(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2017-162346
(P2017-162346A)
(43)公開日　平成29年9月14日(2017.9.14)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 17/30	(2006.01)	G06F	17/30	340B

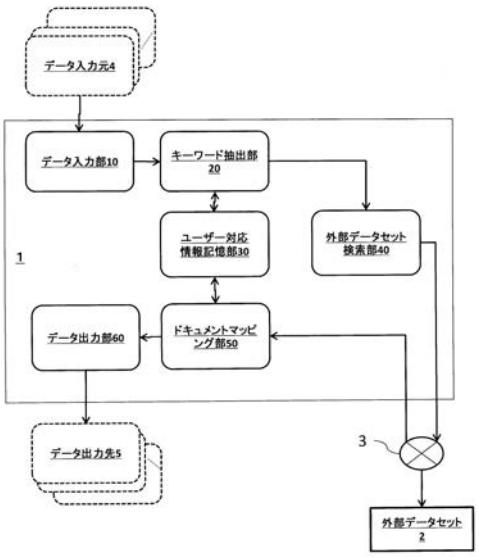
審査請求 未請求　請求項の数 1　OL　（全 13 頁）

(21)出願番号　特願2016-48109(P2016-48109)
(22)出願日　平成28年3月11日(2016.3.11)

(71)出願人　000163833
金剛株式会社
熊本県熊本市西区上熊本３丁目８番１号
(74)代理人　100088856
弁理士　石橋　佳之夫
(72)発明者　長谷川　秀彦
茨城県つくば市天王台１丁目１番１　国立大学法人筑波大学内
(72)発明者　宇陀　則彦
茨城県つくば市天王台１丁目１番１　国立大学法人筑波大学内
(72)発明者　坂井　一隆
熊本県熊本市西区上熊本３丁目８番１号
金剛株式会社内

(54)【発明の名称】情報検索システム

(57)【要約】
【課題】　ユーザーからの直接的な反応入力が無くても最適な検索結果を出力できる情報検索システムを提供する。
【解決手段】　ユーザーからの明示的な検索語の入力を要求せずに検索語に適合する文書情報を検索するために、複数の入力データソースとの接続を選択的に切り換える入力データソースセレクタと、入力データセレクタが生成したテキストデータとユーザー対応情報記憶部に記憶されているデータに基づいて検索語を生成するキーワード抽出部と、複数の外部データセットとの接続を検索語に応じて選択的に切り換える外部データセットセレクタと、複数のドキュメントマッピングを選択的に切り換えるドキュメントマッピングセレクタと、複数の出力先との接続を選択的に切り換える出力コンバータセレクタと、を有する情報検索システムによる。
【選択図】図１



【特許請求の範囲】
【請求項１】
ユーザーからの明示的な検索語の入力を要求せずに前記検索語に適合する文書情報を検索する検索システムであって、
複数の入力データソースとの接続を選択的に切り換える入力データソースセレクタと、
前記入力データソースセレクタが生成したテキストデータとユーザー対応情報記憶部に記憶されているデータに基づいて前記検索語を生成するキーワード抽出部と、
複数の外部データセットとの接続を前記検索語に応じて選択的に切り換える外部データセットセレクタと、
複数のドキュメントマッピングを選択的に切り換えるドキュメントマッピングセレクタと、
複数の出力先との接続を選択的に切り換える出力コンバータセレクタと、
を有することを特徴とする情報検索システム。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、情報検索システムに関するものである。
【背景技術】
【０００２】
デジタルデータ化した文書の情報（以下、文書情報という）を検索する情報検索システムが知られている。情報検索システムは、ユーザーが検索したい文書に適合すると考えた単語等（以下、検索語という）を入力して、予め用意してあるデータ群に対する検索処理を実行させるシステムである。すなわち、従来の情報検索システムでは、ユーザーが能動的に検索語を選択し、かつ、入力する必要がある。
【０００３】
情報検索システムは、概ね以下のように動作する。まず、検索対象とする文書に関連する索引語を予め作成し、情報検索システムの内部に記憶しておく。その後、情報検索システムは、ユーザーからの検索語の入力に応じて、この検索語と上記の索引語とのマッチング処理を実行する。その後、情報検索システムは、検索語にマッチングした索引語に関連付けられている文書情報を「検索結果」としてディスプレイなどの表示装置に出力する。
【０００４】
情報検索システムにおける上記のようなマッチングの精度を高めるには、検索対象とする文書情報から生成される索引語の生成処理を工夫する手法や、検索語として入力される上記の単語等と索引語とのマッチング処理を工夫する手法がある。例えば、索引語に重み付けをして、索引語と検索語とのマッチング結果を索引語の重みに相関させてランク付けする情報検索システムが知られている。また、マッチング結果の方にランク付けを行う手法も複数知られている。
【０００５】
例えば、ユーザーが１回目の検索結果に対して適合する文書情報（ドキュメント）に印をつける（チェックする）手法の場合、チェックした結果を用いて索引語への重み付けを再計算した上で、再検索処理を実行する。このような処理方法を、適合性フィードバックという。適合性フィードバックは、ユーザーが能動的に印を付けなければならないので、検索作業が煩雑になる。
【０００６】
ユーザーが印をつけるのではなく、１回目の計算の上位数件を適合文献とみなして、再計算を行う処理方法も知られている。このような処理方法に係る手法は、疑似的適合性フィードバックと呼ばれる。この擬似的適合性フィードバックには、さらに、適合度の計算において複数の処理方法が知られている。例えば、印が少なくても（フィードバックが少なくても）適合度を上げる処理方法（例えば、非特許文献１を参照）や、閲覧履歴に基づ

