(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5238105号
(P5238105)
(45)発行日　平成25年7月17日(2013.7.17)　(24)登録日　平成25年4月5日(2013.4.5)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 340Z

請求項の数 4　(全 32 頁)

(21)出願番号　特願2007-549011 (P2007-549011)
(86)(22)出願日　平成17年12月9日(2005.12.9)
(86)国際出願番号　PCT/JP2005/022699
(87)国際公開番号　WO2007/066414
(87)国際公開日　平成19年6月14日(2007.6.14)
審査請求日　平成20年6月13日(2008.6.13)
審判番号　不服2011-16153(P2011-16153/J1)
審判請求日　平成23年7月26日(2011.7.26)

(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　100074099
弁理士　大菅　義之
(74)代理人　100133570
弁理士　▲徳▼永　民雄
(72)発明者　松浦　正卓
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　林　宏也
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】プログラム、及びデータ抽出方法

(57)【特許請求の範囲】
【請求項１】
取得可能なデータのなかから、指定された第一の抽出条件を満たすデータを抽出できるデータ抽出装置を実現させるためにコンピュータに実行させるプログラムであって、
前記データを取得する機能と、
前記第一の抽出条件を入力する機能と、
前記入力する機能により、二つ以上、入力された前記第一の抽出条件それぞれを、該第一の抽出条件それぞれに含まれる複数の部分条件に分割し、該分割によって得られる部分条件を複数の前記第一の抽出条件に共通して含まれる共通条件と該共通条件以外の非共通条件との組み合わせで表現する第二の抽出条件に前記第一の抽出条件それぞれを変換し、前記第二の抽出条件の前記部分条件それぞれを前記データの形式に対応する照合用オートマトンにそれぞれ変換し、前記データより前記照合用オートマトンを用いて前記部分条件単位で該部分条件を満たすデータをそれぞれ抽出し、該抽出したデータそれぞれから、前記第二の抽出条件それぞれを満たすデータを抽出することにより、前記第一の抽出条件を満たすデータを抽出する機能と、
を実現させるためのプログラム。
【請求項２】
請求項１記載のプログラムであって、
前記入力する機能は、前記第一の抽出条件それぞれと併せて、該第一の抽出条件それぞれと対応付けたデータの出力先に関する出力条件をそれぞれ入力することができ、

前記出力条件に従って、該出力条件と対応付けられた抽出条件を満たすデータを出力するプログラム。

【請求項３】

請求項１記載のプログラムであって、

前記抽出する機能により前記第二の抽出条件毎に抽出したデータはそれぞれ異なる出力先に出力するプログラム。

【請求項４】

取得可能なデータのなかから、指定された第一の抽出条件を満たすデータを抽出するためのデータ抽出方法において、

前記第一の抽出条件を二つ以上、取得した場合に、前記第一の抽出条件それぞれを、該第一の抽出条件それぞれに含まれる複数の部分条件に分割し、該分割によって得られる部分条件を複数の前記第一の抽出条件に共通して含まれる共通条件と該共通条件以外の非共通条件との組み合わせで表現する第二の抽出条件に前記第一の抽出条件それぞれを変換し、前記第二の抽出条件の前記部分条件それぞれを前記データの形式に対応する照合用オートマトンにそれぞれ変換し、前記データより前記照合用オートマトンを用いて前記部分条件毎に該部分条件を満たすデータをそれぞれ抽出して該抽出によって得たデータを記憶ユニットに記憶し、

前記抽出によって得たデータを、前記記憶ユニットから読み出し、該読み出したデータそれぞれから、前記第二の抽出条件それぞれを満たすデータを抽出することにより、前記第一の抽出条件を満たすデータを抽出する、

処理を情報処理装置に実行させることを特徴とするデータ抽出方法。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、取得可能なデータのなかから指定された抽出条件を満たすデータを抽出するための技術に関する。

【背景技術】

【０００２】

取得可能なデータのなかから任意のデータを抽出することができるデータ抽出装置は、現在、様々な用途に広く用いられている。インターネットで公開されている情報の検索では、検索エンジンとして用いられている。ユーザはそのデータ抽出装置を用いることにより、大量のデータのなかから所望のデータを迅速に得ることができる。

【０００３】

データ抽出装置は、予め定められた単位でデータを抽出する。その単位となるのは、例えばファイル、或いはレコードである。文書、及びインターネット上のＷｅｂページはファイルに相当する。顧客の利用実績ＰＯＳ（Point Of Sales）データやＨＨＴ（Hand Held Terminal）データなどはレコード単位で管理されるのが普通である。

