(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7670951号
(P7670951)

(45)発行日　令和7年5月1日(2025.5.1)

(24)登録日　令和7年4月22日(2025.4.22)

(51)国際特許分類
Ｇ０６Ｆ　16/33　(2025.01)

ＦＩ
Ｇ０６Ｆ　16/33

請求項の数　11　（全21頁）

(21)出願番号　特願2020-178643(P2020-178643)
(22)出願日　令和2年10月26日(2020.10.26)
(65)公開番号　特開2022-69790(P2022-69790A)
(43)公開日　令和4年5月12日(2022.5.12)
審査請求日　令和5年10月25日(2023.10.25)

(73)特許権者　390002761
キヤノンマーケティングジャパン株式会社
東京都港区港南２丁目１６番６号
(73)特許権者　592135203
キヤノンＩＴソリューションズ株式会社
東京都港区港南２丁目１６番６号
(74)代理人　100189751
弁理士　木村 友輔
(72)発明者　松田 雄介
東京都港区港南２丁目１６番６号 キヤノンＩＴソリューションズ株式会社内
(72)発明者　福田 直之
東京都港区港南２丁目１６番６号 キヤノンＩＴソリューションズ株式会社内

最終頁に続く

(54)【発明の名称】　情報処理装置、情報処理方法、プログラム

(57)【特許請求の範囲】

【請求項１】

文書のカテゴリ毎に、文書のフィールドと、当該フィールドから抽出する文字列と、当該フィールドの重みと、を定義した定義情報を受け付ける定義情報受付手段と、

複数の検索対象の文書のそれぞれについて、当該検索対象の文書に係るカテゴリに対応する前記定義情報に基づき、当該検索対象の文書から当該検索対象の文書のフィールド毎に前記文字列を抽出する抽出手段と、

ユーザから前記検索対象の文書を検索するための検索語を受け付ける検索語受付手段と、

複数の検索対象の文書のそれぞれについて、前記検索語受付手段により受け付けた検索語と、前記抽出手段により当該検索対象の文書から抽出された文字列と、前記定義情報により定義された当該文字列に係るフィールドの重みと、に基づき、当該検索対象の文書のフィールド毎に算出された検索スコアに基づく当該検索対象の文書の検索スコアを算出するスコア算出手段と、

を備えることを特徴とする情報処理システム。

【請求項２】

前記スコア算出手段は、前記フィールド毎に算出された検索スコアの合計値を当該検索対象の文書の検索スコアとして算出することを特徴とする請求項１に記載の情報処理システム。

【請求項３】

前記スコア算出手段は、前記フィールド毎の検索スコアを、当該フィールドにおける前

記検索語と一致する前記抽出された文字列の数と、当該フィールドの重みとに基づき算出することを特徴とする請求項１または２に記載の情報処理システム。
【請求項４】
前記スコア算出手段は、前記フィールド毎の検索スコアを、当該フィールドにおける前記検索語と一致する前記抽出された文字列の数に、当該フィールドの重みを掛け合わせることで算出することを特徴とする請求項１乃至３のいずれか１項に記載の情報処理システム。
【請求項５】
前記定義情報は、前記フィールドから抽出する文字列の抽出方法として、前記フィールド毎に、形態素解析、キーワードマッチ、正規表現パターンのいずれかの方法が定義されていることを特徴とし、
前記抽出手段は、前記定義情報に定義された抽出方法に基づき、当該検索対象の文書のフィールド毎に前記文字列を抽出することを特徴とする請求項１乃至４のいずれか１項に記載の情報処理システム。
【請求項６】
前記スコア算出手段により算出された検索スコアに基づき、前記検索語による検索結果を表示するよう制御する表示制御手段をさらに備えることを特徴とする請求項１乃至５のいずれか１項に記載の情報処理システム。
【請求項７】
前記検索語により検索された文書の閲覧実績に基づき、当該検索された文書に係るカテゴリに対応する定義情報での前記フィールドの重みを更新する更新手段をさらに備えることを特徴とする請求項１乃至６のいずれか１項に記載の情報処理システム。
【請求項８】
前記更新手段は、閲覧された文書に係るカテゴリに対応する定義情報での前記フィールドの重みを高くすることを特徴とする請求項７に記載の情報処理システム。
【請求項９】
前記更新手段は、閲覧されなかった文書に係るカテゴリに対応する定義情報での前記フィールドの重みを低くすることを特徴とする請求項７または８に記載の情報処理システム。
【請求項１０】
情報処理システムの定義情報受付手段が、文書のカテゴリ毎に、フィールドと、当該フィールドから抽出する文字列と、当該フィールドの重みと、を定義した定義情報を受け付ける定義情報受付工程と、
前記情報処理システムの抽出手段が、複数の検索対象の文書のそれぞれについて、当該検索対象の文書に係るカテゴリに対応する前記定義情報に基づき、当該検索対象の文書から当該検索対象の文書のフィールド毎に前記文字列を抽出する抽出工程と、
前記情報処理システムの検索語受付手段が、ユーザから前記検索対象の文書を検索するための検索語を受け付ける検索語受付工程と、
前記情報処理システムのスコア算出手段が、複数の検索対象の文書のそれぞれについて、前記検索語受付工程により受け付けた検索語と、前記抽出工程により抽出された文字列と、前記定義情報により定義された当該文字列に係るフィールドの重みと、に基づき、当該検索対象の文書のフィールド毎に算出された検索スコアに基づく当該検索対象の文書の検索スコアを算出するスコア算出工程と、
を備えることを特徴とする情報処理方法。
【請求項１１】
コンピュータを請求項１乃至９のいずれか１項に記載の各手段として機能させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、情報処理装置、情報処理方法、プログラムに関する。

【背景技術】
【０００２】
従来の単語頻度のみによる全文検索では単語の重要度や意味というものが考慮されない。そのため、出現頻度は低いが重要な単語ではヒットしても検索上位に現れなかったり、字面は同じだがニュアンスが異なる単語にヒットした文書が検索結果に現れたりするという問題があった。
【０００３】
特許文献１には、文書データのフィールド情報を検索スコアの計算に用いて、ユーザの検索意図に近い検索結果を得るための技術について開示されている。
【先行技術文献】
【特許文献】
【０００４】
【文献】特開２００５－０６３４６８号公報
【発明の概要】
【発明が解決しようとする課題】
【０００５】
特許文献１には、文書データのフィールド情報を検索スコアの計算に用いて、ユーザの検索意図に近い検索結果を得るための技術が記載されている。
【０００６】
しかし、フィールドごとのスコア反映割合を検索のたびにユーザが入力する必要があり、フィールド数が多くなった場合に煩雑である。また、フィールドが事前に文書のメタデータとして用意されていない部分についてはフィールドに格納されない問題がある。さらに文書のカテゴリについての概念がないため、各カテゴリに応じたフィールド情報の抽出やスコア計算を行うことができないという課題がある。
【０００７】
そのため、文書データにカテゴリ情報を付与し、カテゴリごとにフィールド抽出情報を定義することが望まれる。
【０００８】
そこで、本発明は、文書データのカテゴリとフィールド情報を検索スコアの計算に用いて、検索ユーザの意図に近い検索結果を得られるよう検索精度の向上を行うことを目的とする。
【課題を解決するための手段】
【０００９】
本発明の情報処理システムは、文書のカテゴリ毎に文書に含まれる各フィールドに設定される重みを記憶する記憶手段と、前記フィールドに設定された重みと、当該フィールドと検索語との関係とに基づき、当該文書のスコアを算出する算出手段と、を備えることを特徴とする。
【００１０】
本発明の情報処理方法は、文書のカテゴリ毎に文書に含まれる各フィールドに設定される重みを記憶する記憶ステップと、前記フィールドに設定された重みと、当該フィールドと検索語との関係とに基づき、当該文書のスコアを算出する算出ステップと、を備えることを特徴とする。
【００１１】
本発明のプログラムは、コンピュータを、文書のカテゴリ毎に文書に含まれる各フィールドに設定される重みを記憶する記憶手段と、前記フィールドに設定された重みと、当該フィールドと検索語との関係とに基づき、当該文書のスコアを算出する算出手段として機能させることを特徴とする。
【発明の効果】
【００１２】
本発明によれば、文書データのカテゴリとフィールド情報を検索スコアの計算に用いて