いたプロファイルを用いることで個人ごとに適合度をあげる処理方法（例えば、非特許文献２を参照）などが知られている。また、連続して行われる検索質問から推測して先読みを行う処理方法（例えば、非特許文献３を参照）や、潜在情報を用いてドキュメントに現れない単語を推定して計算に用いる処理方法（例えば、非特許文献４を参照）も知られている。

【先行技術文献】

【非特許文献】

【０００７】

【非特許文献１】岩山真、「適合性フィードバックの効率化について」、情報処理学会情報学基礎研究会、2000、FI-57-1、p.1-8

【非特許文献２】杉山一成ほか、「ユーザからの負担なく構築したプロファイルに基づく適応的Web情報検索」、電子情報通信学会論文誌 D-I、2004ｖｏｌ.J87-D-1、no.11、p.975-990

【非特許文献３】藤田遼治ほか、「検索質問と検索結果の変化を利用した先読み検索」、情報処理学会論文誌データベース、2010、vol.3、no.3、p.78-87

【非特許文献４】原島純、黒橋禎夫、「テキストの表層情報と潜在情報を利用した適合性フィードバック」、自然言語処理、2012、vol.19、no.3、p.121-143

【０００８】

従来から、表示された検索結果に対してユーザーが能動的に反応することで、その反応を利用した再検索を行うことで、検索結果の精度向上を図る処理方法は知られている。しかし、ユーザーの明示的な反応を用いることなく検索結果の精度を向上させる処理手法を備える情報検索システムは知られていない。

【発明の概要】

【発明が解決しようとする課題】

【０００９】

本発明は、ユーザーからの直接的な反応を用いずに、ユーザーの間接的な情報の示唆に基づいて、その場に適合する文書情報を自動的に検索し、かつ、次の検索結果の適合度を向上させることができる情報検索システム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、ユーザーからの明示的な検索語の入力を要求せずに前記検索語に適合する文書情報を検索する検索システムであって、複数の入力データソースとの接続を選択的に切り換える入力データソースセレクタと、前記入力データセレクタが生成したテキストデータとユーザー対応情報記憶部に記憶されているデータに基づいて検索語を生成するキーワード抽出部と、複数の外部データセットとの接続を前記検索語に応じて選択的に切り換える外部データセットセレクタと、複数のドキュメントマッピングを選択的に切り換えるドキュメントマッピングセレクタと、複数の出力先との接続を選択的に切り換える出力コンバータセレクタと、を有することを主な特徴とする。

【発明の効果】

【００１１】

本発明によれば、ユーザーからの直接的な反応を用いずに、ユーザーの間接的な情報の示唆に基づいて、その場に適合する文書情報を自動的に検索し、かつ、次の検索結果の適合度を向上させることができる。

【図面の簡単な説明】

【００１２】

【図１】本発明に係る情報検索システムの実施形態である文書情報検索システムの例を示す構成図である。

【図２】上記文書情報検索システムの機能構成の一部を詳細に示す機能構成図である。

【図３】上記文書情報検索システムの機能構成の別の一部を詳細に示す機能構成図である。

【図４】上記文書情報検索システムの機能構成のさらに別の一部を詳細に示す機能構成図である。
【図５】上記文書情報検索システムの機能構成のさらに別の一部を詳細に示す機能構成図である。
【図６】本発明に係る情報検索プログラムの実施形態である文書情報検索プログラムの処理の流れの例を示すフローチャートである。
【図７】上記文書情報検索プログラムの一部における詳細な処理の流れの例を示すフローチャートである。
【図８】上記文書情報検索プログラムの別の一部における詳細な処理の流れの例を示すフローチャートである。
【図９】上記文書情報検索プログラムに対する入力テキストデータの例を示す図である。
【図１０】上記文書情報検索プログラムにおいて処理された上記入力テキストデータの例を示す図である。
【図１１】上記文書情報検索プログラムにおける、（ａ）形態素解析処理の結果の例を示す図と、（ｂ）固有表現抽出処理の結果の例を示す図である。
【図１２】上記文書情報検索プログラムにおいて処理された上記入力テキストデータから生成される検索語の例を示す図である。
【図１３】上記文書情報検索プログラムにおいてユーザー対応情報を更新するためのデータの例を示す図である。
【図１４】上記文書情報検索プログラムにおいて出力される検索結果の例を示す図である。
【発明を実施するための形態】
【００１３】
以下、図面を参照しながら本発明に係る情報検索システムの実施形態として、文書情報検索システム１を例に用いて説明する。
【００１４】
●情報検索システムの構成
図１に示すように文書情報検索システム１は、データ入力部１０と、キーワード抽出部２０と、ユーザー対応情報記憶部３０と、外部データセット検索部４０と、ドキュメントマッピング部５０と、データ出力部６０と、を有してなる。
【００１５】
文書情報検索システム１は、コンピュータ上で実行される文書情報検索プログラム１００（後述する）と、同コンピュータのハードウェア資源によって実現される。したがって、図１に示した上記の各部は、当該コンピュータのハードウェア資源と文書情報検索プログラム１００に係るコンピュータソフトウェアとの協働により実行される各機能ブロックの例である。文書情報検索システム１を構成するコンピュータのハードウェア資源には、例えば、ＣＰＵ、ＲＡＭ、ＲＯＭ、データバス、記憶装置、入力装置、通信装置、出力装置などが含まれる。文書情報検索システム１を動作させる文書情報検索プログラム１００は、不揮発性メモリである上記のＲＯＭの内部や記憶装置に予め記憶されている。この文書情報検索プログラム１００がユーザーの操作によってＲＡＭ上に転送されて、ＣＰＵにおける所定の処理を実行し、以下において説明する各部の機能を提供する。
【００１６】
文書情報検索システム１は、インターネットやＬＡＮ（Ｌｏｃａｌ Ａｒｅａ Ｎｅｔｗｏｒｋ）（以下、ネットワーク３と表記する）に接続されている。文書情報検索システム１は、ネットワーク３を介して外部データセット２と通信可能に構成されている。外部データセット２とは、文書情報検索システム１が情報検索処理を実行するときの検索対象となる情報（データ）群が保持されているコンピュータシステムである。例えば、外部データセット２は、いわゆるインターネット上のデータベースサーバーなどである。ここでは、外部データセット２を、インターネット上に存在する複数のデータベースサーバーを複数束ねた概念上のシステムとする。