【０００４】

図１は、従来のデータ抽出方法を説明する図である。ここで、図１を参照して、そのデータ抽出方法について具体的に説明する。

図１に示す従来のデータ抽出方法は、例えばクレジットカード会社で行われる場合のものである。表記した「ＪＯＵＲＮＡＬ」は、ファクトデータをレコード単位で格納したジャーナルファイルを表している。「ＭＡＳＴＥＲ」は、クレジットカードの所有者である顧客のデータをレコード単位で格納したマスタファイルを表している。それにより、図１に示すデータ抽出方法は、ＳＱＬ（Structured Query Language）を用いて、共に複数、存在するジャーナルファイル、及びマスタファイルのなかから所望のものを連結（ＪＯＩＮ）させ、その連結結果から所望のレコードを抽出する場合の例を表している。

【０００５】

連結させるジャーナルファイル、マスタファイルのそれぞれの条件は、ＦＲＯＭ句内のＷＨＥＲＥ句に記述されている。そこに記述された条件により、マスタファイルは現在の

ものが選択され、ジャーナルファイルは２００４年のものが選択される。そのＦＲＯＭ句内のＦＲＯＭ句には、ファイル間におけるレコードの対応関係はクレジットカードナンバーにより特定することが記述されている。連結結果から抽出されるレコードに格納されるデータの項目は、ＳＥＲＥＣＴ句に記述されている。そこに記述された項目は、顧客の指名（Ｖ．ＮＡＭＥ）、その年齢（Ｖ．ＡＧＥ）、利用回数（Ｖ．ＳＡＬＥＳ＿ＮＵＭ）、売上額（Ｖ．ＳＡＬＥＳ）である。連結結果から抽出するレコードの条件は、ＷＨＥＲＥ句に記述されている。そこに記述された条件は、カードの種類がコールドカード、というものである。このようなことから、２００４年に利用し、現在もゴールドカードを持つ顧客のレコードが検索結果として抽出される。

【０００６】

連結結果から抽出されるレコードを異ならせるには、ＷＨＥＲＥ句に記述する抽出条件を変更すれば良い。シルバーカードを持つ顧客のレコードを抽出させるのであれば、例えば図２に示すように、「ＧＯＬＤ」の記述を「ＳＩＬＶＥＲ」に変更すれば良い。それにより、２００４年に利用し、現在もシルバーカードを持つ顧客のレコードが検索結果として抽出される。

【０００７】

このように、従来のデータ抽出方法では、所望のデータを得るための抽出条件を決定し、その抽出条件毎に検索を行わせるようになっていた。このため、データを抽出する目的の数、つまり検索に使用する抽出条件の数が多くなるほど、全ての抽出結果を得るまでに要する時間が長くなり、効率的な作業が行えなくなるという問題点があった。

【０００８】

現在、デジタルデータで扱う情報の種類、及びその量は非常に増大しつつある。そのため、今後は従来のデータ抽出方法では対応するのが非常に困難となるのが予想される。このこともあって、膨大なデータのなかからでも必要な種類のデータを全てより迅速に得られるようにすることが重要であると考えられる。

【特許文献１】特開２００２－２２２１９４号公報
【特許文献２】特開２００５－７０９１１号公報
【特許文献３】特開平６－３１９９０６号公報

【発明の開示】

【０００９】

本発明は、膨大なデータのなかからでも必要な種類のデータを全てより迅速に得られるようにする技術を提供することを目的とする。

本発明の第１、及び第２の態様のプログラムは共に、取得可能なデータのなかから指定された抽出条件を満たすデータを抽出できるデータ抽出装置を実現させるためにコンピュータに実行させることを前提とし、それぞれ以下の機能を実現させる。

【００１０】

第１の態様のプログラムは、データを取得する機能と、抽出条件を入力する機能と、入力する機能により一つ以上、入力された抽出条件を用いて、該抽出条件毎にデータを抽出する機能と、抽出する機能により抽出条件毎に抽出されたデータをそれぞれ異なる出力先に出力する機能と、を実現させる。

【００１１】

第２の態様のプログラムは、データを取得する機能と、抽出条件を入力する機能と、入力する機能により入力された抽出条件を構成する条件式を複数の部分条件式に分割し、該分割によって得られる部分条件式の組み合わせで表現する形式に該抽出条件を変換して、該部分条件式単位で該部分条件式を満たすか否か確認することにより、取得する機能により取得したデータのなかで該抽出条件を満たすデータを抽出する機能と、を実現させる。

【００１２】

本発明のデータ抽出方法は、取得可能なデータのなかから指定された抽出条件を満たすデータを抽出するために適用されることが前提であり、対象となるデータが異なる抽出条件を複数、入力可能とさせ、抽出条件が１つ以上、入力された場合に、該抽出条件毎にデ