、検索ユーザの意図に近い検索結果を得られるよう検索精度の向上を行うことが可能となる。

【図面の簡単な説明】

【００１３】

【図１】情報処理システムのシステム構成を示す図である。

【図２】情報処理装置のハードウェア構成を示す図である。

【図３】本実施例での処理の流れを示すフローチャートである。

【図４】本発明の実施形態における、文書登録処理の一例を示すフローチャートである。

【図５】本発明の実施形態における、フィールド抽出処理の一例を示すフローチャートである。

【図６】本発明の実施形態における、キーワードによるフィールド抽出処理の一例を示すフローチャートである。

【図７】本発明の実施形態における、パターンによるフィールド抽出処理の一例を示すフローチャートである。

【図８】本発明の実施形態における、形態素解析によるフィールド抽出処理の一例を示すフローチャートである。

【図９】本発明の実施形態における、検索処理の一例を示すフローチャートである。

【図１０】本発明の実施形態における、検索セッション統計情報の更新処理の一例を示すフローチャートである。

【図１１】本発明の実施形態における、フィールド重みの更新処理の一例を示すフローチャートである。

【図１２】本発明の実施形態における、抽出定義一覧画面の一例を示す図である。

【図１３】本発明の実施形態における、抽出定義詳細画面の一例を示す図である。

【図１４】本発明の実施形態における、フィールド名とキーワードの距離についての説明の図である。

【図１５】本発明の実施形態における、フィールド重み更新処理の一例を示す図である。

【図１６】本発明の実施形態における、検索セッション統計情報のテーブルの一例を示す図である。

【図１７】本発明の実施形態における、フィールドスコアの計算の一例を示す図である。

【発明を実施するための形態】

【００１４】

以下、図面を参照して、本発明の実施形態を詳細に説明する。なお、以下に説明する実施形態は、本発明を具体的に実施した場合の一例を示すもので、特許請求の範囲に記載した構成の一例である。

【００１５】

図１は、本発明の実施形態における文書検索システムのシステム構成の一例を示す図である。本発明における実施形態における文書検索システム２０００は、情報処理装置１００と、文書ＤＢ１０７、クライアントＰＣ１０８で構成される。情報処理装置１００は、文書登録処理部１０１、文書検索処理部１０２、形態素解析辞書１０３、登録文書インデックス１０４、抽出定義ＤＢ１０５、検索セッション統計情報１０６から構成され、外部の文書ＤＢ１０７や文書検索システムに文書を登録する際に使用するクライアントＰＣとネットワークを介して通信可能に接続されている。

【００１６】

文書登録処理部１０１では、ユーザから受け付けた文書に係る処理を実行する機能部である。具体的には、テキスト抽出処理やカテゴリ付与やフィールドの抽出処理を行い、検索インデックスを作成し、登録文書インデックス１０４に格納するなどの処理を行う。

【００１７】

文書検索処理部１０２では、ユーザから受けつけた検索語を用いて、インデックス済みの文書を検索する機能部である。ユーザから検索語を受け付けると、インデックス済みの文書から本文スコアとフィールドスコアを計算して、それぞれを合算して検索結果に反映

させる処理を行う。
【００１８】
形態素解析辞書１０３は、形態素解析を行う際に使用される辞書である。
【００１９】
登録文書インデックス１０４は、登録対象となる文書から抽出した本文及び各フィールドに対する検索インデックスを格納するＤＢである。本ＤＢを用いて、検索処理部１０２による処理が行われる。
【００２０】
抽出定義ＤＢ１０５は、カテゴリ毎に定義づけられる抽出定義を記憶しておくＤＢである。本抽出定義ＤＢに記憶される当該カテゴリの抽出定義として設定された抽出方式により、フィールドの抽出を行う。抽出方式は、キーワードによる抽出を行うか、パターンによる抽出を行うか、形態素解析による抽出などがある。
【００２１】
検索セッション統計情報１０６は、ユーザの検索セッション統計情報を更新するＤＢである。ユーザの検索セッション統計情報の更新を行い、抽出定義のフィールド重みを更新する際に利用する。
【００２２】
文書ＤＢ１０７は、文書が記憶されているＤＢである。クラウドサービスなどの外部ＤＢも含まれる。
【００２３】
クライアントＰＣ１０８は、ユーザから文書登録を受付ける際に使用される。
【００２４】
図２は、本発明の実施形態における情報処理装置のハードウェア構成の一例を示すブロック図である。
【００２５】
図２に示すように、情報処理装置は、システムバス２００を介してＣＰＵ（Ｃｅｎｔｒａｌ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）２０１、ＲＯＭ（Ｒｅａｄ Ｏｎｌｙ Ｍｅｍｏｒｙ）２０２、ＲＡＭ（Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ）２０３、記憶装置２０４、入力コントローラ２０５、音声入力コントローラ２０６、ビデオコントローラ２０７、メモリコントローラ２０８、および通信Ｉ／Ｆコントローラ２０９が接続される。
【００２６】
ＣＰＵ２０１は、システムバス２００に接続される各デバイスやコントローラを統括的に制御する。
【００２７】
ＲＯＭ２０２あるいは外部メモリ２１３は、ＣＰＵ２０１が実行する制御プログラムであるＢＩＯＳ（Ｂａｓｉｃ Ｉｎｐｕｔ／Ｏｕｔｐｕｔ Ｓｙｓｔｅｍ）やＯＳ（Ｏｐｅｒａｔｉｎｇ Ｓｙｓｔｅｍ）や、本情報処理方法を実現するためのコンピュータ読み取り実行可能なプログラムおよび必要な各種データ（データテーブルを含む）を保持している。
【００２８】
ＲＡＭ２０３は、ＣＰＵ２０１の主メモリ、ワークエリア等として機能する。ＣＰＵ２０１は、処理の実行に際して必要なプログラム等をＲＯＭ２０２あるいは外部メモリ２１３からＲＡＭ２０３にロードし、ロードしたプログラムを実行することで各種動作を実現する。
【００２９】
入力コントローラ２０５は、キーボード２１０や不図示のマウス等のポインティングデバイス等の入力装置からの入力を制御する。入力装置がタッチパネルの場合、ユーザがタッチパネルに表示されたアイコンやカーソルやボタンに合わせて押下（指等でタッチ）することにより、各種の指示を行うことができることとする。
【００３０】
また、タッチパネルは、マルチタッチスクリーンなどの、複数の指でタッチされた位置

