(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2014 年 7 月 3 日(03.07.2014)

WIPO | PCT

(10) 国際公開番号
WO 2014/102992 A1

(51) 国際特許分類:
G06F 17/30 (2006.01)

(21) 国際出願番号: PCT/JP2012/084007

(22) 国際出願日: 2012 年 12 月 28 日(28.12.2012)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(71) 出願人: 株式会社日立製作所 (HITACHI, LTD.) [JP/JP]; 〒1008280 東京都千代田区丸の内一丁目６番６号 Tokyo (JP).

(72) 発明者: 藤田 雄介 (FUJITA, Yusuke); 〒1858601 東京都国分寺市東恋ヶ窪一丁目２８０番地 株式会社日立製作所 中央研究所内 Tokyo (JP). 額賀 信尾 (NUKAGA, Nobuo); 〒1858601 東京都国分寺市東恋ヶ窪一丁目２８０番地 株式会社日立製作所 中央研究所内 Tokyo (JP). 児玉 昇司 (KODAMA, Shoji); 〒2440817 神奈川県横浜市戸塚区吉田町２９２番地 株式会社日立製作所 横浜研究所内 Kanagawa (JP).

(74) 代理人: 藤井 正弘 (FUJII, Masahiro); 〒1050001 東京都港区虎ノ門一丁目１６番４号アーバン虎ノ門ビル 後藤特許事務所 Tokyo (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT,

[続葉有]

(54) **Title**: DATA PROCESSING SYSTEM AND DATA PROCESSING METHOD

(54) 発明の名称 : データ加工システムおよびデータ加工方法

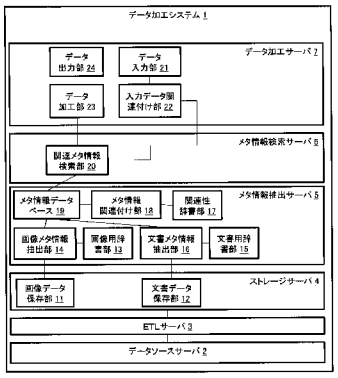


FIG. 1:
1 Data processing system
2 Data source server
3 ETL server
4 Storage server
5 Meta-information extraction server
6 Meta-information search server
7 Data processing server
11 Image data storage unit
12 Document data storage unit
13 Image-use dictionary unit
14 Image meta-information extraction unit
15 Document-use dictionary unit
16 Document meta-information extraction unit
17 Association dictionary unit
18 Meta-information association unit
19 Meta-information database
20 Related meta-information search unit
21 Data input unit
22 Input data association unit
23 Data processing unit
24 Data output unit

(57) **Abstract**: A data processing system which has one or more processors, and one or more storage devices connected to the one or more processors, wherein the data processing system: retains meta-information extraction-use dictionary information which defines conditions to extract meta-information from a plurality of types of data, and association dictionary information which defines conditions associating the meta-information extracted from the plurality of types of data; extracts meta-information from the plurality of types of data on the basis of the meta-information extraction-use dictionary information; extracts meta-information from data that have been input; associates meta-information extracted from the data that have been input with the meta-information extracted from the plurality of types of data on the basis of the association dictionary information; and, on the basis of the results of the association, outputs information indicating association of any combination among the plurality of types of data, the data that have been input, and the meta-information extracted from the data.

(57) 要約: 一つ以上のプロセッサと、前記一つ以上のプロセッサに接続される一つ以上の記憶装置と、を有するデータ加工システムであって、複数の種類のデータからメタ情報を抽出する条件を定義するメタ情報抽出用辞書情報と、前記複数の種類のデータから抽出されたメタ情報を関連付ける条件を定義する関連性辞書情報と、を保持し、前記複数の種類のデータから、前記メタ情報抽出用辞書情報に基づいて前記メタ情報を抽出し、入力されたデータからメタ情報を抽出し、前記関連性辞書情報に基づいて、前記入力されたデータから抽出されたメタ情報と前記複数の種類のデータから抽出されたメタ情報とを関連付け、前記関連付けの結果に基づいて、前記複数の種類のデータと、前記入力されたデータと、それらのデータから抽出されたメタ情報と、のいずれかの組み合わせの関連を示す情報を出力する。

WO 2014/102992 A1

NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:

— 国際調査報告（条約第 21 条(3)）

明　細　書

発明の名称 ： データ加工システムおよびデータ加工方法

技術分野

[0001] 本発明は、複数のモダリティからのデータ抽出および加工技術に関する。

背景技術

[0002] 本技術分野の背景技術として、米国特許出願公開第２０１０／０１８５９３４号明細書（特許文献１）がある。この明細書には、「収集された電子文書から情報を検索及び表示するための、計算機記憶媒体に格納された計算機プログラムを含む方法、システム及び装置。一つの態様は、既存の構造化プレゼンテーションを記述する記述データを受信する動作と、既存の構造化プレゼンテーションの特性と収集された非構造化電子文書に含まれる電子文書の内容とを比較することで、既存の構造化プレゼンテーションに関する新たな属性を示す電子文書を特定する動作と、新たな属性の識別子を既存の構造化プレゼンテーションに追加することで、拡張された構造化プレゼンテーションを形成する動作と、拡張された構造化プレゼンテーションを提示する指示を出力する動作と、を含む、計算機に実装された方法によって実施され得る」と記載されている（要約参照）。

先行技術文献

特許文献

[0003] 特許文献１：米国特許出願公開第２０１０／０１８５９３４号明細書

発明の概要

発明が解決しようとする課題

[0004] 従来、データを用いて事象の発生を予測するためのデータマイニングは、表または関係データベースといった形に整理された構造データに基づいて行われる。しかし、構造データとして取り込むことのできるデータは、特定用途向けに、コンピュータシステム内で予め属性名と属性値の付与が行われたものだけであって、画像、音声または非定型の文書といった非構造データを

、直接データマイニングの対象とすることはできない。

[0005] 一方、テキスト文書を対象とした全文検索エンジンは、非構造データ中の単語を高速に検索することができ、単語リストを用いての単純な条件付き検索は可能となった。また、複数の構造データをルールに基づいて結び付けることは可能であり、インターネット上の大量のデータから構造データを検索して、大きな構造データを生成することが可能となった。さらに、非構造データに対し、テキストの構文解析を行うことで、部分的な構造を取得することは可能となった。例えば、上記の特許文献１は、これらの技術を組み合わせ、表の属性名および属性値を追加する方法を開示している。

[0006] しかしながら、非構造データを構造データに結び付けたデータマイニング、および単語リストだけでない、非構造データに対する属性条件を使った検索は、実現されていない。また、非構造データに対して、部分的に構造を与えることができても、構造データ中のどの行または列に結び付けるかを判断する手法は従来なかった。

[0007] 本発明は、このような点に鑑みてなされたものであり、その目的は、複数のモダリティから抽出したデータを用いて、データマイニング・条件付き検索を行うための、データ抽出・加工方法を提供することにある。

課題を解決するための手段

[0008] 上記の課題を解決するために、本発明は、一つ以上のプロセッサと、前記一つ以上のプロセッサに接続される一つ以上の記憶装置と、を有するデータ加工システムであって、複数の種類のデータからメタ情報を抽出する条件を定義するメタ情報抽出用辞書情報と、前記複数の種類のデータから抽出されたメタ情報を関連付ける条件を定義する関連性辞書情報と、を保持し、前記複数の種類のデータから、前記メタ情報抽出用辞書情報に基づいて前記メタ情報を抽出し、入力されたデータからメタ情報を抽出し、前記関連性辞書情報に基づいて、前記入力されたデータから抽出されたメタ情報と前記複数の種類のデータから抽出されたメタ情報とを関連付け、前記関連付けの結果に基づいて、前記複数の種類のデータと、前記入力されたデータと、それらの

データから抽出されたメタ情報と、のいずれかの組み合わせの関連を示す情報を出力する。

発明の効果

[0009] 本発明の一実施形態によれば、入力データに関連するメタ情報を容易に検索して、加工することができる。

[0010] 上記以外の課題、構成及び効果は、以下の実施形態の説明により明らかにされる。

図面の簡単な説明

[0011] [図1]本発明の実施例１のデータ加工システムの全体構成を説明するブロック図である。

[図2]本発明の実施例１のデータ加工システムのハードウェア構成を説明するブロック図である。

[図3]本発明の実施例１におけるメタ情報データベース構築処理の動作を示すフローチャートである。

[図4]本発明の実施例１において画像メタ情報抽出部が画像データに対するメタ情報を抽出する処理の詳細を説明する図である。

[図5]本発明の実施例１において文書メタ情報抽出部が文書データに対するメタ情報を抽出する処理の詳細を説明する図である。

[図6]本発明の実施例１においてメタ情報関連付け部がメタ情報に関連付けを付与する処理の詳細を説明する図である。

[図7]本発明の実施例１におけるデータ加工処理の動作を示すフローチャートである。

[図8]本発明の実施例１において入力データ関連付け部が受け付けた表から属性の関係を抽出する処理の詳細を説明する図である。

[図9]本発明の実施例１において入力データ関連付け部が表構造属性情報に関連付けを付与する処理の詳細を説明する図である。

[図10]本発明の実施例１においてデータ加工部が入力データに属性を追加する処理を説明する図である。

[図11]本発明の実施例２のデータ加工システムの全体構成を説明するブロック図である。

[図12]本発明の実施例２において音声メタ情報抽出部が音声データに対するメタ情報を抽出する処理の詳細を説明する図である。

[図13]本発明の実施例２においてテキストメタ情報抽出部がテキストデータに対するメタ情報を抽出する処理の詳細を説明する図である。

[図14]本発明の実施例２におけるデータ加工処理の動作を示すフローチャートである。

[図15]本発明の実施例２において入力データ関連付け部が抽出されたキーワードに関連付けを付与する処理の詳細を説明する図である。

[図16]本発明の実施例２においてデータ加工部が音声メタ情報とテキストメタ情報を関連付ける処理を説明する図である。

発明を実施するための形態

[0012] 以下、実施例を、図面を用いて説明する。

実施例 1

[0013] 本実施例では、予め構築しておいた画像データおよび文書データに関するメタ情報データベースに基づいて、入力された表を拡張する、データ加工システムの例を説明する。本システムは、例えば、建造物または機械等を製造する際に発行される、設計図面および設計文書を管理するために用いることができる。設計に関する表を入力すると、設計図面および設計文書から抽出されたメタ情報を用いて、表が自動的に拡張される。こうして設計に関わる大規模な表が得られるため、本実施例は、設計の不具合分析または不良予測など、データマイニングに応用することが可能となる。

[0014] 図１は、本発明の実施例１のデータ加工システムの全体構成を説明するブロック図である。

[0015] データ加工システム１は、データソースサーバ２、ＥＴＬ（Ｅｘｔｒａｃｔ　Ｔｒａｎｓｆｏｒｍ　Ｌｏａｄ）サーバ３、ストレージサーバ４、メタ情報抽出サーバ５、メタ情報検索サーバ６、およびデータ加工サーバ７によ

って構成される。

[0016] データソースサーバ２は、画像および文書を管理する装置である。データソースサーバ２は、図面をＩＤ（識別情報）と結び付けて管理するリレーショナルデータベース（図示省略）、およびテキスト文書を保存するファイルサーバ（図示省略）を備える。

[0017] ＥＴＬサーバ３は、データソースサーバ２に保存されている画像データおよび文書データを、ストレージサーバ４に保存する機能を備える。ここで、画像および文書のフォーマットを統一するなどの変換が行われる。

[0018] ストレージサーバ４は、画像データ保存部１１および文書データ保存部１２を備え、複数のデータソースから収集された画像データおよび文書データを統一された形式で保存する。

[0019] メタ情報抽出サーバ５は、画像用辞書部１３、画像メタ情報抽出部１４、文書用辞書部１５、文書メタ情報抽出部１６、関連性辞書部１７、メタ情報関連付け部１８、およびメタ情報データベース１９を備え、ストレージサーバ４にあるデータから抽出したメタ情報の管理を行う。

[0020] メタ情報検索サーバ６は、関連メタ情報検索部２０を備え、検索要求を受けつけてメタ情報データベース１９を検索した結果を返す。

[0021] データ加工サーバ７は、データ入力部２１、入力データ関連付け部２２、データ加工部２３、およびデータ出力部２４を備え、入力されたデータをメタ情報に基づいて加工し出力する。

[0022] 上記の各部の詳細については後述する。

[0023] 図２は、本発明の実施例１のデータ加工システム１のハードウェア構成を説明するブロック図である。

[0024] データソースサーバ２は、相互に接続された通信部２２１、ＣＰＵ（Ｃｅｎｔｒａｌ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）２２２、メモリ２２３およびディスク２２４を有する計算機である。

[0025] 通信部２２１は、中継装置２８０に接続され、中継装置２８０を介して他のサーバと通信するためのインターフェースである。ＣＰＵ２２２は、メモ

リ２２３に格納されたプログラムを実行することによって所定の機能を実現するプロセッサである。メモリ２２３及びディスク２２４は、ＣＰＵ２２２によって実行されるプログラムおよびＣＰＵ２２２によって参照されるデータ等を格納する記憶装置である。これらはどのような種類の記憶装置であってもよいが、典型的な例を示すと、メモリ２２３がＤＲＡＭ（Ｄｙｎａｍｉｃ　Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ）のような比較的高速の半導体メモリであり、ディスク２２４がハードディスク装置のような比較的大容量の記憶装置である。

[0026] ＥＴＬサーバ３は、相互に接続された通信部２３１、ＣＰＵ２３２、メモリ２３３およびディスク２３４を有する計算機である。これらの各部の説明は、データソースサーバ２の通信部２２１、ＣＰＵ２２２、メモリ２２３およびディスク２２４の説明と同様であるため、省略する。

[0027] ストレージサーバ４は、相互に接続された通信部２４１、ＣＰＵ２４２、メモリ２４３およびディスク２４４を有する計算機である。これらの各部の説明は、データソースサーバ２の通信部２２１、ＣＰＵ２２２、メモリ２２３およびディスク２２４の説明と同様であるため、省略する。ただし、ディスク２４４は、画像データを保存する画像データ保存部１１および文書データを保存する文書データ保存部１２を含む。画像データ保存部１１および文書データ保存部１２に保存されたデータの一部又は全部が必要に応じてメモリ２４３にコピーされてもよい。

[0028] メタ情報抽出サーバ５は、相互に接続された通信部２５１、ＣＰＵ２５２、メモリ２５３およびディスク２５４を有する計算機である。これらの各部の説明は、データソースサーバ２の通信部２２１、ＣＰＵ２２２、メモリ２２３およびディスク２２４の説明と同様であるため、省略する。ただし、メモリ２５３は、画像メタ情報抽出部１４、文書メタ情報抽出部１６およびメタ情報関連付け部１８を含む。これらは、ＣＰＵ２５２によって実行されるプログラムである。