ータの抽出を行い、該抽出によって得たデータを、該データが満たす抽出条件に応じた出力先に出力する。

【００１３】

本発明では、対象となるデータが異なる抽出条件を複数、入力可能とさせ、抽出条件が１つ以上、入力された場合に、抽出条件毎にデータの抽出を行い、それによって得たデータを、そのデータが満たす抽出条件に応じた出力先にそれぞれ出力する。このため、ユーザは、複数の抽出条件を定義して入力することにより、１度に複数の抽出結果を得ることができる。それにより、必要な全ての抽出結果をより迅速に得ることができる。この結果、高い作業効率も容易に実現させることができる。

【００１４】

本発明では、入力された抽出条件は、それを構成する条件式を複数の部分条件式に分割し、その分割によって得られる部分条件式の組み合わせで表現する形式に変換して、部分条件式単位でその部分条件式を満たすか否か確認することにより、データのなかで抽出条件を満たすデータを抽出する。部分条件式の組み合わせで表現する形式に抽出条件を変換することにより、異なる条件式に同じ部分条件式が存在していても、条件式毎に部分条件式をデータが満たすか否かの確認を行う必要性を回避できるようになる。このため、より小さい負荷でデータ抽出を行えることとなる。

【図面の簡単な説明】

【００１５】

【図１】従来のデータ抽出方法を説明する図である。

【図２】従来のデータ抽出方法で異なる種類のデータを抽出させるための抽出条件の相違を説明する図である。

【図３】本実施の形態によるデータ抽出装置の昨日構成を説明する図である。

【図４】本実施の形態によるデータ抽出装置１００が可能なデータ抽出を説明する図である。

【図５】本実施の形態によるデータ集計装置を実現できるコンピュータのハードウェア構成の一例を示す図である。

【図６】ＸＭＬデータの構成例を説明する図である。

【図７】ＣＳＶデータの構成例を説明する図である。

【図８】抽出条件群の内容例を説明する図である。

【図９】タグＤＦＡ例を説明する図である。

【図１０】階層照合ＮＦＡ例を説明する図である。

【図１１】ＣＳＶ解析ＤＦＡ例を説明する図である。

【図１２】キーワードＤＦＡ例を説明する図である。

【図１３】論理テーブル例を説明する図である。

【図１４】出力バッファの管理方法を説明する図である。

【図１５】抽出条件入力部１１０が実行する処理のフローチャートである。

【図１６】データ入力構造検索部１２０が実行する処理のフローチャートである。

【図１７】抽出条件判定部１３０が実行する処理のフローチャートである。

【図１８】データ判定部１４０が実行する処理のフローチャートである。

【図１９】本実施の形態によるデータ抽出装置の適用例を説明する図である（その１）。

【図２０】本実施の形態によるデータ抽出装置の適用例を説明する図である（その２）。

【図２１】本実施の形態によるデータ抽出装置の適用例を説明する図である（その３）。

【図２２】本実施の形態によるデータ抽出装置の適用例を説明する図である（その４）。

【図２３】本実施の形態によるデータ抽出装置の適用例を説明する図である（その５）。

【図２４】本実施の形態によるデータ抽出装置の適用例を説明する図である（その６）。

【発明を実施するための最良の形態】

【００１６】

以下、本発明の実施の形態について、図面を参照しながら詳細に説明する。

図３は、本実施の形態によるデータ抽出装置の機能構成を説明する図である。

そのデータ抽出装置１００は、入力装置２１０からデータ２１１としてテキストデータを入力し、そのデータ２１１を指定された抽出条件群２２０により振り分けて出力するものとして実現されている。そのために、抽出条件入力部１１０、データ入力構造検索部１２０、抽出条件判定部１３０、データ判定部１４０、外部出力用の出力バッファ１５０、及びデータ出力部１６０を備えている。ここでは便宜的に、入力装置２１０から入力するデータ２１１として、図６に示すようなＸＭＬ（eXtensible Markup Language ）データ、及び図７に示すようなＣＳＶ（Comma Separated Values）データのみを想定する。それらのデータは共にテキストデータである。