を検出することが可能なタッチパネルであってもよい。
【００３１】
ビデオコントローラ２０７は、ディスプレイ２１２などの外部出力装置への表示を制御する。ディスプレイは本体と一体になったノート型パソコンのディスプレイも含まれるものとする。なお、外部出力装置はディスプレイに限ったものははく、例えばプロジェクタであってもよい。また、前述のタッチ操作を受け付け可能な装置については、入力装置も提供する。
【００３２】
なおビデオコントローラ２０７は、表示制御を行うためのビデオメモリ（ＶＲＡＭ）を制御することが可能で、ビデオメモリ領域としてＲＡＭ２０３の一部を利用することもできるし、別途専用のビデオメモリを設けることも可能である。
【００３３】
メモリコントローラ２０８は、外部メモリ２１３へのアクセスを制御する。外部メモリとしては、ブートプログラム、各種アプリケーション、フォントデータ、ユーザファイル、編集ファイル、および各種データ等を記憶する外部記憶装置（ハードディスク）、フレキシブルディスク（ＦＤ）、或いはＰＣＭＣＩＡカードスロットにアダプタを介して接続されるコンパクトフラッシュ（登録商標）メモリ等を利用可能である。
【００３４】
通信Ｉ／Ｆコントローラ２０９は、ネットワークを介して外部機器と接続・通信するものであり、ネットワークでの通信制御処理を実行する。例えば、ＴＣＰ／ＩＰを用いた通信やＩＳＤＮなどの電話回線、および携帯電話の３Ｇ回線を用いた通信が可能である。
【００３５】
尚、ＣＰＵ２０１は、例えばＲＡＭ２０３内の表示情報用領域へアウトラインフォントの展開（ラスタライズ）処理を実行することにより、ディスプレイ２１２上での表示を可能としている。また、ＣＰＵ２０１は、ディスプレイ２１２上の不図示のマウスカーソル等でのユーザ指示を可能とする。
【００３６】
本発明を実現するための後述する各種プログラムは、外部メモリ２１３に記憶されており、必要に応じてＲＡＭ２０３にロードされることによりＣＰＵ２０１によって実行されるものである。さらに上記プログラムの実行時に用いられる定義ファイル及び各種情報テーブル等も外部メモリ２１３に格納されており、これらについての詳細な説明も後述する。
【００３７】
次に図３を用いて、本願発明における処理の流れについて説明する。
【００３８】
ステップＳ３０１では、事前設定として、カテゴリ毎のフィールド抽出定義情報（フィールド重みを含む）とデフォルトの抽出定義（カテゴリが設定されていないファイルやフィールド重みセットで指定しなかったフィールドに使う抽出定義）の設定を受け付ける。フィールド抽出定義情報とは、抽出定義詳細画面１３００に示すように、カテゴリ毎に、フィールド名と当該フィールドを抽出する方法と抽出定義が対応付けられた情報である。例えば図１３に示す抽出定義情報によれば、「工事概要」というカテゴリの文書については、「事務所」や「病院」といったキーワードにより抽出されるフィールドを「建物用途」というフィールドとして抽出することが可能となる。
【００３９】
設定された抽出定義情報は、抽出定義ＤＢ１０５に保存される。
【００４０】
ステップＳ３０２では、ユーザから受け付けた文書（検索対象文書）に対して、文書登録処理を実行する。文書登録処理では、検索対象文書の本文抽出やカテゴリの付与、検索対象文書のフィールド抽出、本文及びフィールドに対する検索インデックスの構築などが行われる。文書登録処理の詳細については、図４を用いて後述する。
【００４１】

ステップＳ３０３では、ユーザから受け付けた検索語に基づき、文書検索処理を実行する。文書検索処理では、ステップＳ３０２で構築した検索インデックスを用いた検索処理が行われる。文書検索処理の詳細については、図９を用いて後述する。

【００４２】

次に図４～図８のフローチャートを用いて、本発明の実施形態における文書登録処理部が実行する文書登録処理について説明する。

【００４３】

図４のフローチャートは、文書登録処理部１０１において文書を登録する処理を示すフローチャートである。

【００４４】

ステップＳ４０１では、登録対象となる文書全てに対して処理が終了したかどうかを判定する。処理が終了していれば（Ｓ４０１のＹｅｓ）該フローチャートの処理を終了し、処理の終了していない文書が残っていれば（Ｓ４０１のＮｏ）ステップＳ４０２に進む。

【００４５】

ステップＳ４０２では該文書に対してテキスト抽出処理を行う。該テキスト抽出処理は一般に開示されている技術により実現されるものであり、どのような技術・方法を用いても構わない。

【００４６】

ステップＳ４０３では該文書に対するカテゴリ付与を行う。カテゴリとは、その文書がいかなるタイプの文書であるかを分類するために付与され、本実施例であれば工事概要、注文書、議事録などがカテゴリの分類例である。ここでのカテゴリ付与は計算機によって自動で行ってもよいし、ユーザによって手動で行っても構わない。

【００４７】

ステップＳ４０４ではフィールド抽出処理を行う。フィールド抽出処理については、図５を使い後述する。

【００４８】

ステップＳ４０５ではステップＳ４０２で抽出したテキスト及びステップＳ４０４で抽出した各フィールドに対する検索インデックスの作成を行い登録文書インデックス１０４に格納する。検索インデックスとは、図９で示す文書検索処理の処理時に使用する検索インデックスである。

【００４９】

図５のフローチャートは、文書からフィールドを抽出する処理を示すフローチャートである。

【００５０】

ステップＳ５０１では、ステップＳ４０３で付与された該文書のカテゴリを取得する。

【００５１】

ステップＳ５０２では、抽出定義ＤＢ１０５からステップＳ３０１で設定された該カテゴリの抽出定義情報を取得する。ステップＳ５０１でカテゴリが取得できなかった場合はデフォルトの抽出定義を取得する。

【００５２】

ステップＳ５０３では、該抽出定義に定義された全てのフィールドに対して抽出処理が終了したかどうかを判断する。終了していれば（Ｓ５０３のＹｅｓ）該フローチャートを終了し、そうでなければ（Ｓ５０３のＮｏ）処理をステップＳ５０４に進める。

【００５３】

ステップＳ５０４では、該抽出定義情報に設定された処理対象のフィールドの抽出方式に応じて、処理を分岐する。例えば、図１３の例では、「住所」のフィールドについては形態素解析により抽出することを意味している。抽出方式が「キーワード」であればステップＳ５０５に、「パターン」であればステップＳ５０６に、「形態素解析」であればステップＳ５０７に処理を進める。

【００５４】

ステップＳ５０５では、キーワードによる抽出処理を行う。キーワードによる抽出処理の詳細は、図６のフローチャートを用いて後述する。

【００５５】

ステップＳ５０６では、パターンによる抽出処理を行う。パターンによる抽出処理の詳細は、図７のフローチャートを用いて後述する。

【００５６】

ステップＳ５０７では、形態素解析による抽出処理を行う。形態素解析による抽出処理の詳細は、図８のフローチャートを用いて後述する

ステップＳ５０８では、抽出されたフィールドを該文書のフィールドとして記録しておく。このとき、該抽出定義情報のフィールド名と関連付けて記録する。

【００５７】

図６のフローチャートは、文書からキーワード方式でフィールドを抽出する処理を示すフローチャートである。

【００５８】

ステップＳ６０１では、該抽出定義の全てのキーワードを処理したかどうかを判断する。全て処理していれば（Ｓ６０１のＹｅｓ）該フローチャートを終了し、そうでなければ（Ｓ６０１のＮｏ）処理をステップＳ６０２に進める。

【００５９】

ステップＳ６０２では、該抽出定義から未処理のキーワードを取得する。

【００６０】

ステップＳ６０３では、該文書に対するキーワードマッチを実行する。このキーワードマッチにはどのような手法を用いても構わない。

【００６１】

ステップＳ６０４では、該文書に該キーワードが存在するかどうかを判断する。存在しない場合（Ｓ６０４のＮｏ）処理をステップＳ６０１に進め、存在する場合（Ｓ６０４のＹｅｓ）処理をステップＳ６０５に進める。

【００６２】

ステップＳ６０５では、存在キーワードの近くにフィールド名が存在するかどうかを判定する。キーワードによる抽出処理に関する抽出定義は、図１３に示すように、フィールド名とキーワードとが対応付けて登録されたものである。検出されたキーワードの近くに、当該キーワードに対応付けられたフィールド名が存在する場合（ステップＳ６０５：ＹＥＳ）は処理をステップＳ６０６に進め、存在しない場合（ステップＳ６０５：ＮＯ）は処理をステップＳ６０１に戻す。