[0029] 以下、画像メタ情報抽出部１４、文書メタ情報抽出部１６またはメタ情報

関連付け部１８が実行する処理について説明する場合があるが、そのような処理は、実際には上記のプログラムに従ってＣＰＵ２５２が必要に応じてメモリ２５３、ディスク２５４及び通信部２５１等を制御することによって実行する処理である。後述するメモリ２６３及びメモリ２７３に格納されたプログラムが実行する処理も、実際には、上記と同様にそれぞれの計算機のＣＰＵによって実行される。

[0030] なお、画像メタ情報抽出部１４、文書メタ情報抽出部１６およびメタ情報関連付け部１８は、ディスク２５４に格納され、必要に応じてメモリ２５３にコピーされてもよい。後述するメモリ２６３及びメモリ２７３に格納されたプログラムについても同様である。

[0031] メタ情報検索サーバ６は、相互に接続された通信部２６１、ＣＰＵ２６２、メモリ２６３およびディスク２６４を有する計算機である。これらの各部の説明は、データソースサーバ２の通信部２２１、ＣＰＵ２２２、メモリ２２３およびディスク２２４の説明と同様であるため、省略する。ただし、メモリ２６３は、関連メタ情報検索部２０を含む。これは、ＣＰＵ２６２によって実行されるプログラムである。

[0032] データ加工サーバ７は、相互に接続された通信部２７１、ＣＰＵ２７２、メモリ２７３およびディスク２７４を有する計算機である。これらの各部の説明は、データソースサーバ２の通信部２２１、ＣＰＵ２２２、メモリ２２３およびディスク２２４の説明と同様であるため、省略する。ただし、メモリ２７３は、データ入力部２１、入力データ関連付け部２２、データ加工部２３及びデータ出力部２４を含む。これらは、ＣＰＵ２７２によって実行されるプログラムである。

[0033] データ加工サーバ７は、さらに、ＣＰＵ２７２に接続され、データ入力部２１およびデータ出力部２４によって制御される入力部２７５および出力部２７６を有する。入力部２７５は、例えばキーボードおよびポインティングデバイスのような入力デバイスであり、出力部２７６は、例えば画像表示装置のような出力デバイスである。

[0034] 図２では省略されているが、データ加工システム１を構成するデータ加工サーバ７以外の各サーバも、データ加工サーバ７と同様の入力部および出力部を有してもよい。

[0035] 中継装置２８０は、各サーバの通信部に接続され、サーバ間の通信を中継する装置である。

[0036] 図２には、各サーバが一つのＣＰＵおよび一つ以上の記憶装置を備える独立した計算機によって実現されるハードウェア構成の例を示したが、このようなハードウェア構成は一例であり、実際には一つ以上のＣＰＵおよび一つ以上の記憶装置を有する種々の形態の計算機システムによって本実施例を実現することができる。例えば、上記の全てのプログラム及びデータを格納する記憶装置と、少なくとも一つのＣＰＵとを含む一つの計算機によってデータ加工システム１が実現されてもよい。あるいは、例えばデータソースサーバ２が一つの計算機によって実現され、ＥＴＬサーバ３及びストレージサーバ４が別の一つの計算機によって実現され、メタ情報抽出サーバ５、メタ情報検索サーバ６及びデータ加工サーバ７がさらに別の一つの計算機によって実現されてもよい。このような場合、各サーバは、仮想化技術を利用して生成された仮想サーバであってもよい。

[0037] 次に、上記のように構成される、本実施例に係るデータ加工システム１の動作を説明する。本システムの動作は、メタ情報データベース構築処理とデータ加工処理とに分けられる。

[0038] まず、メタ情報データベースの構築処理に関する動作を説明する。

[0039] 図３は、本発明の実施例１におけるメタ情報データベース構築処理の動作を示すフローチャートである。

[0040] まず、ＥＴＬサーバ３は、データソースサーバ２から、画像データおよび文書データを取得する（ステップＳ３０１）。続いて、ＥＴＬサーバ３は、画像データおよび文書データに対し、必要な変換を施す（ステップＳ３０２）。例えば、後述する画像メタ情報抽出部１４が、特定の形式の画像データしか受け付けない場合には、画像データの形式を当該特定の形式に変換する

処理が行われる。続いて、ＥＴＬサーバ３は、変換された画像データおよび文書データを、それぞれ、ストレージサーバ４の画像データ保存部１１及び文書データ保存部１２へ保存する（ステップＳ３０３）。

[0041] 次に、メタ情報抽出サーバ５は、ストレージサーバ４から画像データおよび文書データを取得する（ステップＳ３０４）。

[0042] 次に、メタ情報抽出サーバ５の画像メタ情報抽出部１４は、画像用辞書部１３に基づいて、画像データに対するメタ情報を抽出する（ステップＳ３０５）。

[0043] 図４は、本発明の実施例１において画像メタ情報抽出部１４が画像データに対するメタ情報を抽出する処理（図３のステップＳ３０５）の詳細を説明する図である。

[0044] 画像用辞書部１３は、画像の形状・色情報等に基づいて認識・分類を行うためのモデル４０１を保持している。また、各モデル４０１には、ラベル４０２が対応づけられている。例えば筒（ｔｕｂｅ）状の図形の画像に相当するモデル４０１にはラベル４０２として“ｓｈａｐｅ：ｔｕｂｅ”が対応づけられる。画像メタ情報抽出部１４は、画像用辞書部１３に基づく画像認識技術によって、画像データ保存部１１から取得された画像データ４０３に対するラベル４０２を付与する。具体的には、例えば、画像メタ情報抽出部１４は、画像データ４０３と各モデル４０１との類似度を公知の画像認識技術によって比較し、最も類似度の高いモデル４０１に対応するラベル４０２を画像データ４０３に付与してもよい。図４の例では、画像メタ情報抽出部１４は、画像データ４０３に“ｓｈａｐｅ：ｔｕｂｅ”のラベルを付与し、その結果を、画像メタ情報４０４として出力する。

[0045] 画像メタ情報４０４は、「Ｃ０１．ｊｐｇのｓｈａｐｅがｔｕｂｅである」ことを示す表現であり、例えば、ＲＤＦ（Ｒｅｓｏｕｒｃｅ　Ｄｅｓｃｒｉｐｔｉｏｎ　Ｆｒａｍｅｗｏｒｋ）を用いて「ＡのＢはＣである」といった三項関係の表現を記述することができる。

[0046] さらに、ステップＳ３０５において、文書メタ情報抽出部１６は、文書用

辞書部１５に基づいて、文書データに対するメタ情報を抽出する。

[0047] 図５は、本発明の実施例１において文書メタ情報抽出部１６が文書データに対するメタ情報を抽出する処理（ステップＳ３０５）の詳細を説明する図である。

[0048] 文書用辞書部１５は、文書中の単語に対し、属性名にふさわしい単語を集めた属性名リスト５０１および属性値にふさわしい単語を集めた属性値リスト５０２を保持している。文書メタ情報抽出部１６は、文書用辞書部１５と文書のレイアウトを解析した結果とに基づいて、文書データ保存部１２から取得された文書データ５０３の構造を解析することによって、文書メタ情報５０４を生成する。

[0049] 図５の例では、文書データ５０３は、「通知：Ｃ０１」および「場所：東京」等の文字列を含む。一方、属性名リスト５０１は「参照」、「場所」および「作成者」等の単語を含み、属性値リスト５０２は「Ｃ０１」、「Ｃ０２」、「東京」および「Ａｌｉｃｅ」等の単語を含む。

[0050] この例において、文書メタ情報抽出部１６は、文書データ５０３から、属性名リスト５０１に基づいて「参照」および「場所」といった属性名を抽出し、属性値リスト５０２に基づいて「Ｃ０１」および「東京」といった属性値を抽出し、それらの単語がコロンを挟んで並べられているといったレイアウト情報に基づいて例えば「参照」と「Ｃ０１」、「場所」と「東京」を対応付けることによって、表の構造化を行った結果として文書メタ情報５０４を生成する。文書メタ情報５０４は、画像メタ情報４０４と同様に、ＲＤＦを用いて記述することができる。

[0051] 次に、画像メタ情報抽出部１４および文書メタ情報抽出部１６は、抽出したメタ情報を、メタ情報データベース１９に保存する（ステップＳ３０６）。ここで、ＲＤＦを管理するデータベースを用いると、画像メタ情報と文書メタ情報を同一のデータベースの中で管理することができる。

[0052] 次に、メタ情報関連付け部１８は、関連性辞書部１７を用いて、メタ情報データベース１９の中に保存されているメタ情報に関連付けを付与する（ス

テップＳ３０７）。

[0053] 図６は、本発明の実施例１においてメタ情報関連付け部１８がメタ情報に関連付けを付与する処理（ステップＳ３０７）の詳細を説明する図である。

[0054] 関連性辞書部１７は、同義語辞書６０１を保持しており、同義語関係として、例えば「図面」と「参照」、「タイプ」と「ｓｈａｐｅ」、「Ｃ０１」と「Ｃ０１．ｊｐｇ」などが、システムの設計者によって予め構築される。同義語辞書６０１は、別途用意された翻訳辞書に基づいて、異なる言語の単語間の同義関係を保持するものであってもよい。

[0055] なお、同義語辞書は、あるメタ情報と置換可能な情報を定義する情報である限り、どのようなものであってもよい。具体的には、同義語辞書は、あるモーダルのメタ情報に含まれる情報と、それと同義の、他のモーダルのメタ情報に含まれる情報とが置換可能であることを示す情報であり、例えば、上記のような同義語辞書または翻訳辞書のほか、話し言葉と書き言葉との同義関係を定義する辞書（実施例２参照）であってもよい。

[0056] メタ情報関連付け部１８は、この同義語辞書６０１に存在する単語をメタ情報データベース１９から検索し、その単語を同義語関係にある単語に置き換えた変更メタ情報６０２に変換し、メタ情報データベース１９を更新する。図６の例では、同義語辞書６０１内の単語「ｓｈａｐｅ」に対して、「Ｃ０１．ｊｐｇのｓｈａｐｅはｔｕｂｅである」という画像メタ情報が検索される。この場合、メタ情報関連付け部１８は、同義語辞書６０１に基づいて「ｓｈａｐｅ」をその同義語である「タイプ」に置き換えることによって、検索された画像メタ情報を「Ｃ０１．ｊｐｇのタイプはｔｕｂｅである」というメタ情報に変換する。

[0057] 同様にして、「＿：ｒ１の参照はＣ０１である」が「＿：ｒ１の図面はＣ０１．ｊｐｇである」に、「＿：ｒ１の場所は東京である」が「＿：ｒ１の場所はＪＰ１である」に変換される。

[0058] ここでは、検索されたメタ情報を新たなメタ情報に更新する例を説明したが、これは、同義語関係に従って、検索されたメタ情報に含まれる属性名ま

たは属性値の単語と置換可能な単語を記述する方法の一例である。ステップＳ３０７では、例えば、同義語関係に従って、メタ情報データベース中の単語「ｓｈａｐｅ」と単語「タイプ」が置換可能であることと、単語「Ｃ０１」と単語「Ｃ０１．ｊｐｇ」が置換可能であることを、記述することができればよい。具体的には、メタ情報関連付け部１８は、例えば、検索されたメタ情報「Ｃ０１．ｊｐｇのｓｈａｐｅはｔｕｂｅである」を書き換える代わりに、同義語辞書６０１に基づいて、単語「ｓｈａｐｅ」と単語「タイプ」が置換可能であることを示す情報、及び、単語「Ｃ０１」と単語「Ｃ０１．ｊｐｇ」が置換可能であることを示す情報を当該メタ情報に追加してもよい。

[0059] なお、上記のように検索されたメタ情報を同義語辞書に基づいて新たなメタ情報に更新する場合には、同義関係にあるメタ情報に単一の表現が与えられるように更新する必要がある。例えば、あるメタ情報に単語「Ｃ０１」が含まれ、別のメタ情報に単語「Ｃ０１．ｊｐｇ」が含まれ、さらに別のメタ情報にそれらと同義の別の単語が含まれる場合に、単語「Ｃ０１．ｊｐｇ」と同義の単語が全てそれらを代表する単語「Ｃ０１．ｊｐｇ」に変換される必要がある。どの単語が代表として使用されてもよく、例えば同義関係にある複数の単語のリストの先頭の単語が使用されてもよい。

[0060] ここで、図６中に点線で示したように、「Ｃ０１．ｊｐｇ」を介して画像メタ情報と文書メタ情報とが関連付けられる。すなわち、「＿：ｒ１の図面はＣ０１．ｊｐｇであり、Ｃ０１．ｊｐｇのタイプはｔｕｂｅである」言い換えると「＿：ｒ１の図面のタイプはｔｕｂｅである」という意味付けがされたことになる。このような場合、属性値「Ｃ０１．ｊｐｇ」を介して関連付けられた二つの属性名「図面」および「タイプ」を結合した「図面ータイプ」が新たな属性名として扱われてもよい（図１０参照）。

[0061] 以上のようにして、メタ情報データベース１９が構築される。

[0062] 例えば、複数のモーダルのデータから抽出されたメタ情報が同一の（または関連する）概念を含んでいるが、モーダルの性質に応じてそれぞれのモー

ダルのメタ情報に異なる表現が与えられている場合に、上記のようなメタ情報の更新又は置換可能なメタ情報の追加等によって、それらの表現が統一される。これによって、後述する処理において、関連するメタ情報を漏れなく検索することが可能になる。

[0063] 次に、データ加工処理に関する動作を説明する。

[0064] 図７は、本発明の実施例１におけるデータ加工処理の動作を示すフローチャートである。

[0065] まず、データ入力部２１は、ユーザから表データの入力を受けつける（ステップＳ７０１）。ここで入力される表データ（例えば後述する入力データ８０１）は、データ加工システム１の内部にあらかじめ保持された（または、例えば入力部２７５もしくは通信部２７１を介してデータ加工システム１の外部から取得された）任意の構造データであり、これが以下の処理によって非構造データと対応付けられる。

[0066] 次に、入力データ関連付け部２２は、受け付けた表から属性の関係を抽出する（ステップＳ７０２）。

[0067] 図８は、本発明の実施例１において入力データ関連付け部２２が受け付けた表から属性の関係を抽出する処理（ステップＳ７０２）の詳細を説明する図である。

[0068] 入力データ関連付け部は２２、入力データ８０１に対し、レコードの情報を抽出し、属性と属性名の関係を抽出し、表構造属性情報８０２を出力する。表構造属性情報８０２は、画像メタ情報および文書メタ情報の記述と同様に、ＲＤＦを用いて記述することができる。

[0069] 例えば、入力データ８０１の一つのレコードに、項目「ＩＤ」、「部品」、「場所」および「図面」のそれぞれに対応する値として「Ｍ００１」、「Ａ」、「ＪＰ１」および「Ｃ０１」が含まれる場合、そのレコードを「＿：ｒ２」と識別して、「＿：ｒ２のＩＤはＭ００１である」、「＿：ｒ２の部品はＡである」、「＿：ｒ２の場所はＪＰ１である」および「＿：ｒ２の図面はＣ０１である」といったＲＤＦによる記述を含む表構造属性情報８０２