【００１７】
また、文書情報検索システム１は、複数のデータ入力元４が接続されている。データ入力元４は、例えば、特定のアプリケーションを用いて作成されたテキストデータ群や、ユーザーの音声を入力データとして用いる場合の音声データなどである。文書情報検索システム１は、これら情報源のそれぞれからの入力を検知し、検知した情報源の特性に応じて、適宜、後段の処理を切り換える。
【００１８】
また、文書情報検索システム１は、複数のデータ出力先５が接続されている。文書情報検索システム１は、出力対象の情報の特性に応じて、適宜、出力先を切り換える。
【００１９】
データ入力元４から入力された情報をデータ入力部１０がテキストデータに変換して、後段のキーワード抽出部２０に渡す。
【００２０】
キーワード抽出部２０は、上記のテキストデータを用いて検索語の候補を複数生成する。生成した検索語候補を所定の処理によって優先順位を付加し、その優先順位に基づいて後段の外部データセット検索部４０に検索語を渡す。また、検索語候補をユーザー対応情報記憶部３０に渡す。
【００２１】
外部データセット検索部４０は、複数の外部データセット２への接続を選択的に切り換える外部データセットセレクタである。外部データセット検索部４０は、前段から渡された検索語に応じて外部データセット２を選択する。
【００２２】
ユーザー対応情報記憶部３０は、文書情報検索システム１を利用可能なユーザー（登録ユーザー）ごとに区別して、当該登録ユーザーによる情報検索のときに用いるキャッシュデータ群を記憶する。ユーザー対応情報記憶部３０には、登録ユーザーごとにユーザー対応記憶領域が設けられる。ユーザー対応記憶領域のそれぞれに記憶されているキャッシュデータは、登録ユーザーの過去の検索結果から機械学習によって得られたデータ群である。以下、このユーザー対応情報記憶部３０に記憶されているデータ群をユーザー対応情報という。
【００２３】
なお、ユーザー対応情報記憶部３０は、文書情報検索システム１が備える外部記憶装置でもよい。この場合、外部記憶装置として、文書情報検索システム１のハードウェアに着脱可能な記憶デバイスを用いれば、複数のユーザー対応情報を適時、交換して使用することができる。
【００２４】
外部データセット検索部４０は、キーワード抽出部２０において生成された検索語を用いて外部データセット２を検索する処理を実行する。
【００２５】
ドキュメントマッピング部５０は、外部データセット検索部４０における検索処理の結果を分析して、ユーザー対応情報記憶部３０の該当するデータを更新する処理を実行する。また、ドキュメントマッピング部５０は、検索結果をユーザーに表示するための出力処理を実行するデータ出力部６０に当該データを渡す処理を実行する。ドキュメントマッピング部５０は、後述するドキュメントマッピングを選択的に切り換えるドキュメントマッピングセレクタである。
【００２６】
データ出力部６０は、ドキュメントマッピング部５０から渡されたデータをデータ出力先５に出力する。データ出力先５は、複数あってもよい。この場合、出力対象のデータの種別を判定し、その種別に合ったデータ出力先５を選択して出力してもよい。データ出力部６０は、複数のデータ出力先５との接続を選択的に切り換える出力コンバータセレクタである。