【００６３】

ここで、キーワード（Ｖａｌｕｅ）とフィールド名（Ｋｅｙ）の距離について、図１４を用いて具体的に説明する。

【００６４】

図１３のようにフィールド名「建物用途」には事務所、病院、飲食店、駐車場、ホテルの５つのキーワードが対応付けられているため、図１４に示す文書において抽出されるキーワードは、Ｖ１「病院」、Ｖ２「事務所」、Ｖ３「駐車場」となる。このうちＶ１とＶ２はキーであるＫ１「建物用途」と同じ行にあり、距離的に近いと言える。一方でＶ３はＫ１と５行離れており、距離的には遠く、「建物用途」とは異なる文脈で使用されていると考えられる。したがって、キーワード抽出の際にはこのキーワードとフィールド名の距離を考慮し、遠いものを抽出対象としないようにすることで誤抽出を防ぐことができる。

【００６５】

図６の説明に戻る。

【００６６】

ステップＳ６０６では、ステップＳ６０５で抽出されたキーワードを該文書のフィールドの値として記録する。

【００６７】

図７は、正規表現パターンによるフィールドの抽出処理を示すフローチャートである。
【００６８】
ステップＳ７０１では、抽出定義情報に設定された正規表現パターンを取得する。正規表現パターンの例としては、図１３のフィールド名１３０２の関連法令であれば、抽出定義１３０４の「．＋（法｜条例）」となる。これは、「法」または「条例」が後方一致する文字列を検出するための正規表現であり、この条件によれば例えば、「建築基準法」「騒音対策条例」などが抽出可能となる。
【００６９】
ステップＳ７０２では、該文書に対してステップＳ７０１で取得した正規表現のパターンマッチを行う。
【００７０】
ステップＳ７０３では、ステップＳ７０２でマッチした部分全てについて処理が行われたかどうかを判断する。全てのパターンで処理が行われた場合（Ｓ７０３のＹｅｓ）該フローチャートの処理を終了し、そうでない場合（Ｓ７０３のＮｏ）ステップＳ７０４へ処理を進める。
【００７１】
ステップＳ７０４では、マッチした部分を該文書のフィールドの値として記録する。また、グループや名前付き前方参照といった正規表現の機能を用いてマッチした部分の一部をフィールドの値として使ってもよい。
【００７２】
図８は、形態素解析で得られた品詞によるフィールドの抽出処理を示すフローチャートである。
【００７３】
ステップＳ８０１では、抽出定義を取得する。例えば、本実施例であれば、図１３のフィールド名１３０２の住所の抽出定義１３０４に定められる品詞の並びを取得する。この場合であれば、抽出定義は［名詞－固有名詞－地域］の並びで定められている。つまり、これは、名詞の中の固有名詞の中の地域カテゴリに属する単語の並びを抽出することを意味し、「東京都港区港南」といった文字列が抽出される。
【００７４】
ステップＳ８０２では、該文書に形態素解析を実行する。
【００７５】
ステップＳ８０３では、ステップＳ８０１で取得した抽出定義に合致する品詞の並びがあるかどうかを判断する。品詞の並びがない場合（Ｓ８０３のＮｏ）該フローチャートの処理を終了し、そうでない場合（Ｓ８０３のＹｅｓ）処理をステップＳ８０４に進める。
【００７６】
ステップＳ８０４では、マッチした部分を該文書のフィールドの値として記録する。
【００７７】
続けて、図９、図１７を用いて、本発明の実施形態における文書検索処理部が実行する処理について説明する。
【００７８】
図９は、検索処理部１０２において、ユーザからの検索語を入力として受けとり、インデックス済みの文書を検索する処理を示すフローチャートである。
【００７９】
ステップＳ９０１では、ユーザからの検索語を取得する。
【００８０】
ステップＳ９０２では、インデックス済みの全文書に対して文書スコアが未計算の文書が存在するかどうかを判断する。文書スコアが未計算の文書が存在する場合（Ｓ９０２のＹｅｓ）処理をステップＳ９０３に進め、そうでない場合（Ｓ９０２のＮｏ）処理をステップＳ９０８に進める。
【００８１】

ステップＳ９０３では、文書スコア未計算の文書を取得する。
【００８２】
ステップＳ９０４では、該文書の本文に対する検索スコアを計算する。検索スコアとは、検索語との関連度合いを数値で表した値である。本文に対する検索スコアを、本文スコアと呼ぶ。なお、本実施例においては、本文スコアは公知の検索スコア算出方法により算出される値とする。
【００８３】
ステップＳ９０５では、フィールドスコアが未計算のフィールドが存在するかどうかを判断する。存在する場合（Ｓ９０５のＹｅｓ）処理をステップＳ８０６に進め、そうでない場合（Ｓ９０５のＮｏ）処理をステップＳ９０７に進める。
【００８４】
ステップＳ９０６では、フィールドスコア未計算のフィールドを取得し、該フィールドに対する検索スコアを計算する。このスコアをフィールドスコアと呼ぶ。
【００８５】
フィールドスコアの計算の方法の一例を、図１７を用いて説明する。ユーザから「ＡＡＡ株式会社　大阪」という検索語を受け付けた場合について説明する。。
【００８６】
図１７Ａは、大阪府警担当者議事録というタイトルの文書を示した図で、当該文書をフィールド毎に分け、各フィールドの値と重みが対応付けられている。図１７Ｂは、○○プロジェクト概要というタイトルの文書を示した図で、図１７Ａと同様に、フィールド毎に値と重みとが対応付けてある。なお、重みは、当該文書のカテゴリによって定まる値である。なお、図17において各フィールドの値として示している内容は、説明の為に抽出定義に合致しない文字列も含めて示しているが、ステップS506、ステップS604、ステップS704で説明した通り、各フィールドの値として登録されるのは、抽出定義に合致した文字列である。
【００８７】
まず、検索語の出現回数をフィールド毎にカウントする。
【００８８】
図１７Ａの文書であれば、タイトルフィールド１８０３には「大阪」は１回出現、人名フィールド１８０４には「大阪」は０回出現、本文フィールド１７０５には「大阪」は３回出現している。そして、各フィールドでの検索語の出現回数をフィールド毎に設定されている重みとをかけてフィールドスコアを求める。
【００８９】
タイトルフィールド１８０３に設定されている重みは１８０６に示すように２で大阪は１回出現なので、１×２＝２となる。同様に、人名フィールド１８０４は０×５＝０、本文フィールド１８０５は３×１＝３となる。これらの合計値（２＋０＋３＝５）が「大阪府警担当者議事録」という文書のフィールドスコアとして算出される。
【００９０】
同様に図１７Ｂの、○○プロジェクト概要．ＰＤＦのフィールドスコアを計算すると、会社名フィールドで、検索語ＡＡＡ株式会社が１回出現しているので１×５＝５、住所フィールドで大阪が１回出現しているので１×５＝５、本文フィールドでＡＡＡ株式会社と大阪がそれぞれ１回ずつ出現しているので２×１＝２となる。これらの合計値（５＋５＋２＝１２）が○○プロジェクト概要．ＰＤＦのフィールドスコアとして算出される。
【００９１】
ステップＳ９０７では、ステップＳ９０４で算出した該文書の本文スコアと、ステップＳ９０６で算出した該文書のフィールドスコアを合算する。この値を文書スコアと呼ぶ。
【００９２】
なお、本実施例においては、本文スコアとフィールドスコアとを合算したスコアを文書スコアとしたが、各フィールドの重みを考慮したスコアであるフィールドスコアのみを用いても良い。