が出力される。

[0070] 次に、入力データ関連付け部２２は、関連性辞書部１７を用いて、表構造属性情報８０２に関連付けを付与する（ステップＳ７０３）。

[0071] 図９は、本発明の実施例１において入力データ関連付け部２２が表構造属性情報８０２に関連付けを付与する処理（ステップＳ７０３）の詳細を説明する図である。

[0072] メタ情報関連付け部１８の機能と同様、入力データ関連付け部２２は、同義語辞書６０１に存在する単語を表構造属性情報８０２から検索し、その単語をそれと同義語関係にある単語に置き変えることによって変更属性情報９０１を生成し、変更属性情報９０１に基づいて表構造属性情報８０２を変更する。図９の例では、同義語辞書６０１内の単語「Ｃ０１」に対して、表構造属性情報８０２から「＿：ｒ２の図面はＣ０１である」が検索され、それが「Ｃ０１」と「Ｃ０１．ｊｐｇ」との同義語関係に基づいて「＿：ｒ２の図面がＣ０１．ｊｐｇである」という属性情報に変更される。

[0073] なお、メタ情報関連付け部１８が実行する関連付け処理と同様に、入力データ関連付け部２２は、表構造属性情報８０２を変更属性情報９０１に書き換える代わりに、「Ｃ０１」と「Ｃ０１．ｊｐｇ」が置換可能であることを示す情報を表構造属性情報８０２に追加してもよい。

[0074] 次に、関連メタ情報検索部２０は、表構造属性情報８０２を用いて、メタ情報データベース１９を検索する（ステップＳ７０４）。ここでは、メタ情報データベース１９に対する問い合わせとして、入力と同じ属性関係を持つメタ情報が選択される。表構造属性情報８０２に存在する「＿：ｒ２の図面はＣ０１．ｊｐｇである」に対して、メタ情報データベース１９に「＿：ｒ１の図面はＣ０１．ｊｐｇである」という情報が含まれるため（図６参照）、それを入力と同一の属性関係として見つけることができる。また同様に、「＿：ｒ２の場所は東京である」に対して「＿：ｒ１の場所は東京である」も見つけることができる。

[0075] ここで、関連メタ情報検索部２０は、複数のレコードの属性関係が所定の

条件を満たす場合に、それらを同一のレコードと見なすことができる。例えば、複数のレコードが二つ以上の共通する属性関係（すなわち共通する属性名と属性値との組）を有する場合にそれらを同一のレコードと見なすとすれば、上記の例では＿：ｒ１と＿：ｒ２とを同一レコードと見なせる。なお、二つのレコードを同一と見なすための条件は、上記の例に限定されない。例えば、同一レコードとみなす属性名のリストを予め定義し、関連メタ情報検索部２０は、そのリストに含まれる属性名に対応する属性値が同一である複数のレコードを同一のレコードと見なしてもよい。あるいは、関連メタ情報検索部２０は、メタ情報の検索結果から同一のレコードを推定してもよい。

[0076] 次に、関連メタ情報検索部２０は、入力データに含まれる＿：ｒ２と同一レコードとみなされた＿：ｒ１に関する属性関係をメタ情報データベース１９から取得する（ステップＳ７０５）。ここでは、「＿：ｒ１の作成者はＡｌｉｃｅである」、「＿：ｒ１の納期は２０１２／７／３０である」の２つのメタ情報が取得できる（図６参照）。さらに、関連メタ情報検索部２０は、属性値に関する属性があればそれをさらにたどることによって、「＿：ｒ１の図面（はＣ０１．ｊｐｇであり、Ｃ０１．ｊｐｇ）のタイプはｔｕｂｅである」というメタ情報も取得できる。こうして、関連メタ情報検索部２０は、「＿：ｒ２の作成者はＡｌｉｃｅである」、「＿：ｒ２の納期は２０１２／７／３０である」、および「＿：ｒ２の図面（はＣ０１．ｊｐｇであり、Ｃ０１．ｊｐｇ）のタイプはｔｕｂｅである」の３つの属性情報を追加属性情報として決定する。

[0077] 次に、データ加工部２３は、関連メタ情報検索部２０により決定された追加属性情報に基づいて、入力データ８０１に属性を追加する（ステップＳ７０６）。

[0078] 図１０は、本発明の実施例１においてデータ加工部２３が入力データ８０１に属性を追加する処理（ステップＳ７０６）を説明する図である。

[0079] データ加工部２３が表構造属性情報８０２に対して追加属性情報１００２を追加し、表の表現に戻すことによって、加工データ１００３が得られる。

図１０の例では、「＿：ｒ２の作成者はＡｌｉｃｅである」、「＿：ｒ２の納期は２０１２／７／３０である」、および「＿：ｒ２の図面（はＣ０１．ｊｐｇであり、Ｃ０１．ｊｐｇ）のタイプはｔｕｂｅである」の３つの属性情報が表形式で追加される。ここで、上記の「図面」と「Ｃ０１．ｊｐｇ」と「タイプ」との関係のように、ある属性名に対応する属性値にさらに別の属性名が関連付けられる場合、それらの複数の属性名表現を繋げて新たな属性名として表示する。図１０の例では、「図面」の「タイプ」という複数の属性名表現をハイフン記号で繋げることによって、「図面 - タイプ」というように表現される。このような加工データ１００３によって、入力データ８０１と、画像メタ情報４０４と、文書メタ情報５０４との関連が示される。

[0080] 次に、データ出力部２４は、加工データ１００３を出力する。

[0081] 以上のようにして、データ加工処理が行われる。

[0082] 本実施例では、予め構築しておいたメタ情報データベースに基づいて、入力された表を拡張する一例を示した。メタ情報データベースにおける単語の関連付けと、入力された表内の単語の関連付けに、同一の関連性辞書として同義語辞書を用いることによって、文書データや画像データから抽出された情報を、入力された表に関連付けることが可能となる。また、メタ情報をＲＤＦで表現しておくことによって、表と文書と画像という異なるモダリティのデータを、単一の同義語辞書に基づいて関連付けることが可能となる。

[0083] また、本実施例では、画像のタイプについてのメタ情報を利用したが、他の情報抽出による結果を用いることもできる。例えば顔写真のような画像に対しては、顔画像認識を用いて、抽出された人名をメタ情報として利用してもよい。

実施例 2

[0084] 以下、本発明の実施例２を、図面を用いて説明する。

[0085] 本実施例では、音声データとテキストデータから抽出されたメタ情報データベースに基づいて、入力されたテキストデータに関連する情報を提示する、データ加工システムの例を説明する。本システムは、例えば、コールセン

ターにおいて蓄積される、通話録音音声とオペレータの応対ログを管理するために用いることができる。本システムは、テキストを入力されると、録音音声と応対ログから抽出されたメタ情報を用いて、関連する音声および応対ログを表示する。これによって、全ての録音音声を聴くことなく、効率的に情報を探索することが可能となる。

[0086] 図１１は、本発明の実施例２のデータ加工システムの全体構成を説明するブロック図である。

[0087] 本実施例は、実施例１における画像データと文書データを、音声データとテキストデータに置き換え、メタ情報データベースを、音声データ用とテキストデータ用でそれぞれ構築しておき、メタ情報データベース構築時にこれらの関連付けを行なわず、メタ情報検索時に関連付ける構成とした。

[0088] 上記およびこれから詳細に説明する相違点を除き、実施例２のデータ加工システムの各部は、図１に示された実施例１の同一の符号を付された各部と同一の機能を有するため、それらの説明は省略する。

[0089] 本実施例のデータ加工システム１は、実施例１と同様に、データソースサーバ２、ＥＴＬサーバ３、ストレージサーバ４、メタ情報抽出サーバ５、メタ情報検索サーバ６およびデータ加工サーバ７によって構成される。

[0090] データソースサーバ２は、音声およびテキストを管理する装置であり、録音音声データをＩＤと結び付けて管理するリレーショナルデータベースと、音声データファイルとテキストファイルを保存するファイルサーバとを備える。

[0091] ＥＴＬサーバ３は、データソースサーバ２に保存されている音声データおよびテキストデータを、ストレージサーバ４に保存する機能を備える。ここで、音声データのフォーマットを統一するなどの変換が行われる。

[0092] ストレージサーバ４は、音声データ保存部５１およびテキストデータ保存部５２を備え、複数のデータソースから収集された音声データおよびテキストデータを統一された形式で保存する。

[0093] メタ情報抽出サーバ５は、音声用辞書部５３、音声メタ情報抽出部５４、

テキスト用辞書部５５、テキストメタ情報抽出部５６、音声メタ情報データベース５７およびテキストメタ情報データベース５８を備え、ストレージサーバ４にあるデータから抽出したメタ情報の管理を行う。

[0094] メタ情報検索サーバ６は、関連性辞書部１７、メタ情報関連付け部１８および関連メタ情報検索部２０を備え、検索要求を受けつけて、音声メタ情報データベース５７およびテキストメタ情報データベース５８をそれぞれ検索した結果を返す。

[0095] データ加工サーバ７は、データ入力部２１、入力データ関連付け部２２、データ加工部２３およびデータ出力部２４を備え、入力されたデータをメタ情報に基づいて加工し出力する。

[0096] 本実施例のデータ加工システム１のハードウェア構成は、図２に示した実施例１の構成と同様である。ただし、ストレージサーバ４のディスク２４４は、画像データ保存部１１および文書データ保存部１２の代わりに音声データ保存部５１およびテキストデータ保存部５２を含む。メタ情報抽出サーバ５のディスク２５４は、画像用辞書部１３、文書用辞書部１５、関連性辞書部１７およびメタ情報データベース１９の代わりに、音声用辞書部５３、テキスト用辞書部５５、音声メタ情報データベース５７およびテキストメタ情報データベース５８を含む。メタ情報抽出サーバ５のメモリ２５３は、画像メタ情報抽出部１４、文書メタ情報抽出部１６およびメタ情報関連付け部１８の代わりに、音声メタ情報抽出部５４およびテキストメタ情報抽出部５６を含む。メタ情報検索サーバ６のメモリ２６３は、関連メタ情報検索部２０に加えてメタ情報関連付け部１８を含み、ディスク２６４は関連性辞書部１７を含む。

[0097] 次に、上記のように構成される、本実施例に係るデータ加工システム１の動作を説明する。本システムの動作は、実施例１と同様に、メタ情報データベース構築処理とデータ加工処理に分けられる。メタ情報データベースの構築処理は、以下に説明する相違点を除き、図３に示した実施例１におけるメタ情報データベースの構築処理と同様である。

[0098] まず、ＥＴＬサーバ３は、データソースサーバ２から、音声データとテキストデータを取得する（ステップＳ３０１）。続いて、ＥＴＬサーバ３は、音声データとテキストデータに対し、必要な変換を施す（ステップＳ３０２）。続いて、ＥＴＬサーバ３は、変換された音声データとテキストデータを、ストレージサーバ４へ保存する（ステップＳ３０３）。

[0099] 次に、メタ情報抽出サーバ５は、ストレージサーバ４から音声データとテキストデータを取得する（ステップＳ３０４）。

[0100] 次に、音声メタ情報抽出部５４は、音声用辞書部５３に基づいて、音声データに対するメタ情報を抽出する（ステップＳ３０５）。

[0101] 図１２は、本発明の実施例２において音声メタ情報抽出部５４が音声データに対するメタ情報を抽出する処理（ステップＳ３０５）の詳細を説明する図である。

[0102] 音声用辞書部５３は、音声中から検出するキーワードのリストであるキーワードリスト１２０１を保持している。音声メタ情報抽出部５４は、音声用辞書部５３に基づく音声認識技術によって、音声データ１２０３を文単位に分割し、その中に出現するキーワードを付与することによって、音声メタ情報１２０４を生成する。図１２の例では、音声データ１２０３から、２つのキーワード「製品Ａ」と「まいります」が抽出される。音声メタ情報１２０４は、「Ｗ０１．ｗａｖの文（ｓｅｎｔｅｎｃｅ）には__：ｓ１がある」、「Ｗ０１．ｗａｖの文には__：ｓ２がある」、「__：ｓ１のキーワードには「製品Ａ」がある」、「__：ｓ２のキーワードには「まいります」がある」といった４つの関係を記述している。これらの関係は、実施例１と同様にＲＤＦを用いて記述することができる。

[0103] 同様に、ステップＳ３０５において、テキストメタ情報抽出部５６は、テキスト用辞書部５５に基づいて、テキストデータに対するメタ情報を抽出する。

[0104] 図１３は、本発明の実施例２においてテキストメタ情報抽出部５６がテキストデータに対するメタ情報を抽出する処理（ステップＳ３０５）の詳細を

説明する図である。

[0105] テキスト用辞書部５５は、テキスト中から抽出するキーワードリスト１３０１を保持している。テキストメタ情報抽出部５６は、テキスト用辞書部５５と、形態素解析の結果に基づいて、テキストデータ１３０３を解析することによって、テキストメタ情報１３０４を生成する。図１３の例では、テキストデータ１３０３に対して、３つのキーワード「製品Ａ」、「クレーム」および「訪問」が抽出される。テキストメタ情報１３０４は、音声メタ情報１２０４と同様に、ＲＤＦを用いて記述することができる。

[0106] 次に、音声メタ情報抽出部５４は、抽出した音声メタ情報を音声メタ情報データベース５７に保存する（ステップＳ３０６）。同様に、ステップＳ３０６において、テキストメタ情報抽出部５６は、抽出したテキストメタ情報をテキストメタ情報データベース５８に保存する。本実施例では、実施例１と異なり、抽出されたメタ情報が各モーダルで分離したデータベースとして保存される。

[0107] 本実施例では、実施例１と異なり、メタ情報保存（ステップＳ３０６）の後に、メタ情報関連付けを行うステップＳ３０７は行われない。

[0108] 次に、データ加工処理に関する動作を説明する。

[0109] 図１４は、本発明の実施例２におけるデータ加工処理の動作を示すフローチャートである。

[0110] まず、データ入力部２１は、ユーザからテキストデータの入力を受けつける（ステップＳ１４０１）。

[0111] 次に、入力データ関連付け部２２は、受け付けたテキストデータからキーワードを抽出する（ステップＳ１４０２）。ここでは、テキストデータとして「マシンＡＢＣ　訪問」が入力され、キーワード「マシンＡＢＣ」および「訪問」が抽出された場合を例として説明する。あるいは、例えばキーワードを含む自然文のテキストデータが入力された場合等に、入力データ関連付け部２２は、形態素解析等を用いてキーワードを抽出してもよい。

[0112] 次に、入力データ関連付け部２２は、関連性辞書部１７を用いて、抽出さ

れたキーワードに関連付けを付与する（ステップＳ１４０３）。

[0113] 図１５は、本発明の実施例２において入力データ関連付け部２２が抽出されたキーワードに関連付けを付与する処理（ステップＳ１４０３）の詳細を説明する図である。