【００１７】

抽出条件入力部１１０によって入力される抽出条件群２２０は、例えば図８に示すような内容のものである。その図８では、（１）～（３）に分けてそれぞれ抽出条件、及び出力条件を示している。そのように分けて示す抽出条件は全て、ユーザが所望のデータ２１１を抽出するためのものである。抽出条件と併せて示す出力条件は、その抽出条件によって抽出されるデータ２１１の出力先、及びそのファイル名を指定するものである。それにより、抽出条件群２２０は、所望のデータ２１１別に、そのデータ２１１が満たすべき抽出条件、及びその出力先ファイル名を指定するものとなっている。そのようにデータ２１１の出力先を任意に指定できるようにしたのは、データ２１１をより迅速に所望の形で利用するのを可能とさせるためである。以降、（１）に記述された抽出条件は「抽出条件１」と表記する。これは他でも同様である。

【００１８】

図４は、本実施の形態によるデータ抽出装置１００が可能なデータ抽出を説明する図である。ここで図４を参照して、そのデータ抽出について具体的に説明する。

図８に示す抽出条件群２２０は、データ２１１としてＸＭＬデータを想定したものである。図４では、ＣＳＶデータを想定した抽出条件群２２０を示している。「Ｑｕｅｒｙ」は抽出条件に相当し、「ＯｕｔＦｉｌｅ」は出力条件に相当する。Ｑｕｅｒｙ（抽出条件）として表記した「＄Ｘ」は、項目名「Ｘ」を表し、「＄＿」は任意の項目名を表している。それにより、例えばＱｕｅｒｙ１で表記した「＄Ｘ＝＝’Ｘ１’　ＯＲ　＄Ｘ＝＝’Ｘａ’」は、項目名「Ｘ」のデータがＸ１またはＸａであるデータ２１１が抽出の対象であることを示している。その表記が「＄＿＝＝’Ｘａ’」となっているＱｕｅｒｙでは、任意の項目のデータとしてＸａが存在しているデータ２１１が抽出の対象であることを示している。そのデータ２１１はＸＭＬデータ、及びＣＳＶデータの何れであっても、ファイルとしてまとめて入力させても良いが、一つずつ順次、入力させても良い。一つずつ入力させる場合、ＸＭＬデータでは図６に示すようなものとなり、ＣＳＶデータでは、図７において、先頭に「０００００１」～「０００００７」を表記した行のようなものとなる。ここでは便宜的に、それらのデータのまとまりをレコードと呼ぶことにする。また、２つの「’」の間に記述された文字列については「キーワード」と呼ぶことにする。そのキーワードは、図８に示す抽出条件群２２０では２つの「”」の間に記述された文字列が相当する。

【００１９】

本実施の形態では、文字列照合方式を用いて、抽出条件群２２０で指定された抽出条件の何れかを満たすデータ２１１を抽出し、満たす抽出条件に対応付けられた出力条件で指定された出力先ファイル名のファイルに出力する。それにより、Ｑｕｅｒｙ１を満たすデータ２１１はファイル名「ｒｅｓｕｌｔ１．ｃｓｖ」のファイル２３１として、Ｑｕｅｒｙ２を満たすデータ２１１はファイル名「ｒｅｓｕｌｔ２．ｃｓｖ」のファイル２３２として、Ｑｕｅｒｙ３を満たすデータ２１１はファイル名「ｒｅｓｕｌｔ３．ｃｓｖ」のファイル２３３として、それぞれ出力される。入力されたデータ２１１とファイル２３１～３の何れかに出力されるデータ２１１の対応関係は、図中に表記の（１）～（６）により示している。

【００２０】

各抽出条件はそれぞれ単独で考慮されるため、抽出条件は全て任意に定義することがで

きる。このため、ＸＭＬデータやＣＳＶデータなどのデータ２１１の種類毎に１つ以上の抽出条件を定義することもでき、また、その構造別に１つ以上の抽出条件を定義することもできるようになっている。従って、対象とするデータ２１１間でスキーマがどのように相違していても、その相違の影響は確実に回避させることができる。

【００２１】

上述したようなことから、抽出条件間は排他関係としなくとも良い。それにより、Ｑｕｅｒｙ１とＱｕｅｒｙ２では条件式（論理式）「＄Ｘ＝＝’Ｘａ’」を満たすデータ２１１をそれぞれ抽出する内容となっている。同様にＱｕｅｒｙ２とＱｕｅｒｙ３では条件式「＄Ｘ＝＝’Ｘｂ’」を満たすデータをそれぞれ抽出する内容となっている。この結果、ファイル２３１、２３２には共に（４）を表記したデータ２１１が出力され、ファイル２３２、２３３には共に（５）を表記したデータ２１１が出力されている。

【００２２】

このように、抽出条件群２２０により複数の抽出条件が指定されると、抽出条件毎にそれを満たすデータ２１１を振り分けて指定の出力先に出力するようになっている。このため、ユーザは、抽出条件群２２０として複数の抽出条件、及び出力条件を定義するだけで１度に複数の抽出結果を得ることができる。それにより、必要な全ての抽出結果はより迅速に得ることができる。この結果、高い作業効率も容易に実現させることができる。