【００９３】
ステップＳ９０８では、文書スコアの降順で検索結果をユーザに示す。なお、本実施例では検索語との関連性が強い文書の文書スコアが高くなる計算方法を用いたため、降順で検索結果をユーザに示したが、検索語との関連性が強い文書の文書スコアが小さくなる算出方法を用いる場合は、昇順により表示する。すなわち、検索語との関連性が強い文書が検索結果の上位に表示されるようソートして表示する。
【００９４】
以上のように、抽出定義情報で「人名」や「会社名」や「住所」など、当該カテゴリの文書を特徴付けるフィールドに対して大きな重みを設定し、設定されたフィールド毎の重みを考慮して検索スコアを算出することで、検索語が同じ数だけ含まれる文書であっても、よりユーザ（検索者）の意図に合った（ユーザが探し求めている）文書を上位に表示することが可能となる。
【００９５】
ステップＳ９０９では、検索セッション統計情報更新処理を行う。図１０のフローチャートを用いて後述する。
【００９６】
ステップＳ９１０では、フィールド重み更新処理を行う。図１１のフローチャートを用いて後述する。
【００９７】
図１０は、ユーザの検索セッションでの統計情報を更新する処理を示すフローチャートである。なお、検索セッションとはユーザが検索結果を取得して、該検索結果を破棄するまでの期間のことを言う。
【００９８】
ステップＳ１００１では、検索セッション統計情報テーブル図１５の１４００の初期化を行う。該検索結果に含まれる全ての文書情報について、文書ＩＤ、カテゴリを設定しセッション閲覧数を０に設定する。
【００９９】
ステップＳ１００２では、検索セッションが終了しているかどうかを判断する。終了している場合（Ｓ１００２のＹｅｓ）該フローチャートの処理を終了し、そうでない場合（Ｓ１００２のＮｏ）ステップＳ１００３に処理を進める。
【０１００】
ステップＳ１００３では、ユーザが検索結果の文書を選択したかどうかを判断する。選択していない場合（Ｓ１００３のＮｏ）処理をステップＳ１００２に進め、そうでない場合（Ｓ９０３のＹｅｓ）は処理をステップＳ１００４に進める。
【０１０１】
ステップＳ１００４では、ユーザが選択した文書の情報を取得する。
【０１０２】
ステップＳ１００５では、検索セッション統計情報テーブルの該文書のエントリを更新する。この場合、該テーブルのセッション閲覧数に１を加える。
【０１０３】
図１１は、検索セッション統計情報を利用して抽出定義のフィールド重みを更新する処理を示すフローチャートである。検索の情報に応じてフィールド重みを更新していくことで、より検索精度が向上していくことが見込まれる。
【０１０４】
ステップＳ１１０１では、検索セッション統計情報テーブル図１５の１４００を取得する。なお、ここで取得するのは検索セッションの終了した検索セッション統計情報テーブルのみである。
【０１０５】
ステップＳ１１０２では、ヒット文書のカテゴリごとに閲覧数を集計する。ヒット文書とは、検索処理部により検索された文書である。検索の結果ヒットした文書をユーザが閲

覧したかを集計することで、次回の検索精度を上げるために利用される。
【０１０６】
ステップＳ１１０３では、ステップＳ１１０２で集計したカテゴリの中に未処理のカテゴリがあるかどうかを判断する。未処理のカテゴリがある場合（Ｓ１１０３のＹｅｓ）処理をステップＳ１００４に進め、そうでない場合（Ｓ１１０３のＮｏ）処理をステップＳ１１０９に進める。
【０１０７】
ステップＳ１１０４では、未処理のカテゴリの抽出定義を取得する。
【０１０８】
ステップＳ１１０５では、検索語に含まれる未処理のフィールド情報（当該カテゴリの抽出定義として設定されたフィールドのうち、検索語として用いられたワードが該当するフィールドであって、未処理のフィールド）があるかどうかを判断する。未処理のフィールド情報がある場合（Ｓ１１０５のＹｅｓ）処理をステップＳ１１０６へ進め、そうでない場合（Ｓ１１０５のＮｏ）処理をステップＳ１１０３に進める。
【０１０９】
ステップＳ１１０６では、該カテゴリのセッション閲覧数が０より大きいかどうかを判断する。０より大きい場合（Ｓ１１０６のＹｅｓ）処理をステップＳ１１０７に進め、そうでない場合（Ｓ１１０６のＮｏ）処理をステップＳ１１０８に進める。
【０１１０】
ステップＳ１１０７では、該カテゴリのセッション閲覧数が０より大きく、該フィールドが検索に貢献できたと考え、該フィールドのフィールド重みを（セッション閲覧数）×０．０１だけ加算する。この計算式はあくまでも一例であり、その他の計算方法を用いても構わない。
【０１１１】
ステップＳ１１０８では、該カテゴリのセッション閲覧数が０であり、該フィールドが検索に貢献していないと考え、該フィールドのフィールド重みを０．０１だけ減算する。この計算式はあくまでも一例であり、その他の計算方法を用いても構わない。
【０１１２】
ステップＳ１１０９では、不要となった該検索セッション統計情報テーブルを破棄する。
【０１１３】
ここで、図１１を用いて、フィールド重み更新処理の一例を説明する。まず、検索語に「住所」と「建物用途」を含む検索語が使われたとし、検索セッション終了時の検索セッション統計情報テーブルが図１６の１４００であったとする。また、カテゴリ「工事概要」「注文書」「議事録」の抽出定義がそれぞれ、１５００、１６００、１７００であったとする。またフィールド重み更新式はセッション閲覧数が０より大きい場合は（セッション閲覧数）×０．０１を加算、０の場合は０．０１の減算とする。この場合、テーブル１３００より検索結果のカテゴリごとのセッション閲覧数は、工事概要が２、注文書と議事録が０となる。工事概要のフィールド重みの更新は、フィールド「住所」（図１６の１５０１）と「建物用途」（図１６の１５０２）が両方とも定義されていることから、２×０．０１＝０．０２が加算され、更新後のフィールド重みはそれぞれ３．０２と２．０２となる。注文書のフィールド重みの更新は「住所」（図１６の１６０１）のみが定義されていることから、０．０１の減算となり、更新後のフィールド重みは０．０９となる。議事録のフィールド重みの更新は「住所」「建物用途」ともに定義されていないため行われない。
【０１１４】
図１２は、現在定義されている抽出定義の確認と、追加、削除を行う画面である。抽出定義一覧画面１２００は抽出定義追加ボタン１２０１、一括削除ボタン１２０２、チェックボックス１２０３、編集ボタン１２０４、個別削除ボタン１２０５からなる。
【０１１５】
抽出定義追加ボタン１２０１は、押下することで抽出定義詳細画面（図１３）に遷移し

、新規に抽出定義を作成するためのものである。
【０１１６】
一括削除ボタン１２０２は、押下することでチェックボックス１２０３が有効になっている全ての抽出定義を一括削除するものである。
【０１１７】
チェックボックス１２０３は、有効にすることで一括削除ボタン１２０２を用いて一括削除を行えるようにするためのものである。
【０１１８】
編集ボタン１２０４は、押下することで抽出定義詳細画面（図１３）に遷移し、選択した抽出定義を編集するためのものである。
【０１１９】
個別削除ボタン１２０５は、押下することで選択した抽出定義を削除するためのものである。
【０１２０】
図１３は、抽出定義の詳細の追加、確認、編集を行う画面である。抽出定義詳細画面１３００は、カテゴリ名テキストボックス１３０１、フィールド名テキストボックス１３０２、抽出方式プルダウンリスト１３０３、抽出定義テキストボックス１３０４、フィールド重みテキストボックス１３０５、フィールド削除ボタン１３０６、抽出定義フィールド追加ボタン１３０７からなる。
【０１２１】
なお、抽出定義一覧画面１１００の抽出定義追加ボタン１１０１を押下して本画面に遷移した場合は、カテゴリ名テキストボックス１３０１は空欄で、フィールド名テキストボックス１３０２、抽出方式プルダウンリスト１３０３、抽出定義テキストボックス１３０４、フィールド重みテキストボックス１３０５は初期状態では表示されていない。また、抽出定義一覧画面１２００の編集ボタンから本画面に遷移した場合、該抽出定義の内容がカテゴリ名テキストボックス１３０１、フィールド名テキストボックス１３０２、抽出方式プルダウンリスト１３０３、抽出定義テキストボックス１３０４、フィールド重みテキストボックス１３０５に表示される。
【０１２２】
カテゴリ名テキストボックス１３０１は、この抽出定義につける名称を設定するためのものである。
【０１２３】
フィールド名テキストボックス１３０２は、フィールドの名称を設定するためのものである。
【０１２４】
抽出方式プルダウンリスト１３０３は、抽出方式を選択するためのものである。ここでは「キーワード」「パターン」「形態素解析」から選択する。
【０１２５】
抽出定義テキストボックス１３０４は、抽出の定義を設定するためのものである。抽出方式が「キーワード」の場合は抽出するキーワードのリスト、「パターン」の場合は正規表現パターン、「形態素解析」の場合は抽出したい形態素の並びを設定する。
【０１２６】
フィールド重みテキストボックス１３０５は、フィールド重みを設定するためのものである。
【０１２７】
フィールド削除ボタン１３０６は、押下することで該フィールドの抽出定義を削除するためのものである。
【０１２８】
抽出定義フィールド追加ボタン１３０７は、押下することで空欄のフィールド名テキストボックス１３０２、抽出方式プルダウンリスト１３０３、抽出定義テキストボックス１