[0114] 関連性辞書部１７には、話し言葉と書き言葉を相互に変換するための情報等を含む同義語辞書６０１が予め構築されている。これによって「訪問」と「まいります」を対応付けることができる。図１５に示す同義語辞書６０１は、さらに、「マシンＡＢＣ」と「製品Ａ」とを対応付ける情報を含む。この例において「マシンＡＢＣ」および「製品Ａ」は同一の製品の別名（例えば、一方が製造元の社内で使用される識別情報であり、もう一方が顧客向けに使用される商品名であるなど）である。

[0115] 入力データ関連付け部２２は、関連性辞書部１７に含まれる同義語辞書６０１に基づいて、入力されたテキストデータ１５０１から抽出されたキーワード「マシンＡＢＣ」および「訪問」をそれぞれ「製品Ａ」および「まいります」に関連付けることができる。

[0116] 次に、メタ情報関連付け部１８は、上記の関連性辞書部１７を用いた関連付けの結果に基づいて、抽出されたキーワードを展開する（ステップＳ１４０４）。具体的には、メタ情報関連付け部１８は、抽出されたキーワード「マシンＡＢＣ」および「訪問」を、音声メタ情報データベース５７を検索するためのキーワード（すなわち音声用検索クエリ）「製品Ａ」および「まいります」に展開する。同様に、メタ情報関連付け部１８は、抽出されたキーワード「マシンＡＢＣ」および「訪問」を、テキストメタ情報データベース５８を検索するためのキーワード（すなわちテキスト用検索クエリ）「製品Ａ」および「訪問」に展開する。

[0117] 次に、関連メタ情報検索部２０は、キーワード「製品Ａ」および「まいります」によって音声メタ情報データベース５７を検索し、さらに、キーワード「製品Ａ」および「訪問」によってテキストメタ情報データベース５８を検索する（ステップＳ１４０５）。

[0118] 次に、データ加工部２３は、関連メタ情報検索部２０によって検索された音声メタ情報とテキストメタ情報を関連付ける（ステップＳ１４０６）。

[0119] 図１６は、本発明の実施例２においてデータ加工部２３が音声メタ情報とテキストメタ情報を関連付ける処理（ステップＳ１４０６）を説明する図である。

[0120] 例えば、ステップ１４０１においてテキストデータ１５０１「マシンＡＢＣ　訪問」が入力され、ステップＳ１４０５において音声メタ情報１２０４およびテキストメタ情報１３０４が検索された場合、データ加工部２３は、テキストメタ情報１３０４に対応するテキストデータ１３０３に、音声メタ情報１２０４に対応する音声データ１２０３を関連付けることによって加工データ１６０１を生成する。

[0121] 具体的には、例えば、データ加工部２３は、テキストデータ１３０３に含まれるキーワード「製品Ａ」および「訪問」に、それぞれ、音声データ１２０３に含まれるキーワード「製品Ａ」および「まいります」へのリンク（図１６の例ではそれぞれ「ｌｉｓｔｅｎ　Ｗ０１．ｗａｖ＠ｓ１」および「ｌｉｓｔｅｎ　Ｗ０１．ｗａｖ＠ｓ２」）を追加することによって、加工データ１６０１を生成する。言い換えると、加工データ１６０１は、テキストデータ１３０３に含まれる、入力されたキーワードに対応するキーワードと、それに対応する音声データ１２０３の部分と、の関連を示す情報である。

[0122] 次に、データ出力部２４は、加工データ１６０１を出力する（ステップＳ１４０７）。

[0123] 以上のようにして、データ加工処理が行われる。

[0124] 本実施例では、実施例１と異なり、事前に音声メタ情報データベース５７とテキストメタ情報データベース５８との間の関連付けを行っていない。したがって、これらのデータベースは、モーダルに応じて異なる表現（例えば話し言葉と書き言葉等の表現）が与えられた、同一のまたは関連する概念の情報を含んでいる可能性がある。しかし、入力キーワードを関連性辞書部１７に基づいて展開することで、入力キーワードが各モーダルの特徴に合わせ

た表現に変換されるため、各モーダルの特徴（例えば話し言葉または書き言葉のいずれが含まれるか）に合わせた検索が可能となる。その結果、例えば通話応対記録のテキスト上に、関連する音声データのリンクを生成する等の加工を行うことが可能となる。

[0125] 以上説明した本発明の各実施形態によれば、メタ情報データベースに基づいて、入力されたデータを加工することが可能となる。メタ情報データベースにおける単語の関連付けと、入力された表内の単語の関連付けに同一の関連性辞書として、同義語辞書または話し言葉変換辞書といった置換可能な単語の組み合わせを定義する辞書を用いることによって、文書データ・画像データ・音声データなど、マルチモーダルデータから抽出された情報を、入力されたデータに関連付けることが可能となる。

[0126] なお、上述した各実施形態では、各サーバのＣＰＵ上で実行されるプログラムによって、データ加工システムの各種機能を実現しているが、それらの一部又は全部が、例えば集積回路等の電子部品を用いたハードウェアによって実現されてもよい。

[0127] 本発明は上述した実施形態に限定されるものではなく、様々な変形例が含まれる。本実施例では、データマイニングのための表生成システムや、コールセンターの音声ログ解析のシステムを想定したが、例えば、電子カルテデータと医用画像データを管理するシステムや、放送データの編集システムなど、マルチモーダルのデータを管理するための様々なシステムに適用することができる。

[0128] 上記の実施形態の各機能を実現するプログラム、テーブル、ファイル等の情報は、不揮発性半導体メモリ、ハードディスクドライブ、ＳＳＤ（Ｓｏｌｉｄ Ｓｔａｔｅ Ｄｒｉｖｅ）等の記憶デバイス、または、ＩＣカード、ＳＤカード、ＤＶＤ等の計算機読み取り可能な非一時的データ記憶媒体に格納することができる。

[0129] 本発明は上記した実施形態に限定されるものではなく、様々な変形例が含まれる。例えば、上記した実施形態は本発明を分かりやすく説明するために

詳細に説明したものであり、必ずしも説明した全ての構成を備えるものに限定されるものではない。また、ある実施形態の構成の一部を他の実施形態の構成に置き換えることが可能であり、また、ある実施形態の構成に他の実施形態の構成を加えることも可能である。また、各実施形態の構成の一部について、他の構成の追加・削除・置換をすることが可能である。

請求の範囲

［請求項1］ 一つ以上のプロセッサと、前記一つ以上のプロセッサに接続される一つ以上の記憶装置と、を有するデータ加工システムであって、

複数の種類のデータからメタ情報を抽出する条件を定義するメタ情報抽出用辞書情報と、前記複数の種類のデータから抽出されたメタ情報を関連付ける条件を定義する関連性辞書情報と、を保持し、

前記複数の種類のデータから、前記メタ情報抽出用辞書情報に基づいて前記メタ情報を抽出し、

入力されたデータからメタ情報を抽出し、

前記関連性辞書情報に基づいて、前記入力されたデータから抽出されたメタ情報と前記複数の種類のデータから抽出されたメタ情報とを関連付け、

前記関連付けの結果に基づいて、前記複数の種類のデータと、前記入力されたデータと、それらのデータから抽出されたメタ情報と、のいずれかの組み合わせの関連を示す情報を出力することを特徴とするデータ加工システム。

［請求項2］ 請求項１に記載のデータ加工システムであって、

前記複数の種類のデータは、第１の種類のデータおよび第２の種類のデータを含み、

前記メタ情報抽出用辞書情報は、前記第１の種類のデータから第１のメタ情報を抽出する条件を定義する第１のメタ情報抽出用辞書情報と、前記第２の種類のデータから第１のメタ情報を抽出する条件を定義する第２のメタ情報抽出用辞書情報と、を含み、

前記関連性辞書情報は、メタ情報と置換可能な情報を定義する情報を含み、

前記データ加工システムは、

前記第１の種類のデータおよび前記第２の種類のデータから、それぞれ、前記第１のメタ情報抽出用辞書情報および前記第２のメタ情報

抽出用辞書情報に基づいて、前記第１のメタ情報および前記第２のメタ情報を抽出し、

入力されたデータから第３のメタ情報を抽出し、

前記関連性辞書情報に基づいて、前記第１のメタ情報と置換可能な情報、前記第２のメタ情報と置換可能な情報、および前記第３のメタ情報と置換可能な情報の少なくとも一つを特定し、

前記第３のメタ情報または前記第３のメタ情報と置換可能な情報を用いて、前記第１のメタ情報または前記第１のメタ情報と置換可能な情報、および、前記第２のメタ情報または前記第２のメタ情報と置換可能な情報を検索し、

前記検索の結果に基づいて、前記第３のメタ情報または前記第３のメタ情報と置換可能な情報と、前記第１のメタ情報または前記第１のメタ情報と置換可能な情報と、前記第２のメタ情報または前記第２のメタ情報と置換可能な情報と、を関連付けることを特徴とするデータ加工システム。

［請求項3］ 請求項２に記載のデータ加工システムであって、

前記関連性辞書情報に基づいて、前記第１のメタ情報と置換可能な情報を特定し、前記関連性辞書情報に基づいて、前記第２のメタ情報と置換可能な情報を特定し、前記特定された置換可能な情報と前記第１のメタ情報と前記第２のメタ情報とが所定の条件を満たすか否かの判定に基づいて、前記第１のメタ情報と前記第２のメタ情報とを関連付け、

前記関連性辞書情報に基づいて、前記第３のメタ情報と置換可能な情報を特定し、

前記第１のメタ情報または前記第１のメタ情報と置換可能な情報が検索結果として取得された場合、前記第３のメタ情報または前記第３のメタ情報と置換可能な情報と、前記第１のメタ情報または前記第１のメタ情報と置換可能な情報と、前記第１のメタ情報に関連付けられ

た前記第２のメタ情報または前記第２のメタ情報と置換可能な情報と、を関連付けることを特徴とするデータ加工システム。

[請求項4] 請求項２に記載のデータ加工システムであって、

前記関連性辞書情報に基づいて、前記第１のメタ情報と置換可能な情報を特定する手順、および、前記関連性辞書情報に基づいて、前記第２のメタ情報と置換可能な情報を特定する手順を実行せず、前記関連性辞書情報に基づいて、前記第３のメタ情報と置換可能な情報を特定することを特徴とするデータ加工システム。

[請求項5] 請求項２に記載のデータ加工システムであって、

前記関連性辞書情報は、前記メタ情報に含まれる単語と同義の他の単語を定義する情報を含むことを特徴とするデータ加工システム。

[請求項6] 請求項２に記載のデータ加工システムであって、

前記関連性辞書情報は、前記メタ情報に含まれる書き言葉の単語と同義の話し言葉の単語を定義する情報、および、前記メタ情報に含まれる話し言葉の単語と同義の書き言葉の単語を定義する情報を含むことを特徴とするデータ加工システム。

[請求項7] 請求項２に記載のデータ加工システムであって、

前記関連性辞書情報は、前記メタ情報に含まれる第１の言語の単語と同義の第２の言語の単語を定義する情報を含むことを特徴とするデータ加工システム。

[請求項8] 請求項２に記載のデータ加工システムであって、

前記メタ情報は、三項関係を含む情報であり、

前記関連性辞書情報は、抽出されたメタ情報が三項関係に係る所定の条件に合致するときに、合致した三項関係を利用して新たな三項関係を生成するための規則を含むことを特徴とするデータ加工システム。

[請求項9] 請求項２に記載のデータ加工システムであって、

前記第１の種類のデータは、テキスト、表構造、音声、画像または

文書のいずれかである第１のモーダルのデータであり、

前記第２の種類のデータは、テキスト、表構造、音声、画像または文書のうち、前記第１のモーダルとは異なる第２のモーダルのデータであり、

前記関連性辞書情報は、異なるモーダル間で置換される情報を定義する情報を含むことを特徴とするデータ加工システム。

[請求項10] 一つ以上のプロセッサと、前記一つ以上のプロセッサに接続される一つ以上の記憶装置と、を有する計算機システムによるデータ加工方法であって、

前記計算機システムは、複数の種類のデータからメタ情報を抽出する条件を定義するメタ情報抽出用辞書情報と、前記複数の種類のデータから抽出されたメタ情報を関連付ける条件を定義する関連性辞書情報と、を保持し、

前記データ加工方法は、

前記複数の種類のデータから、前記メタ情報抽出用辞書情報に基づいて前記メタ情報を抽出する第１手順と、

入力されたデータからメタ情報を抽出する第２手順と、

前記関連性辞書情報に基づいて、前記入力されたデータから抽出されたメタ情報と前記複数の種類のデータから抽出されたメタ情報とを関連付ける第３手順と、

前記関連付けの結果に基づいて、前記複数の種類のデータと、前記入力されたデータと、それらのデータから抽出されたメタ情報と、のいずれかの組み合わせの関連を示す情報を出力する第４手順と、を含むことを特徴とするデータ加工方法。

[請求項11] 請求項１０に記載のデータ加工方法であって、

前記複数の種類のデータは、第１の種類のデータと、第２の種類のデータと、を含み、

前記メタ情報抽出用辞書情報は、前記第１の種類のデータから第１

のメタ情報を抽出する条件を定義する第１のメタ情報抽出用辞書情報と、前記第２の種類のデータから第１のメタ情報を抽出する条件を定義する第２のメタ情報抽出用辞書情報と、を含み、

前記関連性辞書情報は、メタ情報と置換可能な情報を定義する情報を含み、

前記第１手順は、前記第１の種類のデータおよび前記第２の種類のデータから、それぞれ、前記第１のメタ情報抽出用辞書情報および前記第２のメタ情報抽出用辞書情報に基づいて、前記第１のメタ情報および前記第２のメタ情報を抽出する第５手順を含み、

前記第２手順は、入力されたデータから第３のメタ情報を抽出する第６手順を含み、

前記第３手順は、

前記関連性辞書情報に基づいて、前記第１のメタ情報と置換可能な情報、前記第２のメタ情報と置換可能な情報、および前記第３のメタ情報と置換可能な情報の少なくとも一つを特定する第７手順と、

前記第３のメタ情報または前記第３のメタ情報と置換可能な情報を用いて、前記第１のメタ情報または前記第１のメタ情報と置換可能な情報、および、前記第２のメタ情報または前記第２のメタ情報と置換可能な情報を検索する第８手順と、

前記検索の結果に基づいて、前記第３のメタ情報または前記第３のメタ情報と置換可能な情報と、前記第１のメタ情報または前記第１のメタ情報と置換可能な情報と、前記第２のメタ情報または前記第２のメタ情報と置換可能な情報と、を関連付ける第９手順と、を含むことを特徴とするデータ加工方法。

［請求項12］ 請求項１１に記載のデータ加工方法であって、

前記第７手順は、

前記関連性辞書情報に基づいて、前記第１のメタ情報と置換可能な情報を特定し、前記関連性辞書情報に基づいて、前記第２のメタ情報

と置換可能な情報を特定し、前記特定された置換可能な情報と前記第１のメタ情報と前記第２のメタ情報とが所定の条件を満たすか否かの判定に基づいて、前記第１のメタ情報と前記第２のメタ情報とを関連付ける手順と、