【００２７】
●情報検索システムの詳細構成
次に、上記の文書情報検索システム１を構成する各機能ブロックの詳細について説明する。図２に示すように、データ入力部１０は、入力データソースセレクタ１０１と、入力フォーマット変換部１０２と、データインポート部１０３と、を有してなる。
【００２８】
入力データソースセレクタ１０１は、データ入力部１０に接続されている複数のデータ入力元４から受け取った情報の種類を判別する処理を実行する。また、入力データソースセレクタ１０１は、判別された各情報に対して情報種類を示す判別情報を付与して後段の入力フォーマット変換部１０２に渡す処理を実行する。また、入力データソースセレクタ１０１は、予め特定の情報種類のみを後段の入力フォーマット変換部１０２に渡すように設定しておき、その設定に合致する情報のみを選択して、後段の入力フォーマット変換部１０２に渡す処理を実行する。
【００２９】
入力フォーマット変換部１０２は、入力データソースセレクタ１０１から受け取った入力情報群と判別情報に基づいて、当該入力情報群をテキストデータへと変換する処理を実行する。例えば、入力情報が音声形式の場合は、音声認識処理を実行することで当該音声形式の入力情報からテキストデータを生成する。入力情報がテキスト形式の場合は、変換処理を行うことなく、後段の処理に用いるテキストデータを生成する。図９に、生成されたテキストデータの例を示す。図９に示すテキストデータの例は、入力情報の形式が通常のテキスト形式の場合の例である。図９に示すように、入力フォーマット変換部１０２において生成されたテキストデータは、自然言語によって表現される文字列である
【００３０】
図２に戻る。データインポート部１０３は、入力フォーマット変換部１０２から受け取ったテキストデータを、所定のデータサイズにして後段のキーワード抽出部２０に渡す処理を実行する。データインポート部１０３における「所定のデータサイズ」とは、後段の処理に渡すテキストデータのデータ量が多すぎると、後段の処理が滞る可能性があることから、これを未然に防ぐために設けるものである。ここでの所定のデータサイズとは、例えば１６００バイトである。
【００３１】
次に、キーワード抽出部２０の詳細な機能構成の例について図３を用いて説明する。
図３に示すように、キーワード抽出部２０は、形態素解析部２０１と、固有表現抽出部２０２と、第１フィルタリング部２０３と、を有してなる。
【００３２】
形態素解析部２０１は、データインポート部１０３から受け取ったテキストデータに対して形態素解析処理を実行する。この形態素解析処理によって、テキストデータは、形態素単位に分解される。分解された形態素は、後段の固有表現抽出部２０２に渡される。
【００３３】
図９に示したテキストデータに対して上記の形態素解析処理を実行した結果得られる検索語候補（キーワード候補）の例を図１０に示す。図１０に示すように、検索語候補のそれぞれには、スコアが算出されている。例えば、このスコアの上位からＸ番目までの形態素を検索語候補として後段の処理に用いることができる。
【００３４】
固有表現抽出部２０２は、形態素解析部２０１において分解された形態素の中から人名や地名に当たる固有表現を抽出する処理を実行する。抽出された固有表現に係る形態素は、第１フィルタリング部２０３に渡される。
【００３５】
固有表現抽出部２０２に入力される形態素データの例が図１１（ａ）に示すようなデータであると仮定する。この場合、固有表現抽出処理によって得られるデータは図１１（ｂ）に示すようになる。したがって、固有表現抽出部２０２は、固有表現に該当すると判定

した形態素に対して、後段の処理において検索語候補の優先順位を決めるフィルタリング処理が実行される。

【００３６】

第１フィルタリング部２０３は、形態素解析部２０１においてテキストデータから分解された形態素であって固有表現に該当する形態素と、予め作成しておいた標準語句分布テーブルとを用いたＴＦ－ＩＤＦベクトル演算を実行する。なお、標準語句分布テーブルは、例えば、インターネット上において展開されている百科事典をコーパスとして事前に作成し、文書情報検索システム１が備える記憶部に記憶させておけばよい。

【００３７】

第１フィルタリング部２０３において算出されたＴＦ－ＩＤＦスコアを付加した検索語候補の例を図１２に示す。図１２に示すように、形態素ごとにＴＦ－ＩＤＦスコアに係る情報が付加される。このＴＦ－ＩＤＦスコアを用いたフィルタリングによって、外部データセット検索部４０に渡す検索語候補の優先順位が決められる。

【００３８】

ここで、ＴＦ－ＩＤＦベクトル演算とは、文書情報中に用いられている単語に関する重み付け手法の一種である。ＴＦ（Ｔｅｒｍ　Ｆｒｅｑｕｅｎｃｙ）は単語の出現頻度、ＩＤＦ（Ｉｎｖｅｒｓｅ　Ｄｏｃｕｍｅｎｔ　Ｆｒｅｑｕｅｎｃｙ）は逆文書頻度、を示す指標である。ＴＦ－ＩＤＦベクトル演算は、これら２つの演算に基づくベクトル演算処理であって、文書情報同士や検索語と索引語の類似度を算出する処理である。

【００３９】

次に、ドキュメントマッピング部５０の詳細な機能構成の例について図４を用いて説明する。図４に示すように、ドキュメントマッピング部５０は、クラスタリング部５０１と、第２フィルタリング部５０２と、を有してなる。

【００４０】

クラスタリング部５０１は、外部データセット検索部４０が取得した検索結果を分析し、入力されたテキストデータとユーザー対応情報記憶部３０に既に記憶されているデータとの距離（類似度）を計算してマッピング処理を実行する。

【００４１】

第２フィルタリング部５０２は、上記のマッピング処理を実行するときに、検索結果の元になる外部データセット２の種類や検索結果のデータ種別に基づいて、検索結果の分析内容に重み付けを指定する処理を実行する。ここで、外部データセット２の種類には、検索結果が記憶されていたサーバーを特定する情報が含まれる。また、検索結果のデータ種別には、論文、特許公報、辞書データなどが含まれる。

【００４２】

次に、データ出力部６０の詳細な機能構成の例について図５を用いて説明する。図５に示すように、データ出力部６０は、出力フォーマット変換部６０１と、データエクスポート部６０２と、を有してなる。