３０４、フィールド重みテキストボックス１３０５、フィールド削除ボタン１３０６が最下行に追加され新しいフィールドの定義ができるようになる。

【０１２９】

このようにして、カテゴリごとに抽出定義を設定することにより検索精度の向上が見込まれる。例えば、登録文書内の建築設計書の工事概要と注文書を比較した場合、工事概要の住所（建設場所）の情報は地形や適用される自治体の条例が異なるなど非常に重要な項目であるが、注文書の住所は特に重要な情報でないため、工事概要ではフィールド重みを高め（例えば３）に、注文書では低め（例えば０．１）に設定することで、同じフィールドでのカテゴリごと重要度の違いを表現できる。このように設定することで、住所で検索を行った場合、検索スコアが高めになる工事概要が検索結果上位に、検索スコアが低めになる注文書は検索下位に表示されることが見込め、検索ユーザの意図に沿った検索結果となりやすい。

【０１３０】

仮にカテゴリごとにフィールド重みを設定しなかった場合、住所で検索した場合、どのカテゴリの文書でも住所を重視するよう設定した場合、工事概要と注文書の両方が検索結果に混在することになり利便性が低下すると考えられる。

【０１３１】

図１５の検索セッション統計情報テーブル１４００は、検索セッションの統計情報を保持するためのテーブルであり、文書ＩＤ１４０１、カテゴリ１４０２、セッション閲覧数１４０３の項目からなる

文書ＩＤ１４０１には、検索でヒットした文書を特定するための項目であり、ヒットした文書のＩＤが登録される。

【０１３２】

カテゴリ１４０２には、該文書のカテゴリが登録される。

【０１３３】

セッション閲覧数１４０３には、ユーザが検索セッション中に該文書を閲覧した回数を記録する。

【０１３４】

以上、本実施形態について示したが、本発明は、例えば、システム、装置、方法、プログラムもしくは記録媒体等としての実施態様をとることが可能である。具体的には、複数の機器から構成されるシステムに適用しても良いし、また、一つの機器からなる装置に適用しても良い。

【０１３５】

また、本発明におけるプログラムは、図３～図１１に示すフローチャートの処理方法をコンピュータが実行可能なプログラムであり、本発明の記憶媒体は図３～図１１の処理方法をコンピュータが実行可能なプログラムが記憶されている。

【０１３６】

以上のように、前述した実施形態の機能を実現するプログラムを記録した記録媒体を、システムあるいは装置に供給し、そのシステムあるいは装置のコンピュータ（またはＣＰＵやＭＰＵ）が記録媒体に格納されたプログラムを読み出し、実行することによっても本発明の目的が達成されることは言うまでもない。

【０１３７】

この場合、記録媒体から読み出されたプログラム自体が本発明の新規な機能を実現することになり、そのプログラムを記録した記録媒体は本発明を構成することになる。

【０１３８】

プログラムを供給するための記録媒体としては、例えば、フレキシブルディスク、ハードディスク、光ディスク、光磁気ディスク、ＣＤ－ＲＯＭ、ＣＤ－Ｒ、ＤＶＤ－ＲＯＭ、磁気テープ、不揮発性のメモリカード、ＲＯＭ、ＥＥＰＲＯＭ、シリコンディスク等を用いることが出来る。

【０１３９】

また、コンピュータが読み出したプログラムを実行することにより、前述した実施形態の機能が実現されるだけでなく、そのプログラムの指示に基づき、コンピュータ上で稼働しているＯＳ（オペレーティングシステム）等が実際の処理の一部または全部を行い、その処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【０１４０】

さらに、記録媒体から読み出されたプログラムが、コンピュータに挿入された機能拡張ボードやコンピュータに接続された機能拡張ユニットに備わるメモリに書き込まれた後、そのプログラムコードの指示に基づき、その機能拡張ボードや機能拡張ユニットに備わるＣＰＵ等が実際の処理の一部または全部を行い、その処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【０１４１】

また、本発明は、複数の機器から構成されるシステムに適用しても、ひとつの機器から成る装置に適用しても良い。また、本発明は、システムあるいは装置にプログラムを供給することによって達成される場合にも適用できることは言うまでもない。この場合、本発明を達成するためのプログラムを格納した記録媒体を該システムあるいは装置に読み出すことによって、そのシステムあるいは装置が、本発明の効果を享受することが可能となる。

【０１４２】

さらに、本発明を達成するためのプログラムをネットワーク上のサーバ、データベース等から通信プログラムによりダウンロードして読み出すことによって、そのシステムあるいは装置が、本発明の効果を享受することが可能となる。なお、上述した各実施形態およびその変形例を組み合わせた構成も全て本発明に含まれるものである。

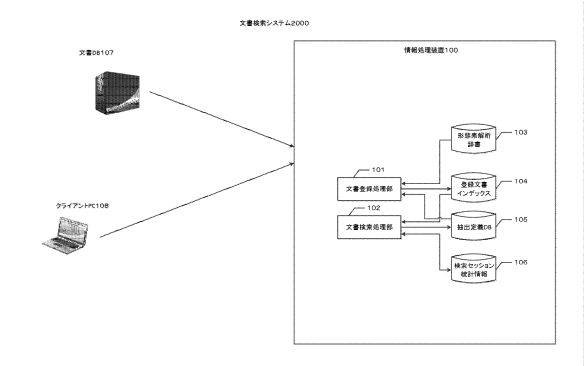
【符号の説明】

【０１４３】

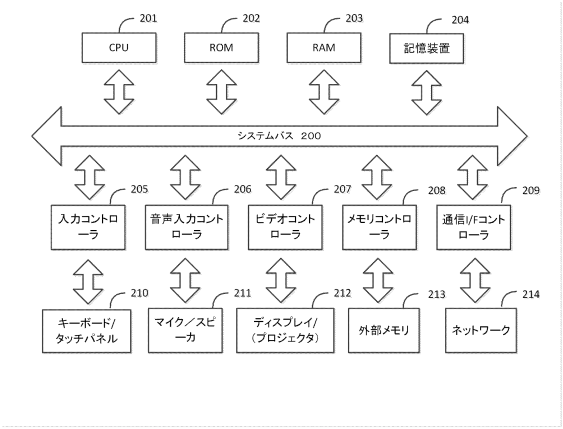
２０００　文書検索システム
１００　情報処理装置
１０１　文書登録処理部
１０２　文書検索処理部
１０３　形態素解析辞書
１０４　登録文書インデックス
１０５　抽出定義ＤＢ
１０６　検索セッション統計情報
１０７　文書ＤＢ
１０８　クライアントＰＣ