前記関連性辞書情報に基づいて、前記第３のメタ情報と置換可能な情報を特定する手順と、を含み、

前記第９手順は、前記第８手順において前記第１のメタ情報または前記第１のメタ情報と置換可能な情報が検索結果として取得された場合、前記第３のメタ情報または前記第３のメタ情報と置換可能な情報と、前記第１のメタ情報または前記第１のメタ情報と置換可能な情報と、前記第１のメタ情報に関連付けられた前記第２のメタ情報または前記第２のメタ情報と置換可能な情報と、を関連付ける手順を含むことを特徴とするデータ加工方法。

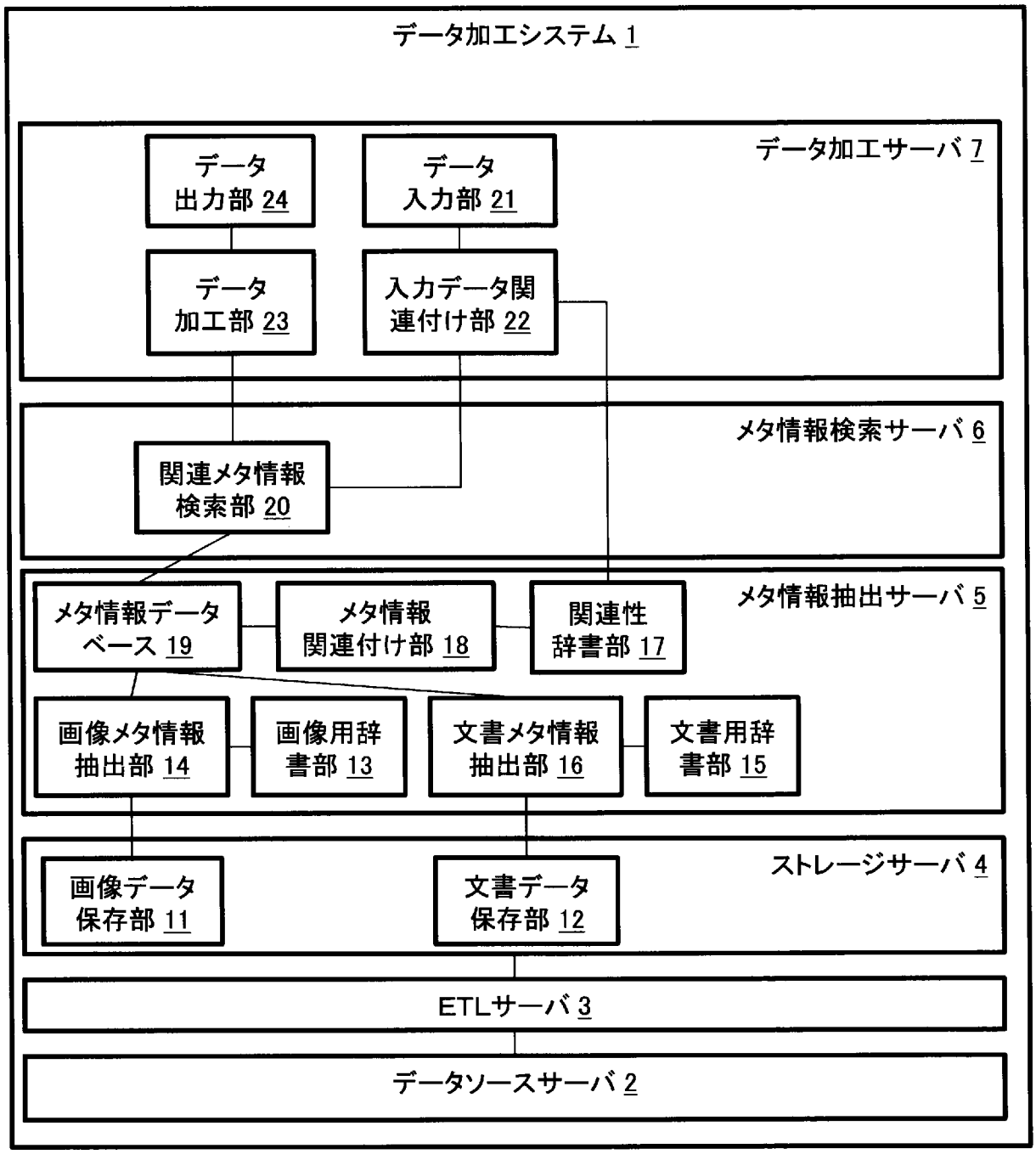
［請求項13］ 請求項１１に記載のデータ加工方法であって、

前記第７手順は、

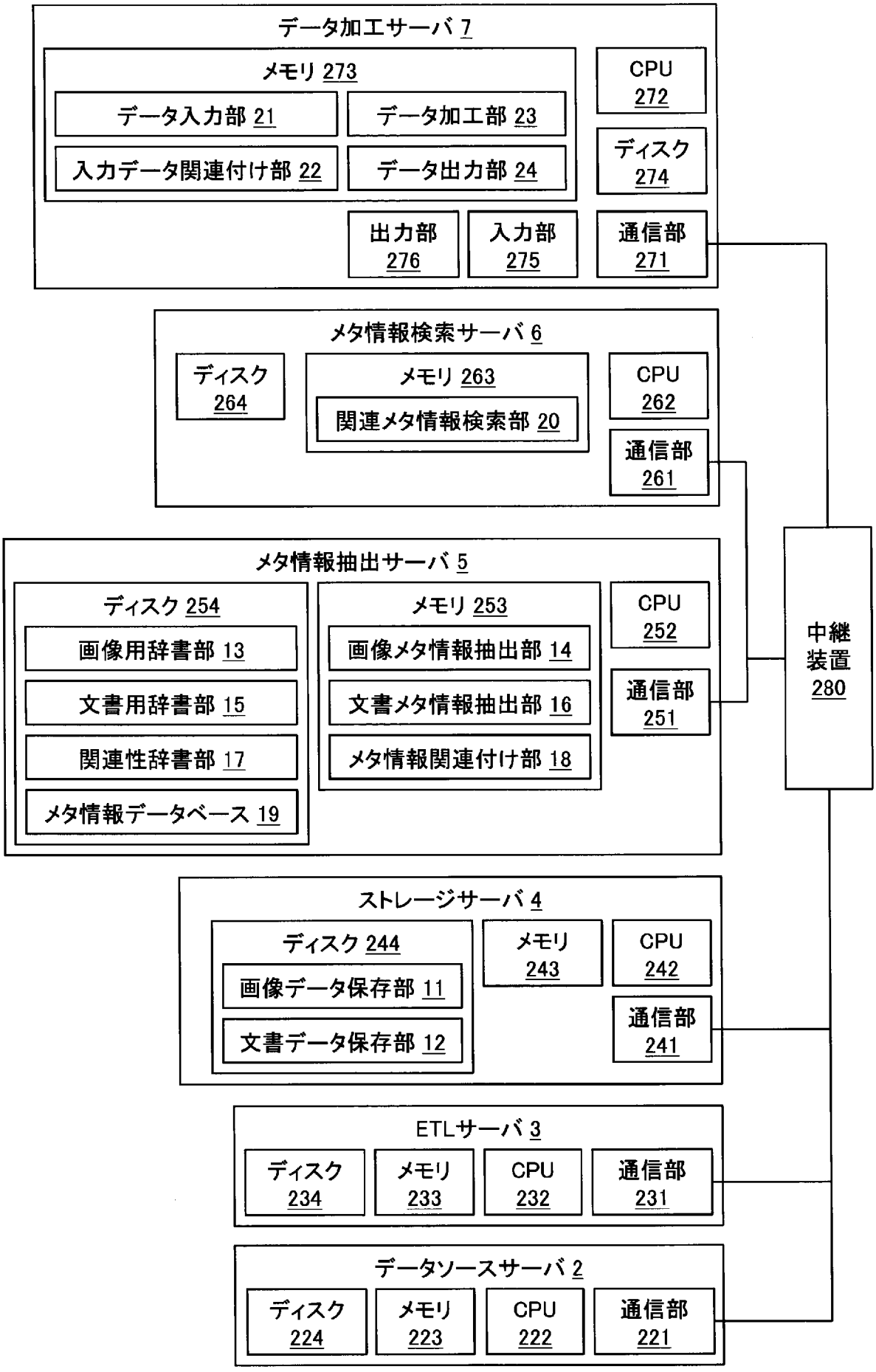
前記関連性辞書情報に基づいて、前記第１のメタ情報と置換可能な情報を特定する手順、および、前記関連性辞書情報に基づいて、前記第２のメタ情報と置換可能な情報を特定する手順を含まず、

前記関連性辞書情報に基づいて、前記第３のメタ情報と置換可能な情報を特定する手順を含むことを特徴とするデータ加工方法。

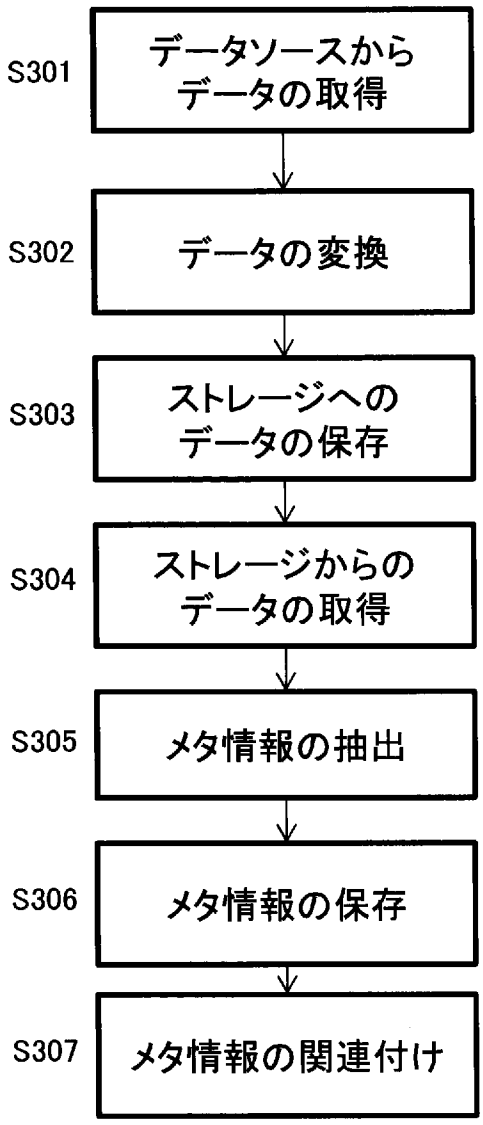
[図1]


データ加工システム 1
データ加工サーバ 7
データ出力部 24
データ入力部 21
データ加工部 23
入力データ関連付け部 22
メタ情報検索サーバ 6
関連メタ情報検索部 20
メタ情報抽出サーバ 5
メタ情報データベース 19
メタ情報関連付け部 18
関連性辞書部 17
画像メタ情報抽出部 14
画像用辞書部 13
文書メタ情報抽出部 16
文書用辞書部 15
ストレージサーバ 4
画像データ保存部 11
文書データ保存部 12
ETLサーバ 3
データソースサーバ 2

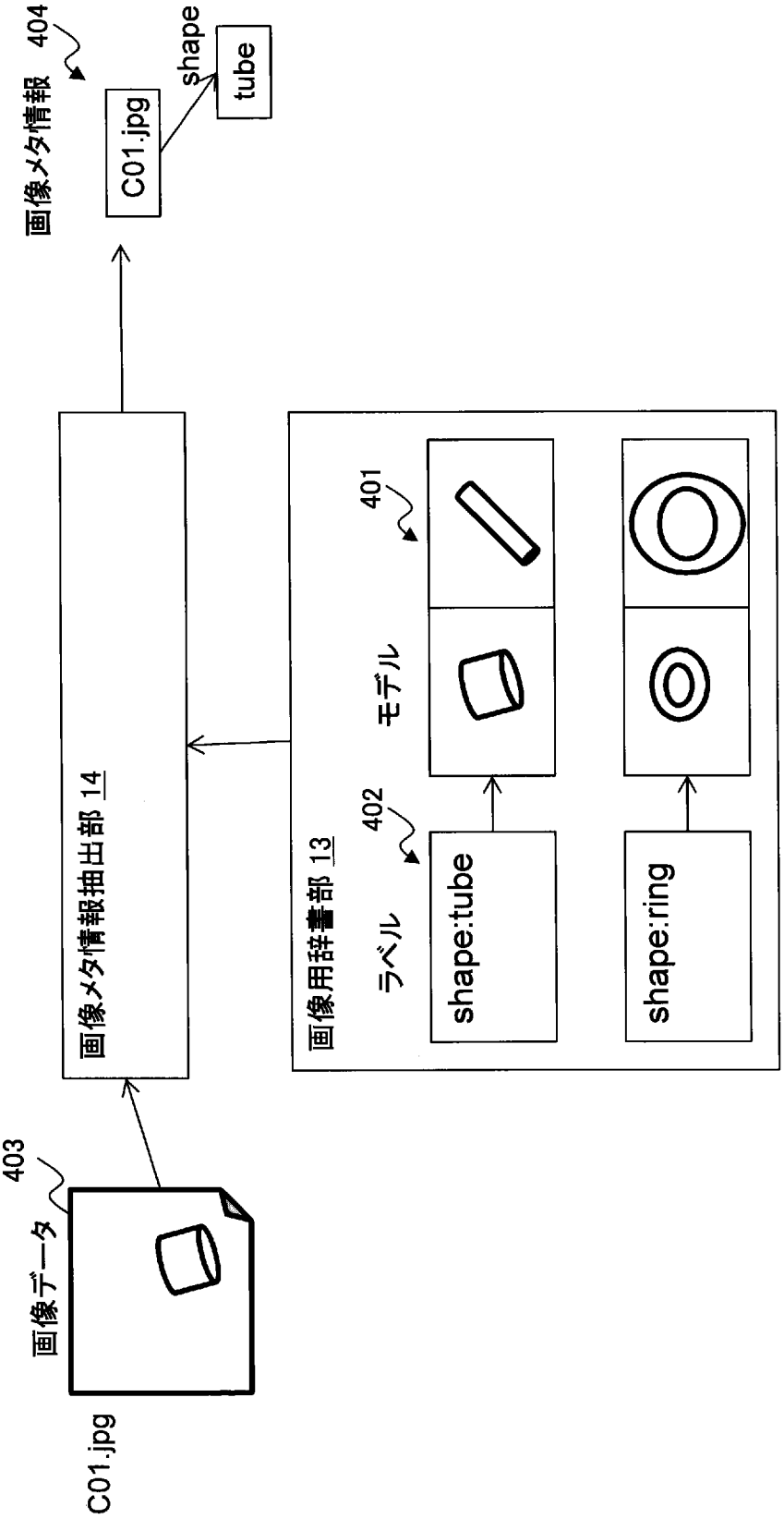
[図2]