【００４３】

出力フォーマット変換部６０１は、ドキュメントマッピング部５０から渡された検索結果に関する情報の種類を判定し、データ出力先５に応じたデータ形式に変換する処理を実行する。例えば、データ出力先５が液晶ディスプレイであれば、テキストデータを出力できるように処理を行う。

【００４４】

データエクスポート部６０２は、出力フォーマット変換部６０１において生成された出力データを、所定のデータ出力先５へと出力する処理を実行する。

【００４５】

上記構成を備える文書情報検索システム１は、ユーザーが能動的に検索のためのデータ入力を実行せずとも、例えば、複数人で会話している声をデータ入力元４とすることで、会話に合った検索を自動的に、繰り返して実行し、その結果を適宜出力できるようになる。ユーザーは、その表示を見ながら、さらに会話を展開したり収束させたりでき、会話を

止めることなく、会話を支援する検索結果を自動的に得ることができる。
【００４６】
●情報検索プログラムの処理フロー
次に、本発明に係る情報検索プログラムの実施形態である文書情報検索プログラム１００における処理の流れの例についてフローチャートを用いて説明する。以下の説明に用いるフローチャートはいずれも、各処理のステップをＳ１０、Ｓ２０・・・のように表記する。
【００４７】
図６に示すように、まず、入力検知処理が実行される（Ｓ１０）。入力検知処理（Ｓ１０）は、データ入力部１０の入力データソースセレクタ１０１がデータ入力元４からのデータ入力を検知するまで処理をループする（Ｓ１０のＮｏ）。データ入力元４からのデータ入力を検知したとき（Ｓ１０のＹｅｓ）、入力されたデータの種類（フォーマット）を判別し、フォーマット識別情報を入力データに付加して後段の処理に渡す。
【００４８】
続いて、フォーマット変換処理が実行される（Ｓ２０）。フォーマット変換処理（Ｓ２０）は、前段の処理において付与されたフォーマット識別情報に基づいて、入力データをテキスト形式のデータ（テキストデータ）へと変換する処理である。フォーマット変換処理（Ｓ２０）は、入力データのデータ形式が音声の場合は、音声認識処理を実行して入力音声からテキストデータを生成する。また、フォーマット変換処理（Ｓ２０）は、入力データが特定のアプリケーションプログラムに紐づくデータ形式であれば、いわゆる「プレーンテキスト」形式に変換して、テキストデータを生成する。
【００４９】
続いて、データインポート処理が実行される（Ｓ３０）。データインポート処理（Ｓ３０）は、フォーマット変換処理（Ｓ２０）において生成されたテキストデータを所定のデータサイズにして、後段の処理に渡す処理である。ここで、「所定のデータサイズ」とは、例えば、１６００バイトである。言い換えると、データインポート処理（Ｓ３０）では、前段の処理で生成されたテキストデータを先頭から１６００バイドで切り出して後段の処理に渡す。
【００５０】
続いて、キーワード抽出処理（Ｓ４０）が実行される。キーワード抽出処理（Ｓ４０）は、入力されたテキストデータに基づいて、検索語を生成する処理である。この検索語を用いて外部データセット２に対する検索処理が実行される。
【００５１】
図７は、キーワード抽出処理（Ｓ４０）における詳細な処理の流れの例を示すフローチャートである。キーワード抽出処理（Ｓ４０）では、まず、形態素解析処理（Ｓ４０１）が実行される。形態素解析処理（Ｓ４０１）によって前段の処理で切りだされたテキストデータが形態素に分解される。次に、固有表現抽出処理（Ｓ４０２）が実行される。固有表現抽出処理（Ｓ４０２）は、前段の処理で得られた形態素の中で、人名や地名などの固有表現に該当する形態素を抽出する処理である。固有表現抽出処理（Ｓ４０２）において抽出された形態素は後段の処理に渡される。
【００５２】
次に、ＴＦ－ＩＤＦベクトル算出処理（Ｓ４０３）が実行される。ＴＦ－ＩＤＦベクトル算出処理（Ｓ４０３）は、前段の処理で得られた形態素と、予め作成しておいた標準語句分布テーブルとを用いてＴＦ－ＩＤＦベクトルを算出する処理である。当該処理により算出されたＴＦ－ＩＤＦベクトルを数値化した「ＴＦ－ＩＤＦスコア」を各形態素に付加し、検索語候補となるデータが生成される。
【００５３】
続いて、ソート処理（Ｓ４０４）が実行される。ソート処理（Ｓ４０４）は、ＴＦ－ＩＤＦベクトルの各要素（語句）を、ユーザー対応情報記憶部３０に記憶されているユーザー対応情報と、上記のＴＦ－ＩＤＦスコアを用いて、フィルリングし、検索語の候補を並

べ替える処理である。

【００５４】

続いて、上位ワード抽出処理（Ｓ４０５）を実行する。上位ワード抽出処理（Ｓ４０５）は、ソート処理で並べ替えられた検索語の候補からＴＦ－ＩＤＦスコアが高い上位５位までの検索語を抽出する処理である。

【００５５】

図６に戻る。上位ワード抽出処理（Ｓ４０５）によって抽出された検索語を用いて外部データセット検索処理が実行される（Ｓ５０）。外部データセット検索処理（Ｓ５０）は、上記の検索語の優先度に従って外部データセット２を検索する処理である。外部データセット検索処理（Ｓ５０）において取得される検索結果は、外部データセット２に記憶されていて上記の検索語に関連する文献情報である。これら文献情報は、ドキュメントマッピング部５０に渡される。

