
よく引用される範囲
（再利用頻度）
内容が類似する
範囲
（内容）
タイトルが連番で
つながっている範囲
（ページ間関連）
「マルチメディア検索とは」を
説明している（説明文）

【図４】

部分文書情報のデータ構成

文書番号	位置	長さ	抽出方法	テキスト	画像
1	2	5	再利用頻度	クロスメディア検索技術…	
1	1	1	内容	情報を眺めて選ぶ…	
1	2	2	内容	クロスメディア検索技術…	
1	5	1	内容	適用事例…	
1	8	8	内容、ページ間関連	デモ手順（1）…	
1	2	3	説明文	クロスメディア検索技術とは…	
2	1	2	再利用頻度	…	
2	3	5	説明文	…	

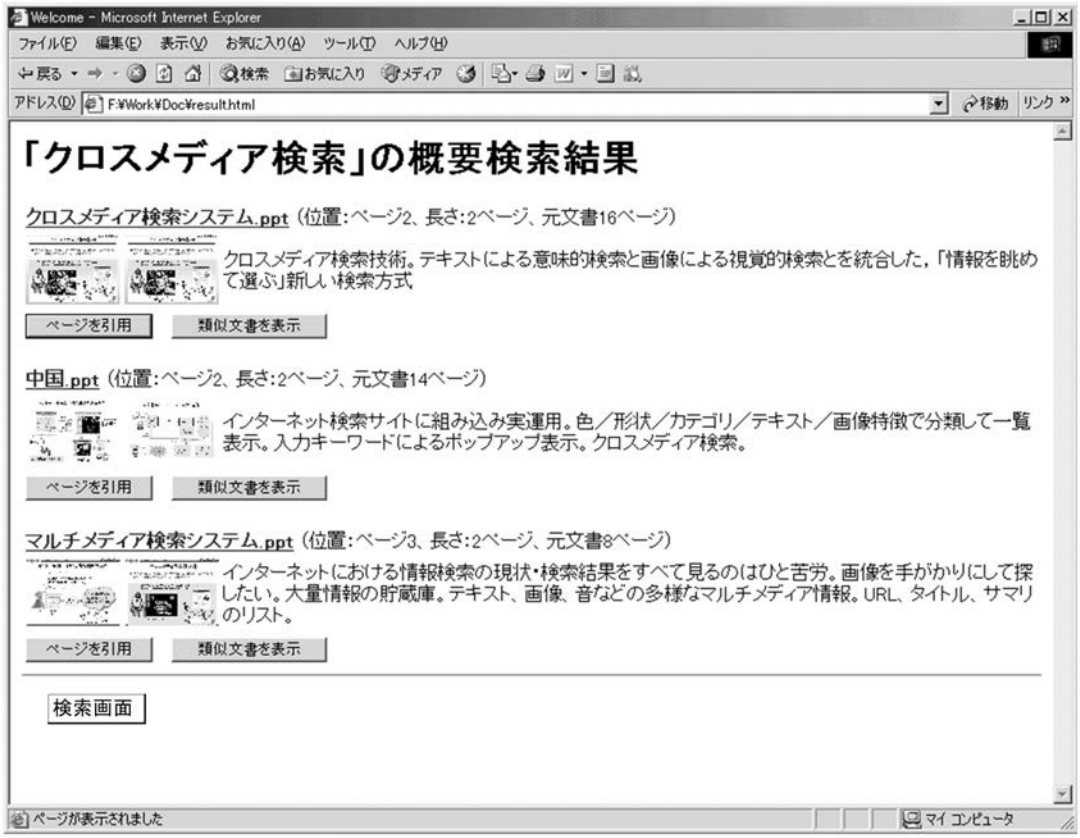
JP 4904920 B2 2012.3.28

【図６】

検索条件入力画面の例

【図７】

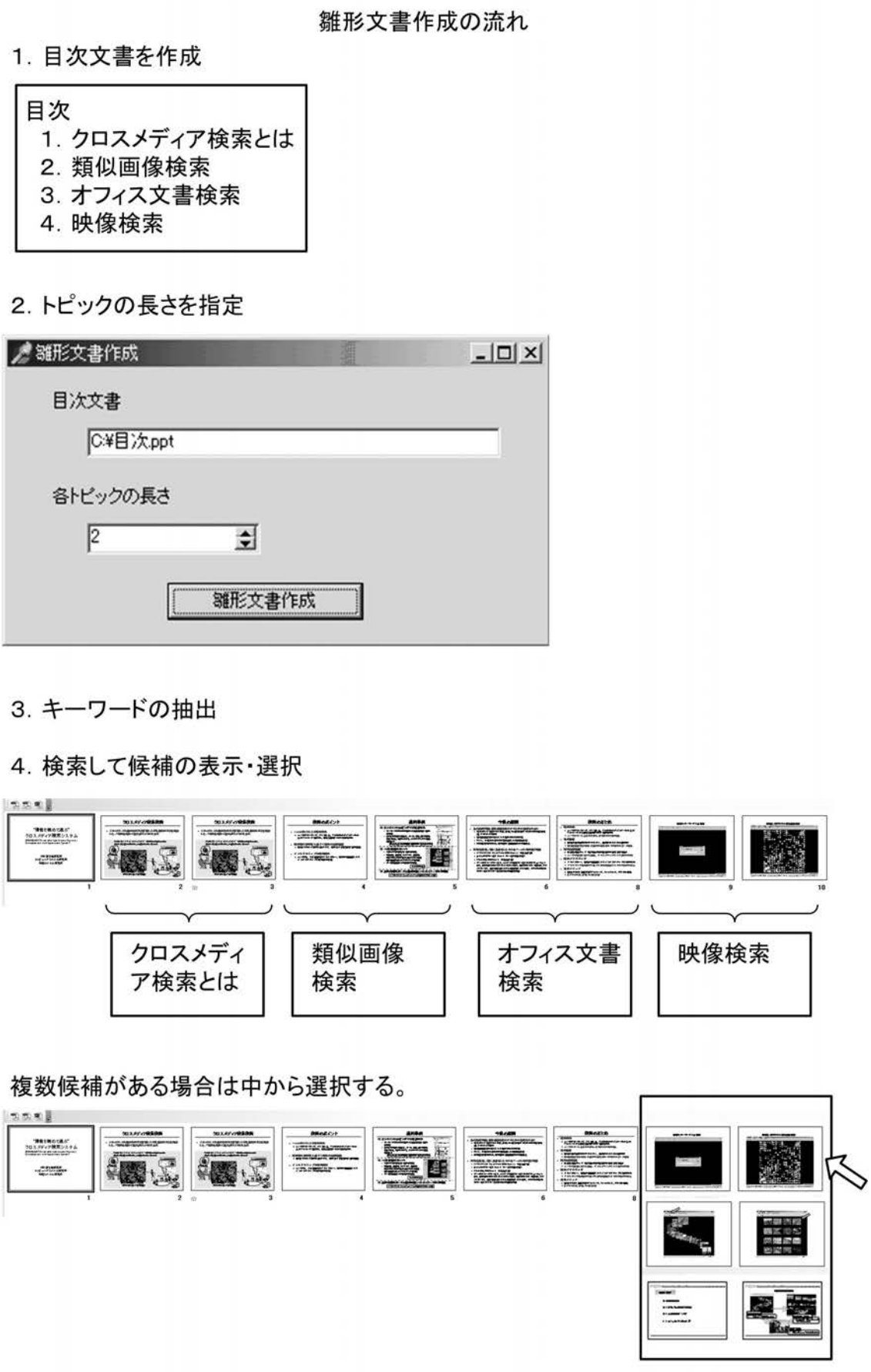
検索結果表示画面の例



【図８】

検索結果表示画面の他の例

【図９】


雛形文書作成の流れ
1. 目次文書を作成
目次
1. クロスメディア検索とは
2. 類似画像検索
3. オフィス文書検索
4. 映像検索
2. トピックの長さを指定
雛形文書作成
目次文書
C:¥目次.ppt
各トピックの長さ
2
雛形文書作成
3. キーワードの抽出
4. 検索して候補の表示・選択
クロスメディア検索とは
類似画像検索
オフィス文書検索
映像検索
複数候補がある場合は中から選択する。

フロントページの続き

(72)発明者　上原　祐介
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(72)発明者　長田　茂美
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

審査官　長　由紀子

(56)参考文献　特開２００４－２５９０３１（ＪＰ，Ａ）
特開平０９－２３１２２８（ＪＰ，Ａ）
特開平７－２７１５６９（ＪＰ，Ａ）
特開平５－１０１０５４（ＪＰ，Ａ）
特開平９－１６０８９６（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／２１　－　２６
Ｇ０６Ｆ　１７／３０