データ加工サーバ 7
メモリ 273
データ入力部 21
データ加工部 23
入力データ関連付け部 22
データ出力部 24
CPU 272
ディスク 274
出力部 276
入力部 275
通信部 271
メタ情報検索サーバ 6
ディスク 264
メモリ 263
関連メタ情報検索部 20
CPU 262
通信部 261
メタ情報抽出サーバ 5
ディスク 254
画像用辞書部 13
文書用辞書部 15
関連性辞書部 17
メタ情報データベース 19
メモリ 253
画像メタ情報抽出部 14
文書メタ情報抽出部 16
メタ情報関連付け部 18
CPU 252
通信部 251
中継装置 280
ストレージサーバ 4
ディスク 244
画像データ保存部 11
文書データ保存部 12
メモリ 243
CPU 242
通信部 241
ETLサーバ 3
ディスク 234
メモリ 233
CPU 232
通信部 231
データソースサーバ 2
ディスク 224
メモリ 223
CPU 222
通信部 221

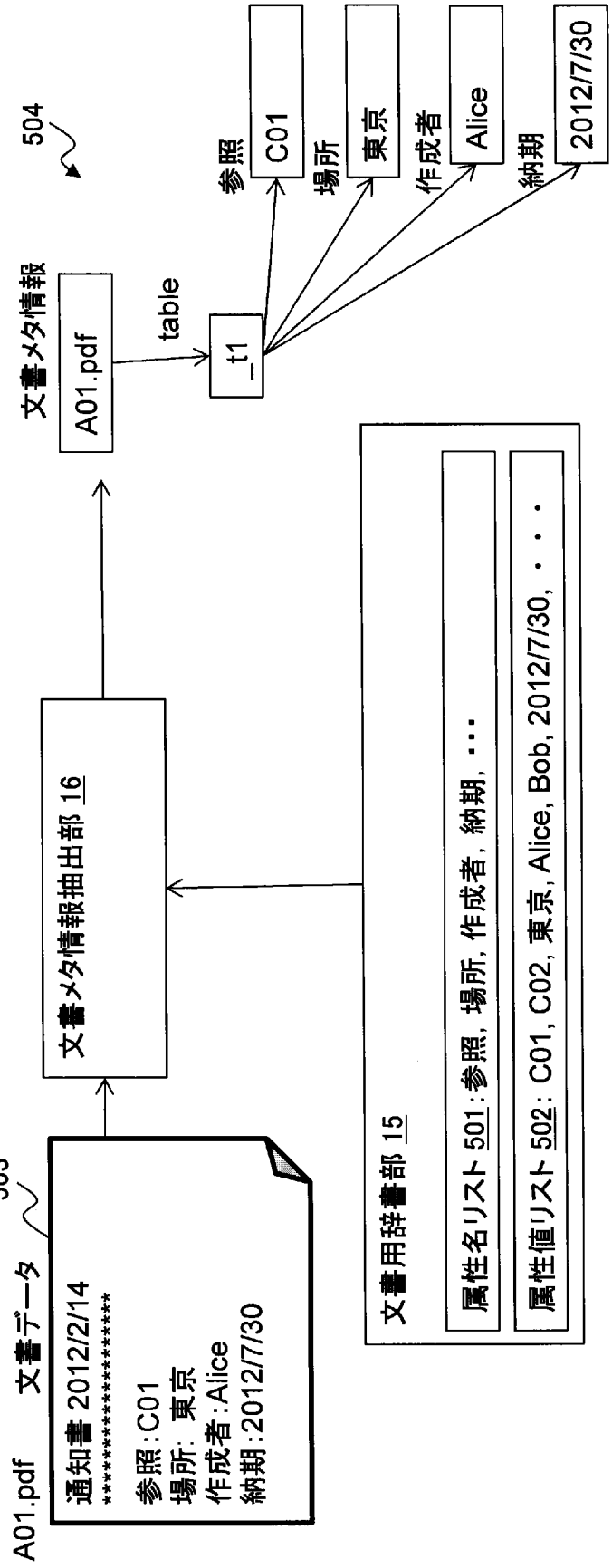
[図3]


S301
データソースから
データの取得
S302
データの変換
S303
ストレージへの
データの保存
S304
ストレージからの
データの取得
S305
メタ情報の抽出
S306
メタ情報の保存
S307
メタ情報の関連付け

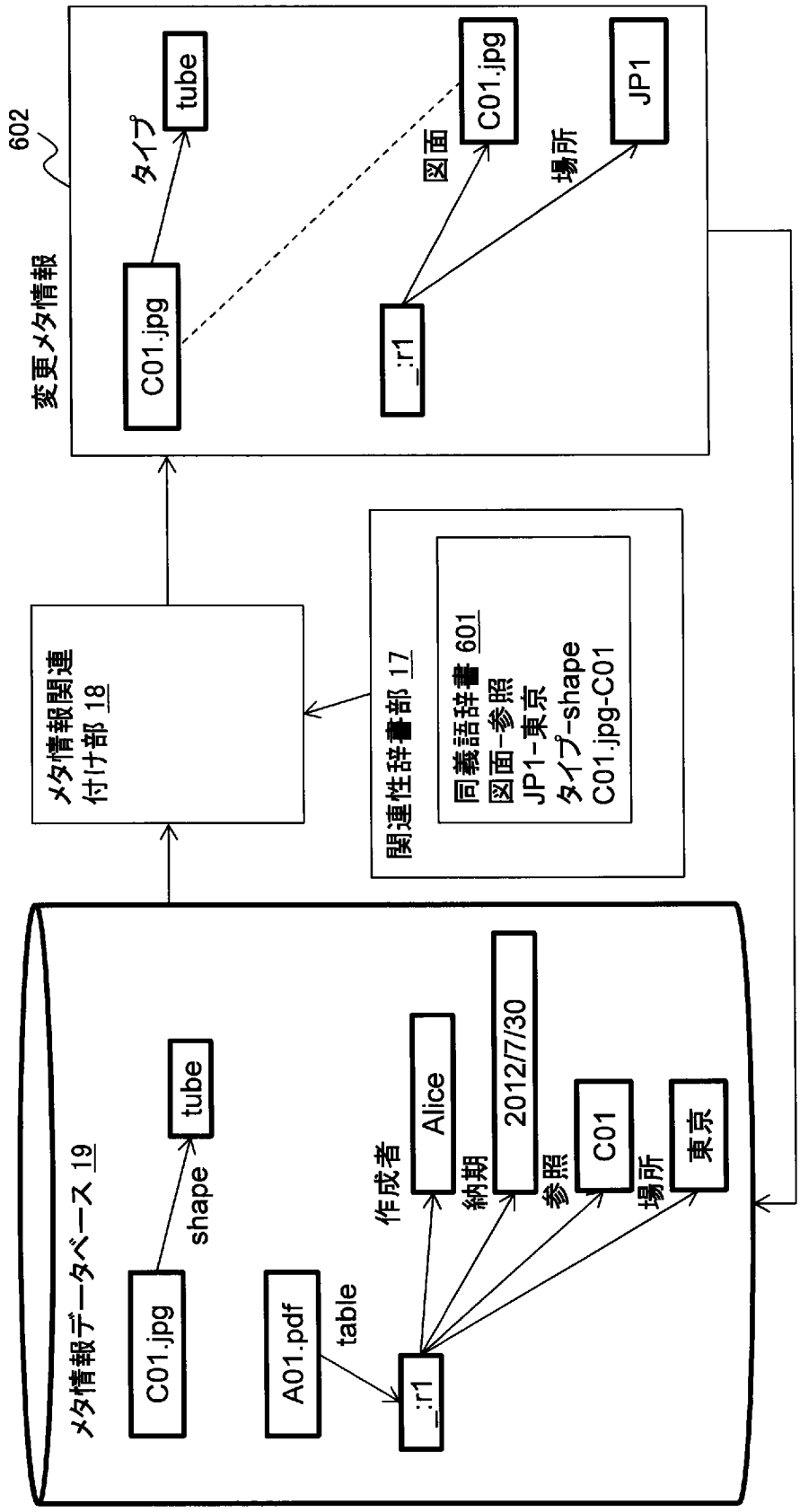
[図4]


404
画像メタ情報
C01.jpg
shape
tube
画像メタ情報抽出部 14
画像用辞書部 13
401
モデル
402
ラベル
shape:tube
shape:ring
403
画像データ
C01.jpg

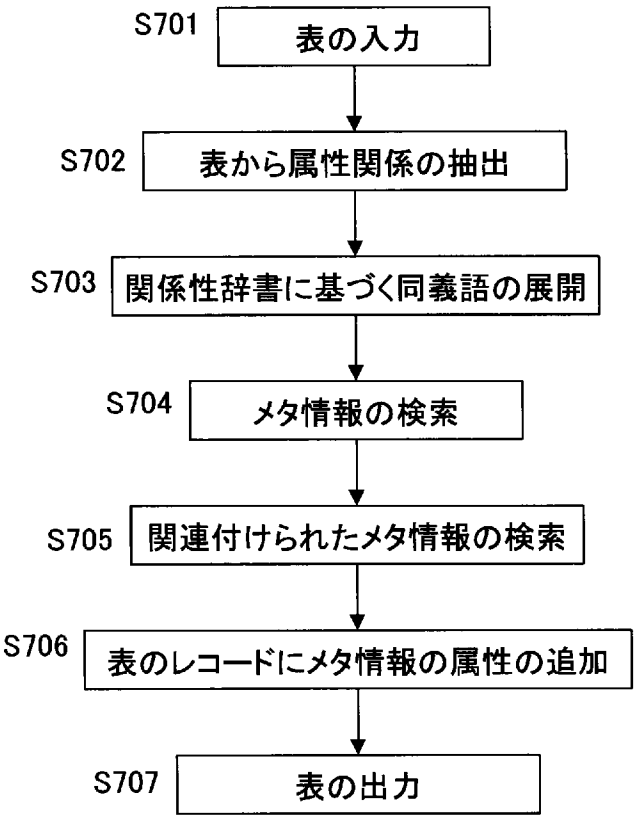
[図5]


504
参照
C01
場所
東京
作成者
Alice
納期
2012/7/30
文書メタ情報
A01.pdf
table
_t1
503
A01.pdf 文書データ
通知書 2012/2/14
****
参照:C01
場所: 東京
作成者:Alice
納期:2012/7/30
文書メタ情報抽出部 16
文書用辞書部 15
属性名リスト 501:参照, 場所, 作成者, 納期, ・・・
属性値リスト 502: C01, C02, 東京, Alice, Bob, 2012/7/30, ・・・

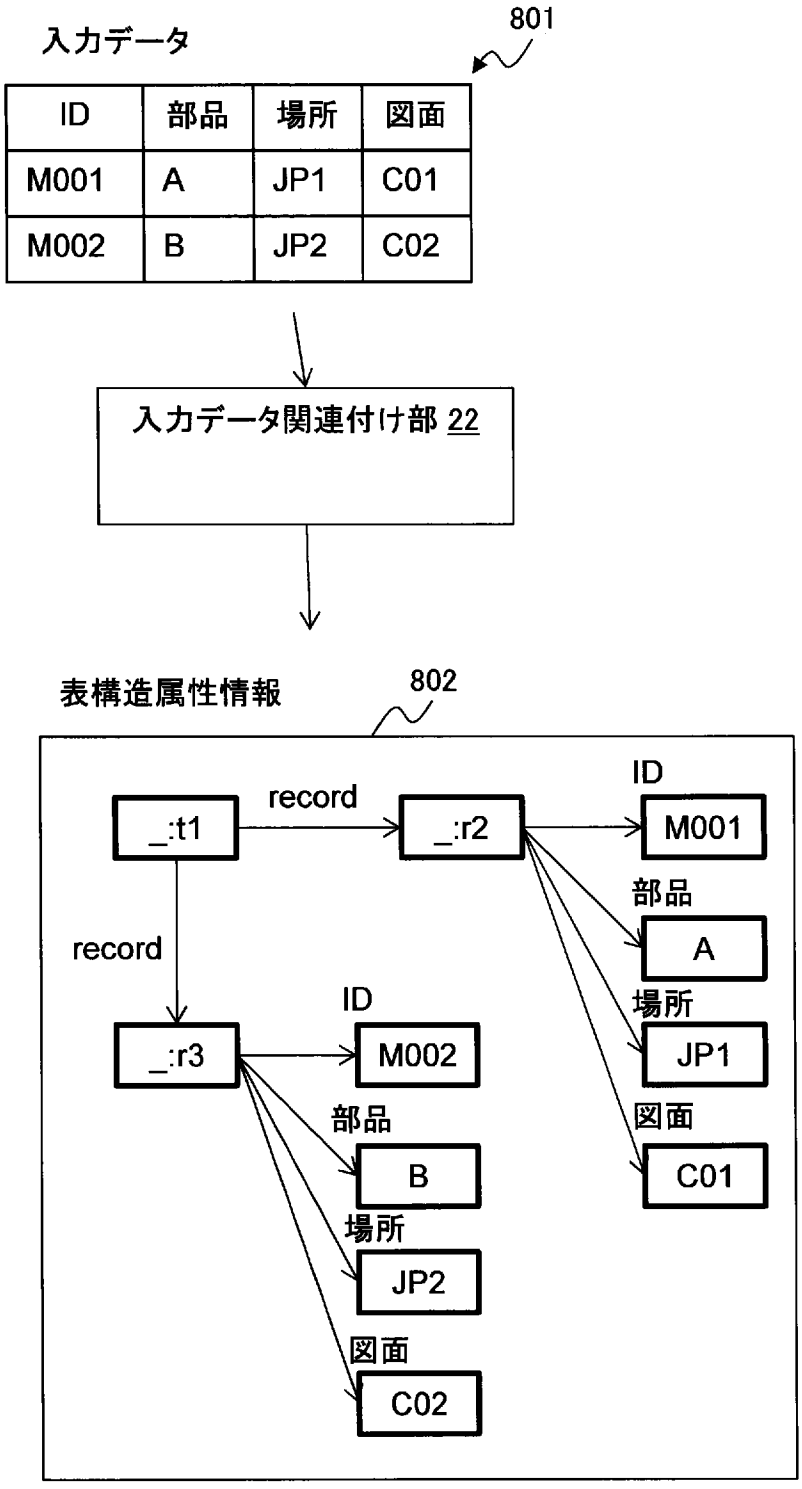
[図6]


メタ情報データベース 19
C01.jpg
shape
tube
A01.pdf
table
_:r1
作成者
Alice
納期
2012/7/30
参照
C01
場所
東京
メタ情報関連
付け部 18
関連性辞書部 17
同義語辞書 601
図面-参照
JP1-東京
タイプ-shape
C01.jpg-C01
602
変更メタ情報
C01.jpg
タイプ
tube
_:r1
図面
C01.jpg
場所
JP1