【００５６】

外部データセット検索処理（Ｓ５０）に続いて、ドキュメントマッピング処理（Ｓ６０）が実行される。ドキュメントマッピング処理（Ｓ６０）は、ドキュメントマッピング部５０によって検索語の出現回数や特徴ベクトル計算、クラスタリング等の機械学習処理を実行する処理である。ドキュメントマッピング処理（Ｓ６０）によって、データ入力部１０から入力されたテキストデータを用いて検索された文献情報に係る抄録や特許文献の解説などの内容と類似度が高い文書を導き出すことができる。

【００５７】

図８は、ドキュメントマッピング処理（Ｓ６０）における詳細な処理の流れの例を示すフローチャートである。まず、キャッシュ処理（Ｓ６０１）が実行される。キャッシュ処理（Ｓ６０１）は、検索結果の上位の文献情報に係る「文献タイトル」、「文献の内容を表すテキスト（抄録、解説など）」、「ソースＵＲＬ」などの情報をキャッシュ情報としてユーザー対応情報記憶部３０に記憶する処理である。

【００５８】

続いて、上記のキャッシュ情報に対して形態素解析処理（Ｓ６０２）が実行される。この形態素解析処理（Ｓ６０２）によって、外部データセット検索処理（Ｓ５０）において取得された検索結果は形態素に分解される。

【００５９】

続いて、分割された形態素と、予め作成しておいた標準語句分布テーブルとを用いてＴＦ－ＩＤＦベクトル算出処理（Ｓ６０３）が実行される。ＴＦ－ＩＤＦベクトル算出処理（Ｓ６０３）は、基本コーパスとして、インターネット上で提供されているデータベースを用いてＴＦ－ＩＤＦベクトルを算出する。ＴＦ－ＩＤＦベクトル算出処理（Ｓ６０３）によって、入力されたテキストデータの形態素とキャッシュ情報との類似度が距離情報として算出される。

【００６０】

続いて、ソート処理（Ｓ６０４）が実行される。ソート処理（Ｓ６０４）は、ＴＦ－ＩＤＦスコアの高い順に上記の形態素を並べ替える処理である。

【００６１】

続いて、ユーザー対応記憶領域更新処理（Ｓ６０５）が実行される。ユーザー対応記憶領域更新処理（Ｓ６０４）は、ユーザー対応情報記憶部３０によって記憶されている、ユーザー対応情報を更新する処理である。ソート処理（Ｓ６０４）において並べ変えられた形態素の上位の要素（語句）を選別してユーザー対応情報を更新する。このユーザー対応情報は、キーワード抽出部２０において検索語を生成するときのフィルタとして機能する。したがって、ユーザー対応情報を更新することで、次回の検索処理における適合度を向上させることができる。

【００６２】

図６に戻る。次に、表示処理（Ｓ７０）が実行される。表示処理（Ｓ７０）は、データ出力部６０が外部データセット検索処理（Ｓ５０）において得られた検索結果から、ドキ

ュメントマッピング処理（Ｓ６０）において更新されたユーザー対応情報に基づいて生成される出力データを表示する処理である。

【００６３】

図１３は、外部データセット検索処理（Ｓ５０）において得られた検索結果（表示すべき文書）と、検索結果から生成されたユーザー対応情報を更新する情報との例を示す図である。

【００６４】

図１３に示すように、文書情報検索プログラム１００の処理によって外部データセット２から得られた検索結果に係るデータのうち、データ出力部６０における選択処理の対象となるデータと、ユーザー対応情報記憶部３０の更新に用いられるデータは同一ではない。ユーザー対応情報記憶部３０の更新に用いられるデータは、出力対象となるデータの一部に相当する。

【００６５】

図１４は、表示処理（Ｓ７０）において、例えばデータ出力先５がディスプレイ出会って、出力データがテキストデータの場合の例である。この場合は、図１４に示すように出力されるテキストデータは、文字列として表示される。

【００６６】

以上説明した文書情報検索システム１を用いることで、例えば、「暑気払い」という語を用いて外部データセット２への検索が行われたとすると、「ビール」と「氷」に関する話題が表示される。この表示された文書の一部（タイトルなど）は、ユーザー対応記憶域に保存される。表示された情報をユーザーの話題が「ビール」に移ると、次の検索のタイミングにおける検索語は「ビール」と「暑気払い」によって構成されるようになる。

【００６７】

したがって、文書情報検索システム１及び文書情報検索プログラム１００を用いることで、ユーザー対応情報記憶部３０に記憶されている内容とのマッチングをとることができ、よりユーザーの意図に関連した話題による検索処理を自動的に実行できるようになる。