【図面】

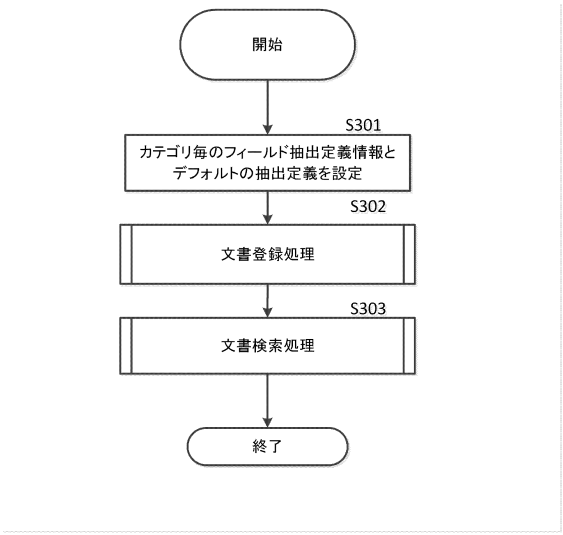
【図１】


文書検索システム2000
文書DB107
情報処理装置100
クライアントPC108
101
文書登録処理部
102
文書検索処理部
103
104
105
106

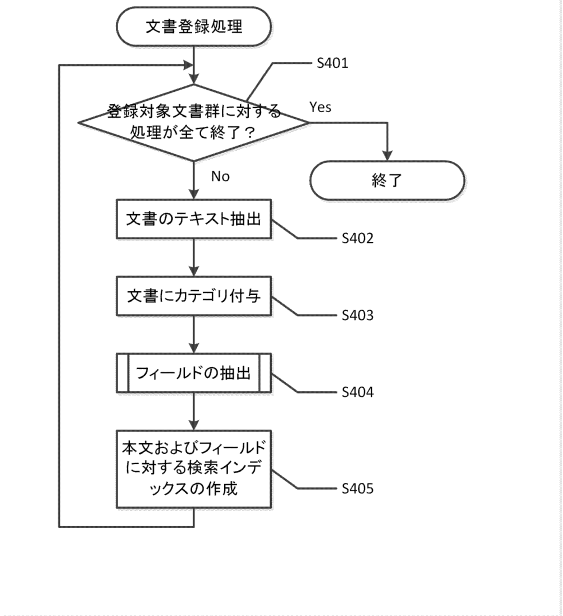
【図２】


201
CPU
202
ROM
203
RAM
204
記憶装置
システムバス 200
205
入力コントローラ
206
音声入力コントローラ
207
ビデオコントローラ
208
メモリコントローラ
209
通信I/Fコントローラ
210
キーボード/タッチパネル
211
マイク／スピーカ
212
ディスプレイ/（プロジェクタ）
213
外部メモリ
214
ネットワーク

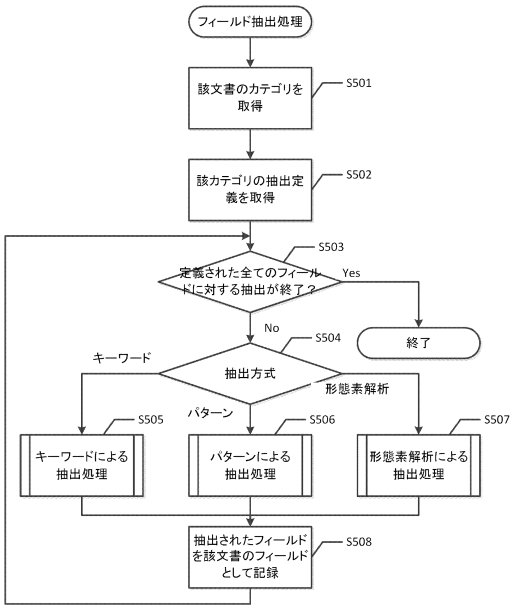
【図３】


開始
S301
カテゴリ毎のフィールド抽出定義情報とデフォルトの抽出定義を設定
S302
文書登録処理
S303
文書検索処理
終了

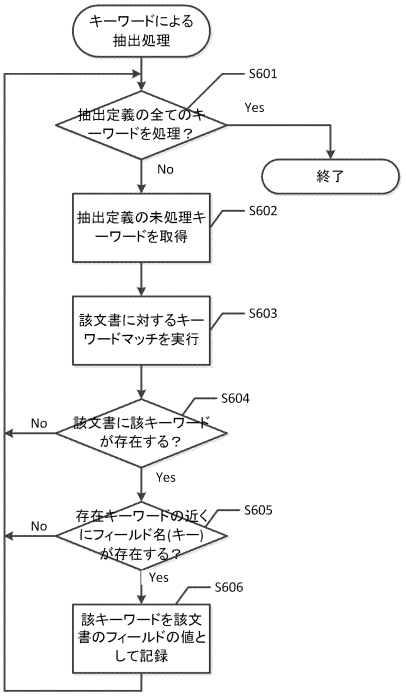
【図４】


文書登録処理
S401
登録対象文書群に対する処理が全て終了？
Yes
No
終了
文書のテキスト抽出
S402
文書にカテゴリ付与
S403
フィールドの抽出
S404
本文およびフィールドに対する検索インデックスの作成
S405

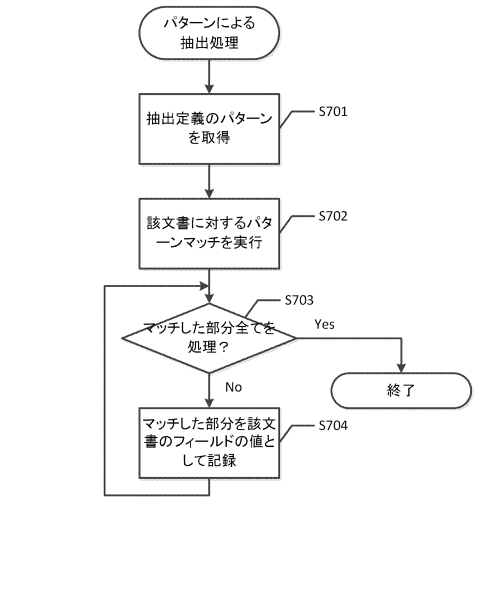
【図５】


フィールド抽出処理
該文書のカテゴリを取得
S501
該カテゴリの抽出定義を取得
S502
S503
定義された全てのフィールドに対する抽出が終了？
Yes
終了
No
S504
抽出方式
キーワード
パターン
形態素解析
S505
S506
S507
キーワードによる抽出処理
パターンによる抽出処理
形態素解析による抽出処理
抽出されたフィールドを該文書のフィールドとして記録
S508

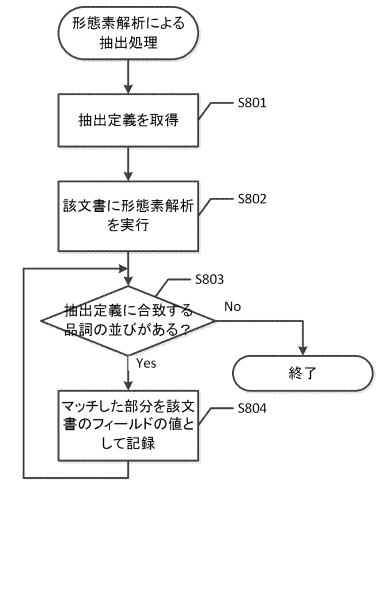
【図６】


キーワードによる抽出処理
S601
抽出定義の全てのキーワードを処理？
Yes
終了
No
抽出定義の未処理キーワードを取得
S602
該文書に対するキーワードマッチを実行
S603
S604
該文書に該キーワードが存在する？
No
Yes
S605
存在キーワードの近くにフィールド名(キー)が存在する？
No
Yes
S606
該キーワードを該文書のフィールドの値として記録