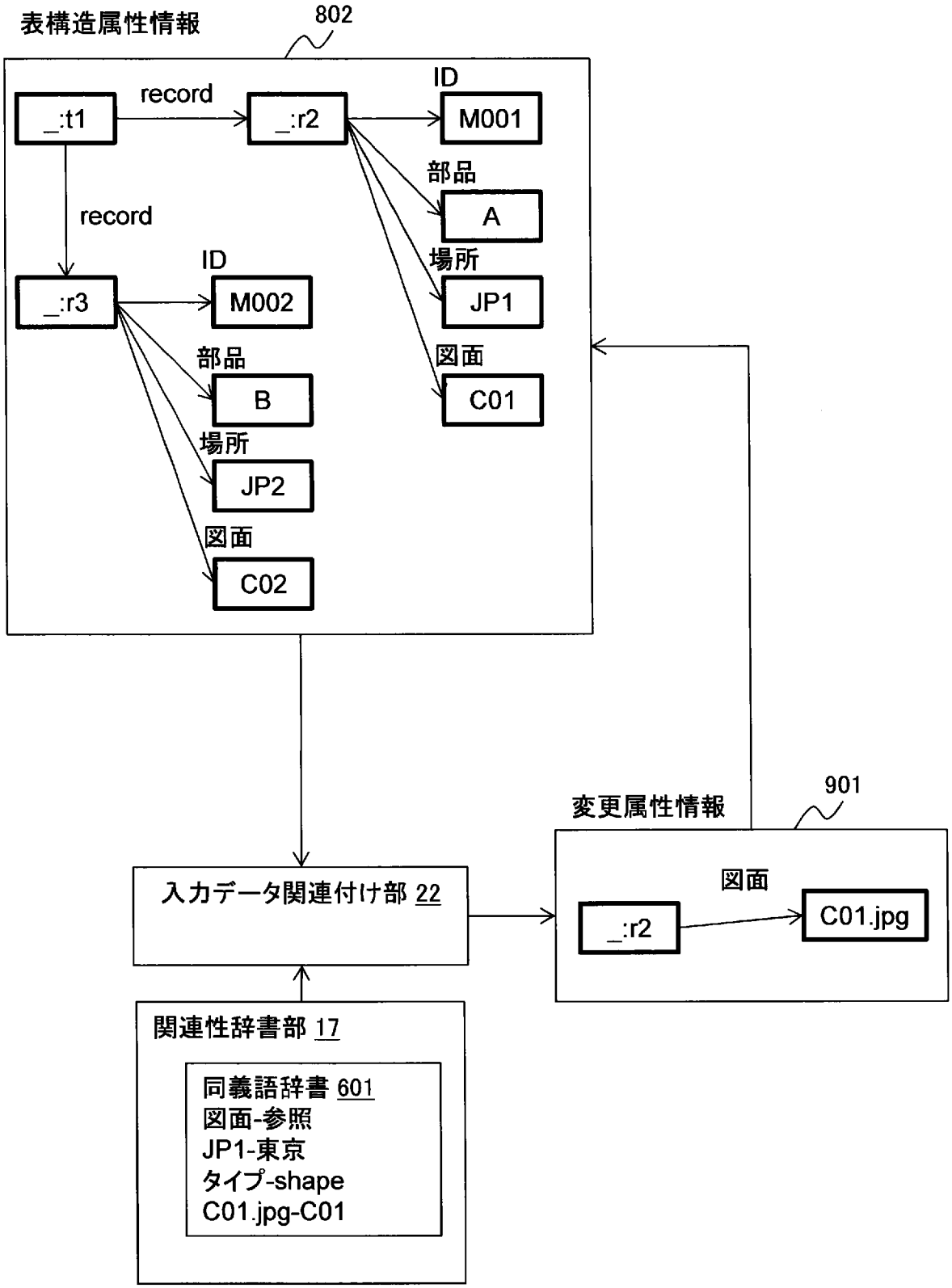
[図7]


S701
表の入力
S702
表から属性関係の抽出
S703
関係性辞書に基づく同義語の展開
S704
メタ情報の検索
S705
関連付けられたメタ情報の検索
S706
表のレコードにメタ情報の属性の追加
S707
表の出力

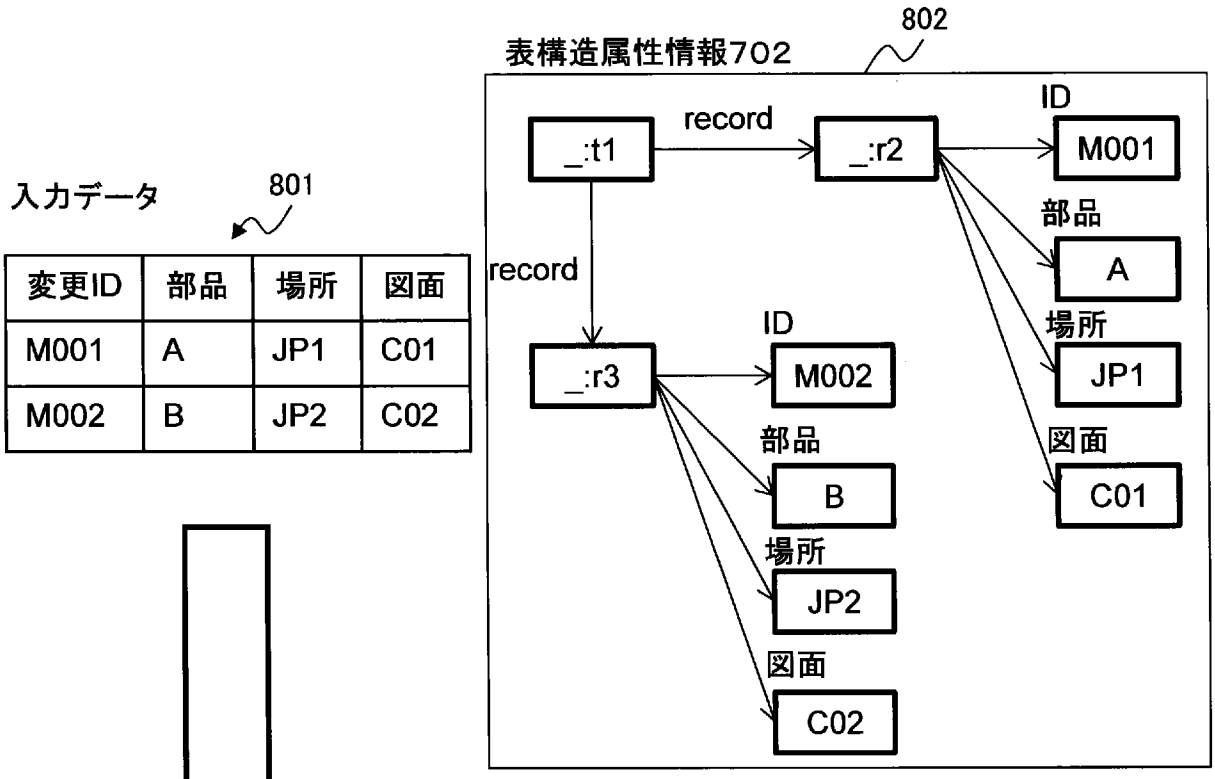
[図8]


入力データ
801
ID
部品
場所
図面
M001
A
JP1
C01
M002
B
JP2
C02
入力データ関連付け部 22
表構造属性情報
802
_:t1
record
_:r2
ID
M001
部品
A
場所
JP1
図面
C01
record
_:r3
ID
M002
部品
B
場所
JP2
図面
C02

[図9]


表構造属性情報
802
_:t1
record
_:r2
ID
M001
部品
A
場所
JP1
図面
C01
record
_:r3
ID
M002
部品
B
場所
JP2
図面
C02
入力データ関連付け部 22
変更属性情報
901
図面
_:r2
C01.jpg
関連性辞書部 17
同義語辞書 601
図面-参照
JP1-東京
タイプ-shape
C01.jpg-C01

[図10]

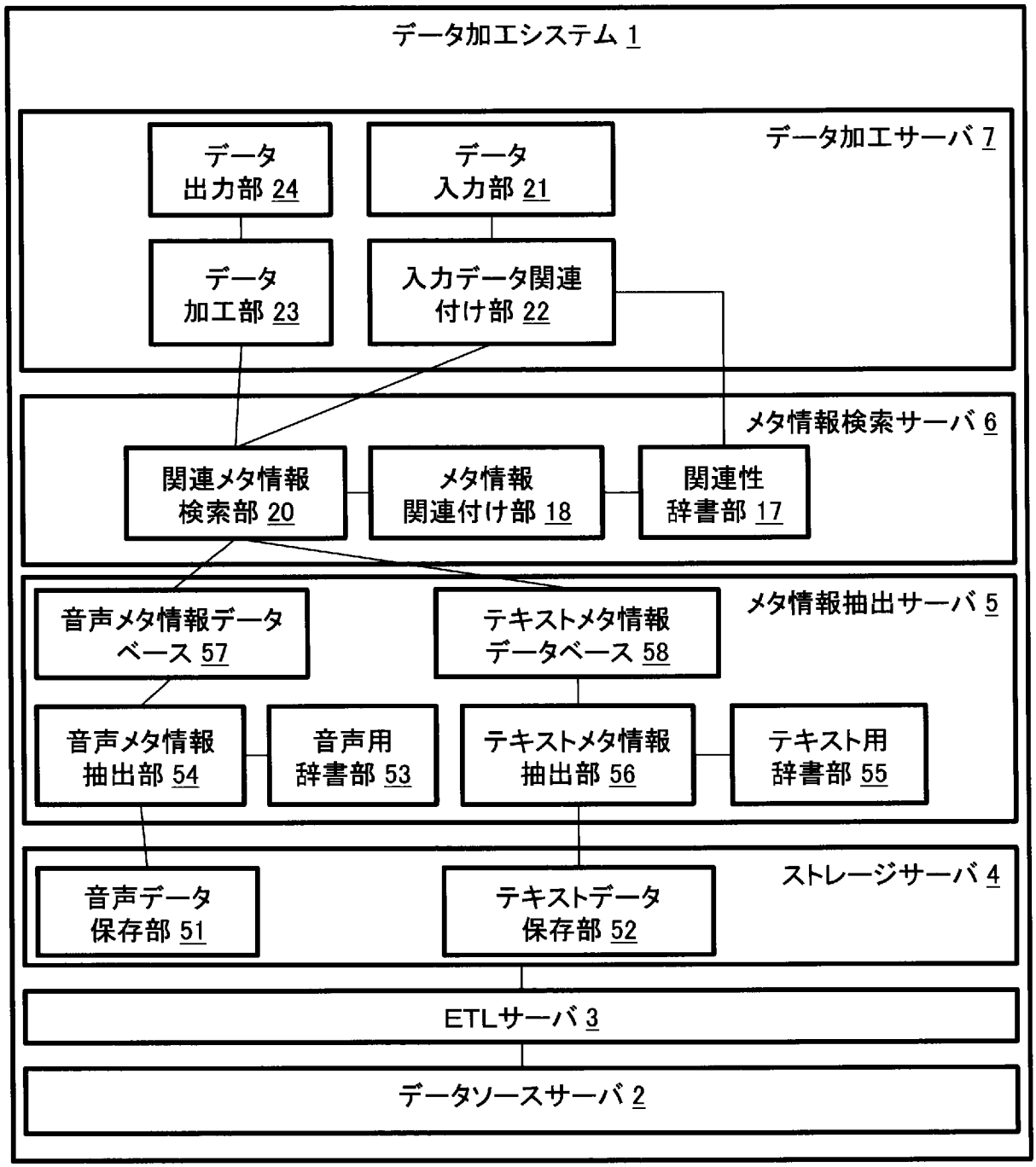


1003

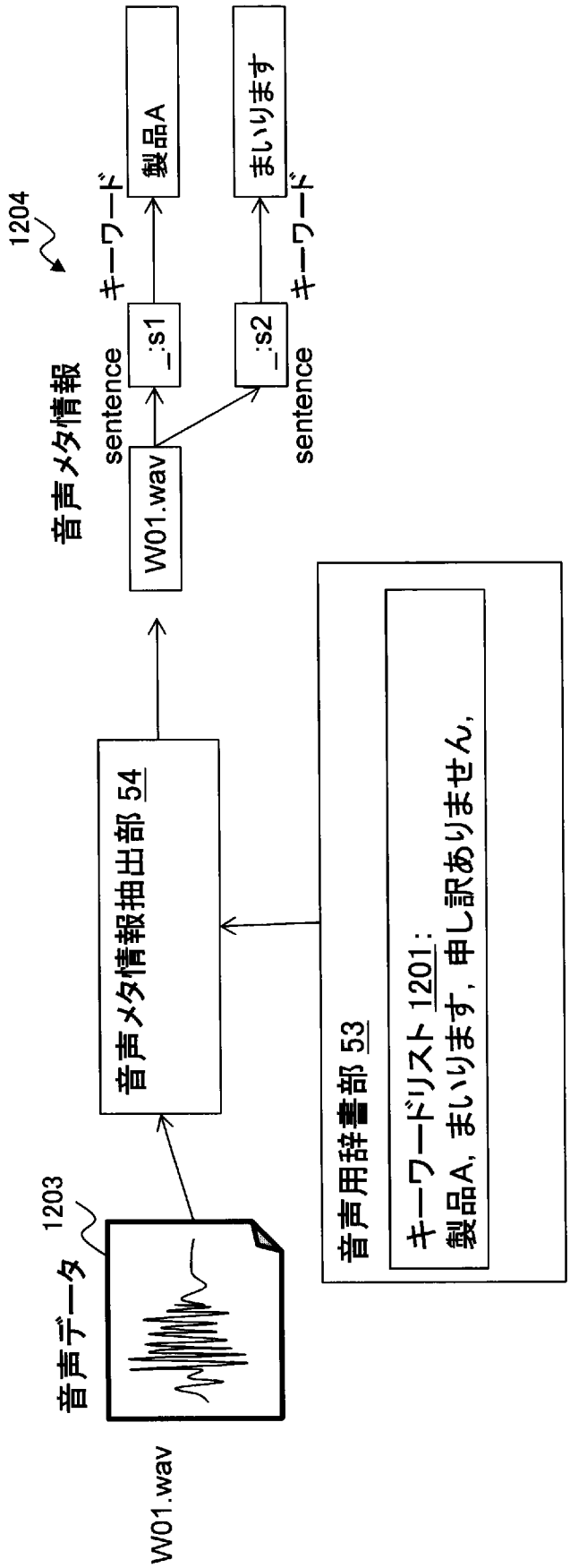
加工データ

変更ID	部品	場所	図面	作成者	納期	図面-タイプ
M001	A	JP1	C01	Alice	2012/7/30	tube
M002	B	JP2	C02			

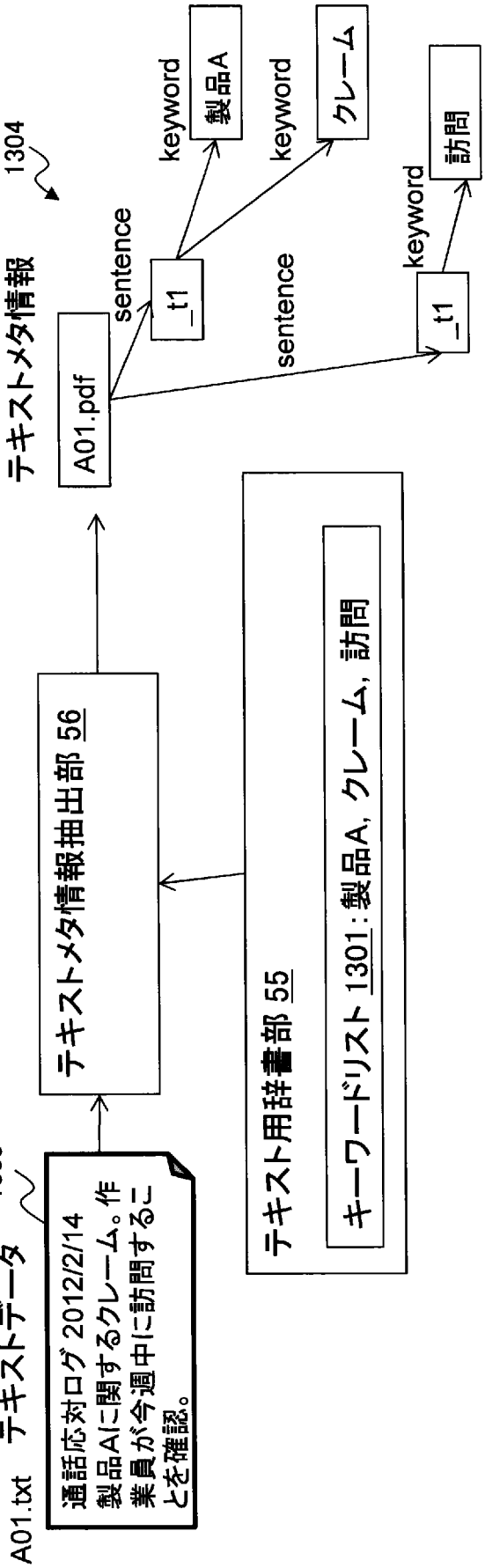
[図11]