【００６８】

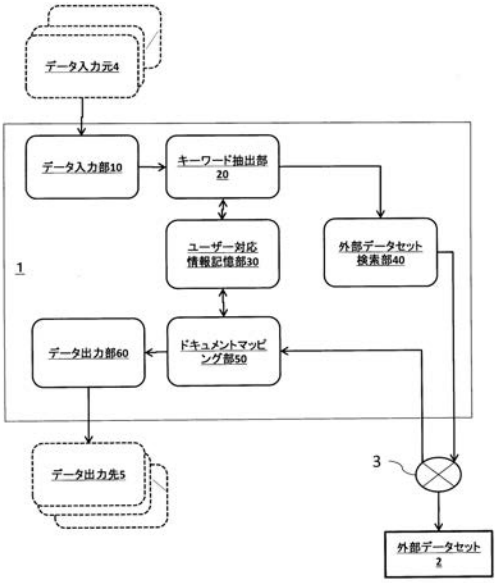
また、文書情報検索システム１及び文書情報検索プログラム１００を用いることで、ユーザーは、ブレインストーミングなどの場での対話内容などを自動的に検索語として検索させることができる。これによって、その場の話題に関連したドキュメント（文書）を自動的に提示することができ、ユーザーからの直接的な反応や入力などを要求することなく、いわば、ユーザーの対話を邪魔することなく適時に適合する文書情報を提示することができる。したがって、毎回、新規の検索語として検索する場合とは異なり、前回の検索と関連を持たせながら話題の方向性を容易に絞り込むことができるようになる。

【符号の説明】

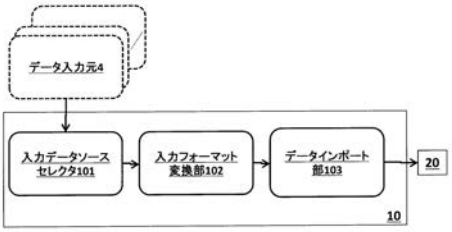
【００６９】

１　　文書情報検索システム
２　　外部データセット
３　　ネットワーク
１０　　データ入力部
２０　　キーワード抽出部
３０　　ユーザー対応情報記憶部
４０　　外部データセット検索部
５０　　ドキュメントマッピング部
６０　　データ出力部
１００　文書情報検索プログラム

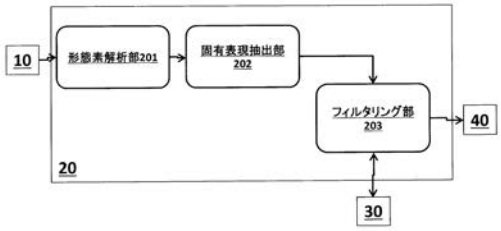
【図１】



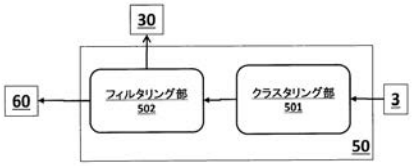
【図２】



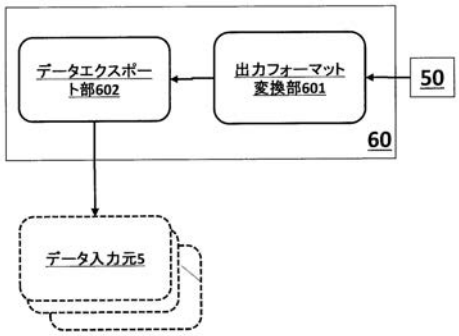
【図３】



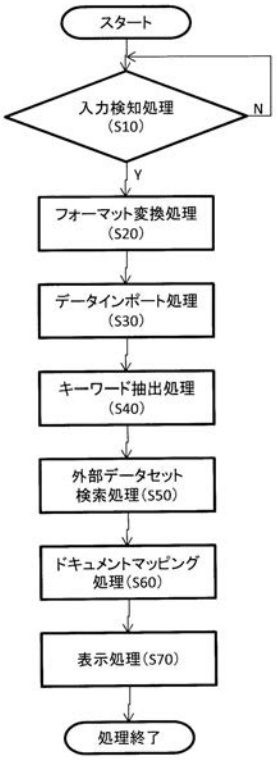
【図４】



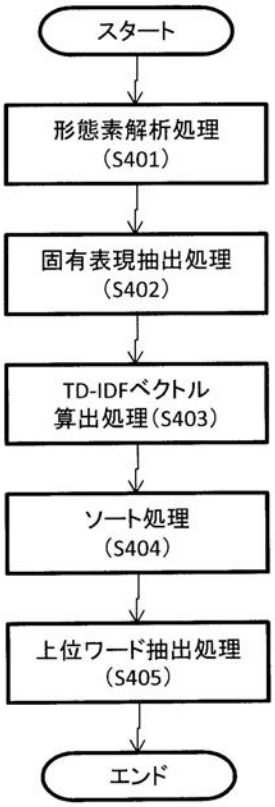
【図５】



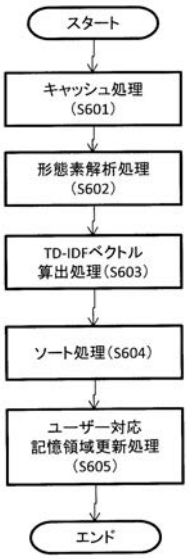
【図６】



【図７】



【図８】



【図９】

『ディープラーニングが注目されるようになったのは、2012年、コンピューターによる物体認識の精度を競う国際コンテスト「ImageNet Large Scale Visual Recognition Challenge（ILSVRC）2012」で、トロント大学のSuperVisionチームが写真データに写っている物体を特定する人工知能をディープラーニングで構成し、「タスク1：分類」「タスク2:局所化と分類」部門で優勝したあたりからでしょうか。他のチームがエラー率26%前後のところ、エラー率17%弱とダントツの認識率をマークしたのです。