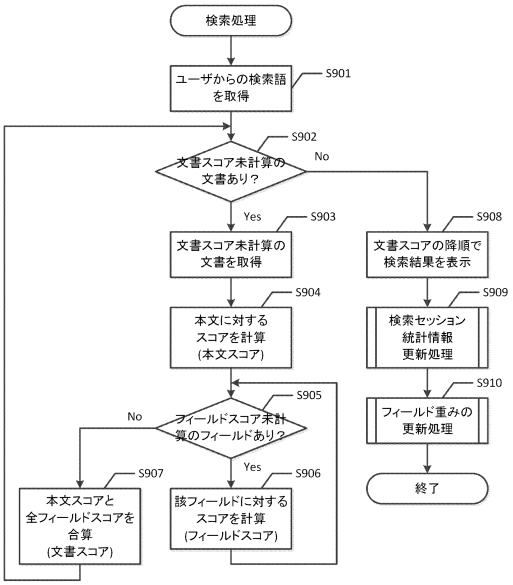
【図７】


パターンによる抽出処理
抽出定義のパターンを取得
S701
該文書に対するパターンマッチを実行
S702
S703
マッチした部分全てを処理？
Yes
終了
No
マッチした部分を該文書のフィールドの値として記録
S704

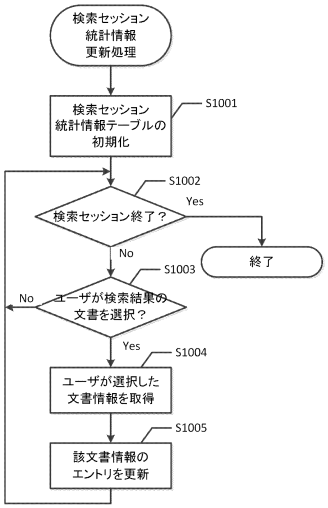
【図８】


形態素解析による抽出処理
抽出定義を取得
S801
該文書に形態素解析を実行
S802
S803
抽出定義に合致する品詞の並びがある？
No
Yes
終了
マッチした部分を該文書のフィールドの値として記録
S804

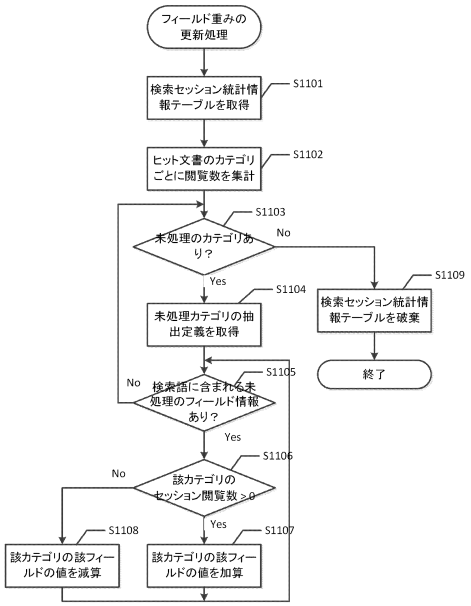
【図９】



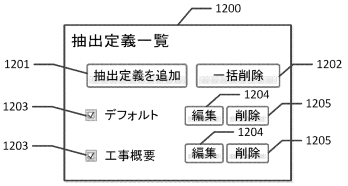
【図１０】



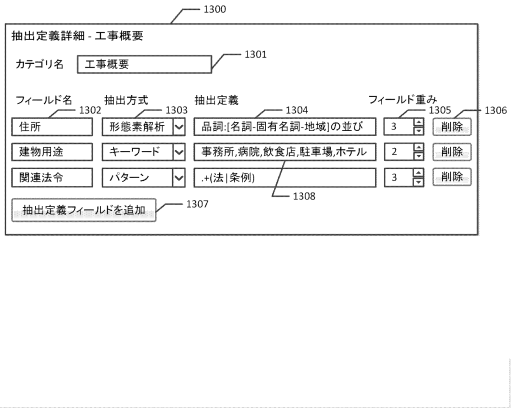
【図１１】



【図１２】

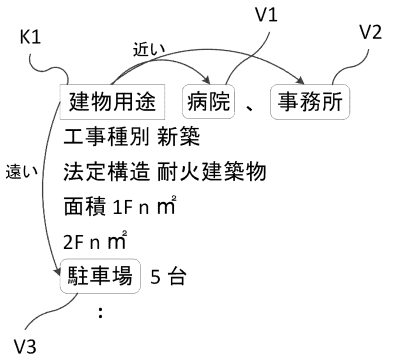


【図１３】

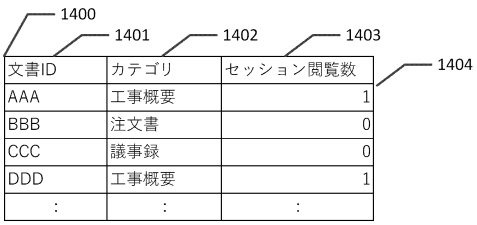


【図１４】

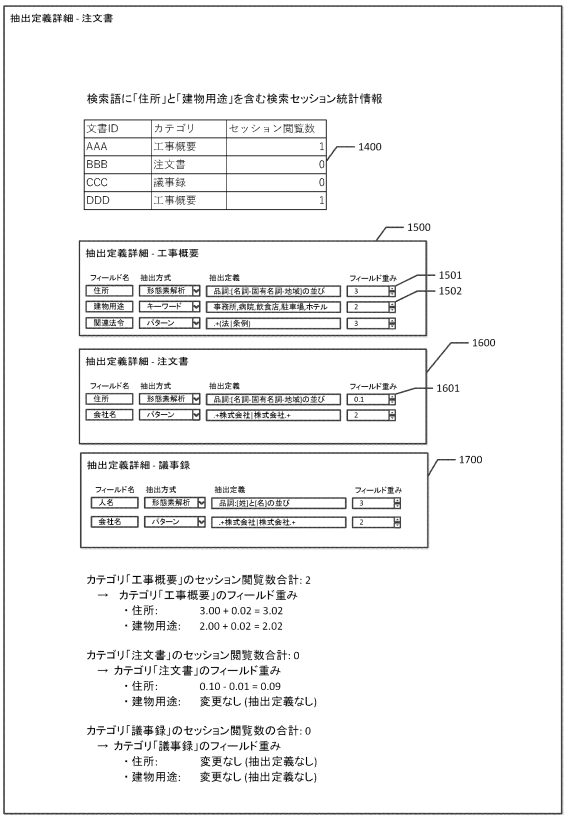
建物用途	病院、事務所	
工事種別	新築	
法定構造	耐火建築物	
面積	1F	n ㎡
	2F	n ㎡
駐車場	5 台	
:	:	



【図１５】



【図１６】



【図１７】

1800 1801 図17A 1802

フィールド	値	重み
タイトル	大阪府警担当者議事録.docx	2
人名	佐藤 一郎 高橋 二郎	5
本文	大阪府警担当者議事録 日時: 20XX/XX/XX 出席者: 佐藤 一郎 (大阪府警)、 高橋 二郎 (A社) 場所: 会議室A 議題1: 大阪での○○について …	1

1803 1804 1805 1806 1807 1808

1806 図17B

フィールド	値	重み
タイトル	○○プロジェクト概要.pdf	2
会社名	AAA株式会社	5
人名	鈴木 太郎	5
住所	大阪府○○市○○0-0-0	5
本文	名称: ○○プロジェクト 顧客会社名: AAA株式会社 顧客責任者: 鈴木 太郎 住所: 大阪府○○市○○0-0-0 …	1

フロントページの続き

審査官　酒井 恭信
(56)参考文献　米国特許出願公開第２０１９／０３２５２１２（ＵＳ，Ａ１）
中国特許出願公開第１１０３９００９４（ＣＮ，Ａ）
特開平１０－０４９５４９（ＪＰ，Ａ）
米国特許第０６１５４７３７（ＵＳ，Ａ）
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１６／００ － １６／９５８