データ加工システム 1
データ加工サーバ 7
データ出力部 24
データ入力部 21
データ加工部 23
入力データ関連付け部 22
メタ情報検索サーバ 6
関連メタ情報検索部 20
メタ情報関連付け部 18
関連性辞書部 17
メタ情報抽出サーバ 5
音声メタ情報データベース 57
テキストメタ情報データベース 58
音声メタ情報抽出部 54
音声用辞書部 53
テキストメタ情報抽出部 56
テキスト用辞書部 55
ストレージサーバ 4
音声データ保存部 51
テキストデータ保存部 52
ETLサーバ 3
データソースサーバ 2

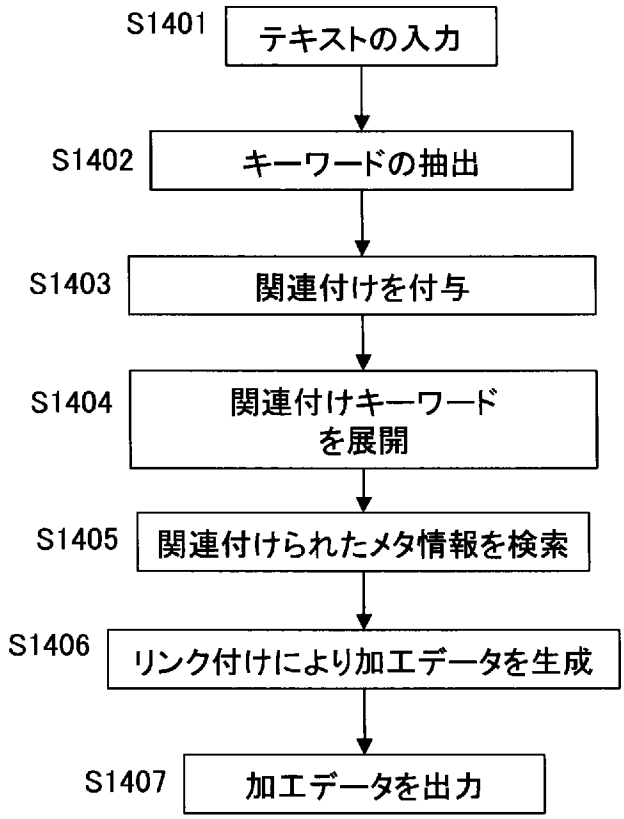
[図12]


1204
音声メタ情報
sentence
_:s1
キーワード
製品A
sentence
_:s2
キーワード
まいります
W01.wav
音声メタ情報抽出部 54
音声用辞書部 53
キーワードリスト 1201:
製品A, まいります, 申し訳ありません,
1203
音声データ
W01.wav

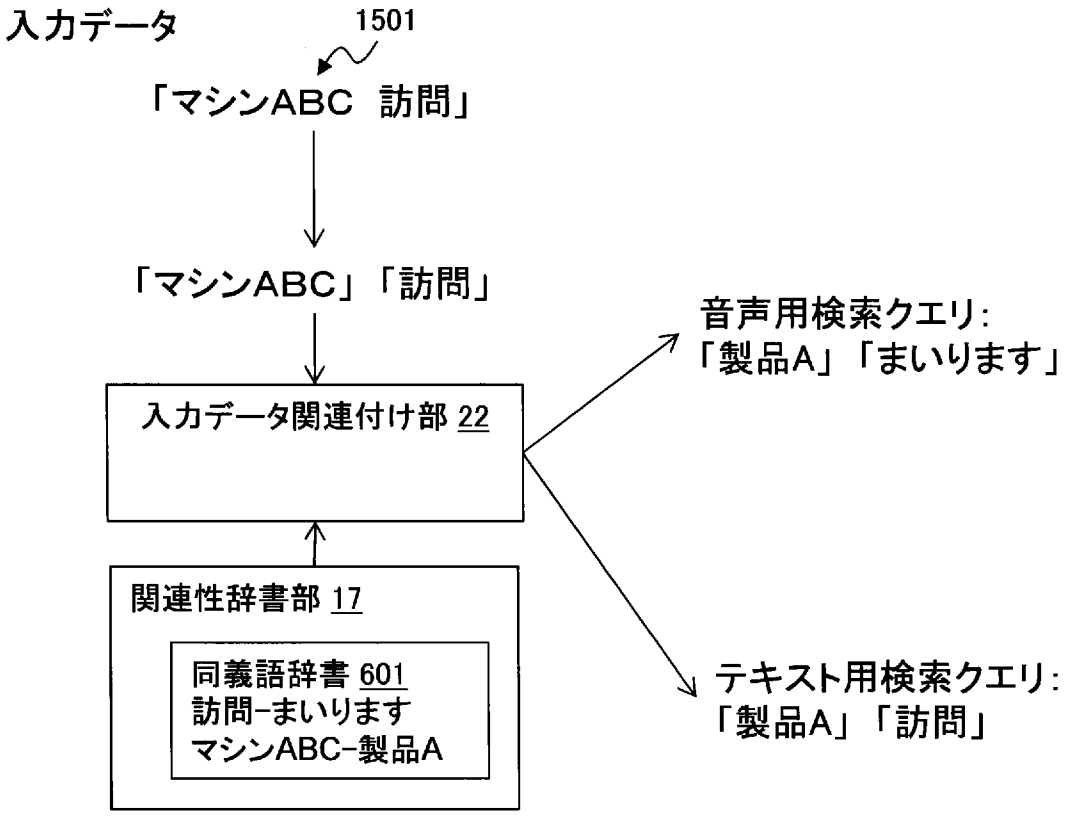
[図13]


A01.txt
テキストデータ
1303
通話応対ログ 2012/2/14
製品Aに関するクレーム。作業員が今週中に訪問することを確認。
テキストメタ情報抽出部 56
テキスト用辞書部 55
キーワードリスト 1301：製品A, クレーム, 訪問
テキストメタ情報
1304
A01.pdf
sentence
_t1
keyword
製品A
keyword
クレーム
sentence
_t1
keyword
訪問

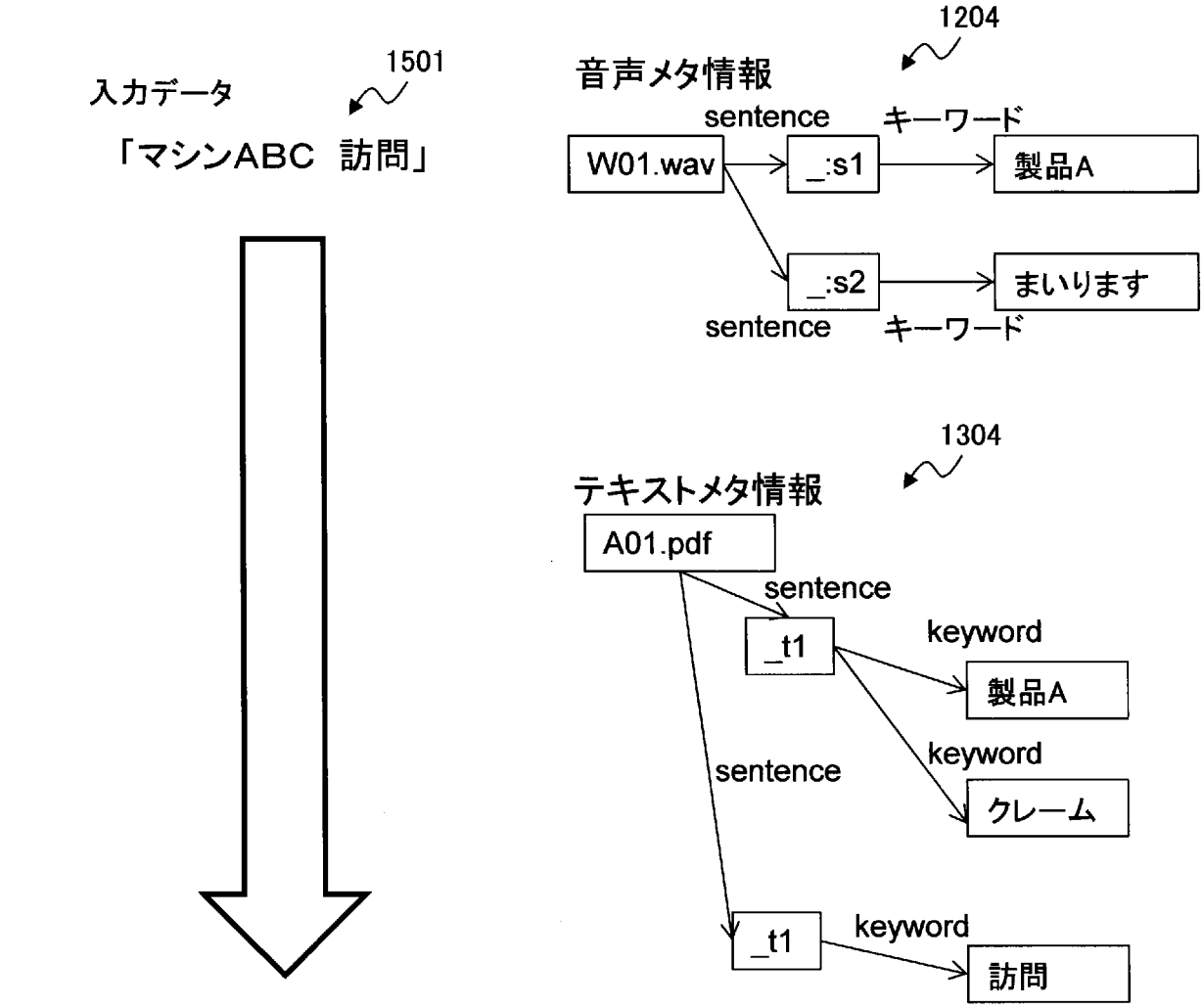
[図14]


S1401
テキストの入力
S1402
キーワードの抽出
S1403
関連付けを付与
S1404
関連付けキーワード
を展開
S1405
関連付けられたメタ情報を検索
S1406
リンク付けにより加工データを生成
S1407
加工データを出力

[図15]


入力データ
1501
「マシンABC 訪問」
「マシンABC」「訪問」
入力データ関連付け部 22
関連性辞書部 17
同義語辞書 601
訪問-まいります
マシンABC-製品A
音声用検索クエリ:
「製品A」「まいります」
テキスト用検索クエリ:
「製品A」「訪問」

[図16]



1601

加工データ

通話応対ログ 2012/2/14
製品A(listen W01.wav@s1)に関するクレーム。
作業員が今週中に訪問(listen W01.wav@s2)することを確認。

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2012/084007

A. CLASSIFICATION OF SUBJECT MATTER
G06F17/30(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F17/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2013
Kokai Jitsuyo Shinan Koho	1971-2013	Toroku Jitsuyo Shinan Koho	1994-2013

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
JSTPlus(JDreamII)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2008-134954 A (Canon Inc.), 12 June 2008 (12.06.2008), entire text; all drawings (Family: none)	1-13
A	JP 2008-192102 A (Sony Computer Entertainment Inc.), 21 August 2008 (21.08.2008), entire text; all drawings (Family: none)	1-13
A	JP 2008-226110 A (Seiko Epson Corp.), 25 September 2008 (25.09.2008), entire text; all drawings (Family: none)	1-13

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search 05 February, 2013 (05.02.13)	Date of mailing of the international search report 19 February, 2013 (19.02.13)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2012/084007

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2008-236373 A (Nippon Hoso Kyokai), 02 October 2008 (02.10.2008), entire text; all drawings (Family: none)	1-13

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１２／０８４００７

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F17/30(2006.01)i

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　G06F17/30

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０１３年
日本国実用新案登録公報	１９９６－２０１３年
日本国登録実用新案公報	１９９４－２０１３年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

JSTPlus(JDreamII)

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
A	JP 2008-134954 A（キヤノン株式会社）2008.06.12，全文、全図（ファミリーなし）	１－１３
A	JP 2008-192102 A（株式会社ソニー・コンピュータエンタイテイメント）2008.08.21，全文、全図（ファミリーなし）	１－１３
A	JP 2008-226110 A（セイコーエプソン株式会社）2008.09.25，全文、全図（ファミリーなし）	１－１３
A	JP 2008-236373 A（日本放送協会）2008.10.02，全文、全図（ファミリーなし）	１－１３

☐　Ｃ欄の続きにも文献が列挙されている。

☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの
「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「Ｏ」口頭による開示、使用、展示等に言及する文献
「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「＆」同一パテントファミリー文献

国際調査を完了した日 ０５．０２．２０１３	国際調査報告の発送日 １９．０２．２０１３
国際調査機関の名称及びあて先 日本国特許庁（ＩＳＡ／ＪＰ） 郵便番号１００－８９１５ 東京都千代田区霞が関三丁目４番３号	特許庁審査官（権限のある職員）　５Ｍ　９２８９ 野崎　大進 電話番号　０３－３５８１－１１０１　内線　３５９９