同じく2012年、米Googleがディープラーニングで構成された人工知能にYouTubeの画像を見せ続け学習させた結果、猫の画像を猫と認識できるようになった、と発表したのも大きなきっかけになっています。1万6千CPUコアのPCクラスタに、300万以上のノード、1億以上のエッジからなる多層ニューラルネットワークを構築し、YouTubeからランダムに抽出した1000万枚の画像をまる3日かけて学習させたところ、人が教えなくても、コンピュータが猫の画像を猫、他のものを他のものと実に2万個もの物体を写真の中から認識できるようになった、というのです。これまではこのような「実世界の画像に対して、その中に含まれる物体を名称で認識する」機械を作るためには「特徴量をベクトル表現で抽出」「学習器によるパターン認識」といったいくつかの人工知能技術を組み合わせて実現していました。しかも、たとえば、画像からどのような特徴量を取り出すかといった部分は人手によるチューニングが必要となっていたのです。ところが「ディープラーニング」ではひとつのアルゴリズムを使い、しかもこれまでの方法よりダントツに高い認識率で実現してしまいました。このことからディープラーニングは人工知能にとって、非常に大きなブレイクスルーとなりました。』

【図１０】

```
keyword_list] => Array
    (
        [0] => Array
            (
                [keyword] => ディープラーニング
            )
        [1] => Array
            (
                [keyword] => 物体認識
            )
        [2] => Array
            (
                [keyword] => 認識率
            )
        [3] => Array
            (
                [keyword] => 人工知能技術
            )
        [4] => Array
            (
                [keyword] => 人工知能
            )
    )
```

【図１１】

(a)

入力テキストデータの例

[{“request_id”:”record002”, “sentence”:”鈴木さんがきょうの9時30分に横浜に行きます。”}

(b)

出力データの例

{“request_id”:”record002”, “ne_list”:[[“鈴木”, “PSN”], [“きょう”, “DAT”], [“9時30分”, “TIM”], [“横浜”, “LOC”]]}

【図１２】

```
keyword_list] => Array
    (
      [0] => Array
        (
          [keyword] => ディープラーニング
          [score] => 0.15105941616182
        )
      [1] => Array
        (
          [keyword] => 物体認識
          [score] => 0.13064159628801
        )
      [2] => Array
        (
          [keyword] => 認識率
          [score] => 0.10764873252922
        )
      [3] => Array
        (
          [keyword] => 人工知能技術
          [score] => 0.10556443713867
        )
      [4] => Array
        (
          [keyword] => 人工知能
          [score] => 0.10556443713867
        )
    )
```

【図１３】

（保持するデータ）表示すべき文書
result_list[n] ['doc_id'] 文書ID
result_list[n] ['title'] 文書タイトル
result_list[n] ['summary'] 検索結果画面におけるサマリ表示テキスト
result_list[n] ['score'] スコア
result_list[n] ['src_type'] 文書が属する情報源
result_list[n] ['doc_type'] 文書の種類
result_list[n] ['src_url'] 情報源URL
result_list[n] ['created_at'] 登録日

（保持するデータ）ユーザ対応記憶域
result_list[n] ['title'] 文書タイトル
result_list[n] ['summary'] 検索結果画面におけるサマリ表示テキスト
result_list[n] ['src_type'] 文書が属する情報源
result_list[n] ['doc_type'] 文書の種類

【図１４】

doc_id : 8501 | created_at : 2015-06-19 21:44:39
src_type : cinii_article | doc_type : doc
title : 人工知能の最前線～人間の脳を真似るコンピューター | score : 0.0019729400885213
summary : 人工知能の実現には様々な技術が利用されているが、特に最近注目を集めているのはディープラーニングと呼ばれる人間の脳のメカニズムを参考にした技術だ。ディープラーニングはコンピューターの類認識性能を人間と同じレベルにまで向上させるなどの成果をあげている。なぜ、ディープラーニングは高い性能を発揮できるのだろうか。それは、コンピューターが概念や意味を理解する力を獲得したからだと言われている。人間は長い生活の中で「生物は生きている」「人間は2本足で歩く」というような概念を獲得していく。これと同じようにコンピューターが多くのデータから「画像に写る物体の見分け方」や「日本語と英語の違い」を学べるようになってきた。これまでは人間がコンピューターに物体の見分け方を教えていたが、それが完全に自動化され、最近では「日本語の良し悪し」のような感覚的なものまでコンピューターが扱えるようになってきている。…
リンク

doc_id : 1013850 | created_at : 2015-07-02 20:06:52
src_type : jg_patent | doc_type : 特許
title : 複合人工知能装置 | score : 0.0019310274265539
summary : 【課題】複数の人工知能装置相互間で認識情報を共有することにより、多方向からの情報を総合的に判断することで認識精度を飛躍的に向上させる。【解決手段】任意の物体に取り付けられる複数の人工知能装置は、それぞれ、取り付けられた物体の属性を記憶する自己物体の属性記憶機能101と、取り付けられた物体の空間位置姿勢を認識する自己物体の空間位置姿勢認識機能102と、取り付けられた物体の周囲の対象物を認識する自己物体の周囲対象物認識機能103と、各認識機能により取り付けられた物体の状況を判断する自己物体の状況判断機能104と、各認識機能の処理過程を記録する機能105と、他の物体に取り付けられ