【００２３】

上述したように、本実施の形態では文字列照合方式を採用している。その文字列照合方式は、抽出条件で指定した文字列と対象のデータ２１１との照合を、そのデータ２１１の先頭より後方に向かって逐次、行っていくことにより、その文字列がデータ２１１中に存在するか否かを調べるものである。その文字列照合方式では、先頭より後方に向かった走査を１回、行うだけで、抽出条件群２２０で定義された抽出条件の何れをデータ２１１が満たしているか確認することができる。そのため、定義された抽出条件の数に係わらず、常に迅速に抽出すべきデータ２１１を抽出することができる。その参考文献としては、例えば特許文献１、及び２が挙げられる。

【００２４】

図３の説明に戻る。

抽出条件入力部１１０は、上述したような抽出条件群２２０を入力し、抽出条件毎に、その抽出条件を解析して対応のオートマトンを生成する。それにより、抽出条件がＸＭＬデータ用のものであればタグＤＦＡ（Deterministic Finite state Automaton）１７０、階層照合ＮＦＡ（Non-deterministic Finite state Automaton）１７１、及びキーワードＤＦＡ１８０が生成される。抽出条件がＣＳＶデータ用のものであればＣＳＶ解析ＤＦＡ１７２、及びキーワードＤＦＡ１８０が生成される。論理テーブル１９０は、キーワードＤＦＡ１７２と同様に、抽出条件が想定するデータ２１１の種類に係わらず生成される。

【００２５】

抽出条件群２２０の作成は基本的に、ユーザによるデータ入力によって行われる。本実施の形態によるデータ抽出装置１００と接続された端末装置で抽出条件群２２０を作成する場合、例えばユーザは抽出条件群２２０作成用の画面を表示させ、その画面上に所望の内容の抽出条件群２２０を入力する。その入力後、データ抽出を指示すると、作成された抽出条件群２２０がデータ抽出装置１００に出力される。

【００２６】

上記論理テーブル１９０としては、抽出条件群２２０が図８に示す内容であった場合、抽出条件入力部１１０によって図１３に示すようなものが生成される。図１３に示すように、その論理テーブル１９０は、Ａ論理テーブル１９０ａ、及びＺ論理テーブル１９０ｂから構成されている。

【００２７】

Ａ論理テーブル１９０ａは、抽出条件を構成する条件式（論理式）を関係演算子（図８中では「＝」及び「＜」が相当）で分解して、その条件式が表現する論理により細分化し（図８では抽出条件２を構成する条件式「／ｒｏｏｔ／Ｃｏｍｐａｎｙ／ｃｏｄｅ　＜

９９」は「／ｒｏｏｔ／Ｃｏｍｐａｎｙ／ｃｏｄｅ」「＜　９９」に分解される）、細分化した条件式（部分条件式）毎に固有の論理番号を付した構成のものである。Ｚ論理テーブル１９０ｂは、条件式、或いは抽出条件を部分条件式、或いは条件式に付した論理番号の組み合わせで表現し、表現した組み合わせ毎に固有の論理番号を付した構成のものである。組み合わせる論理番号はＡ論理テーブル１９０ａ、及びＺ論理テーブル１９０ｂの何れのものであっても良い。その論理番号を用いて条件式、或いは抽出条件を表現することにより、Ａ論理テーブル１９０ａ、或いはＺ論理テーブル１９０ｂで参照すべきレコード（行）を特定できるようにさせている。特には図示していないが、そのＺ論理テーブル１９０ｂには、論理番号の組み合わせ毎に、その組み合わせで表現される条件式、或いは抽出条件が成立しているか否かを示す符号を格納できるようになっている。以降テーブル１９０ａ、及び１９０ｂでそれぞれ割り当てる論理番号を区別するために、Ａ論理テーブル１９０ａの論理番号には「Ａ」、Ｚ論理テーブル１９０ｂの論理には「Ｚ」をそれぞれ先頭に付して表記する。

【００２８】

Ｚ論理テーブル１９０ｂで論理番号Ｚ１が割り当てられた組み合わせは「Ａ１×Ａ２」である。その組み合わせ「Ａ１×Ａ２」は、論理番号Ａ１の部分条件式（／ｒｏｏｔ／ｏｒｉｇｉｎ）が成立し、且つ論理番号Ａ２の部分条件式（”ａｔｃｇ”）が成立するデータ２１１が抽出対象であることを表す形式の論理式となっている。それにより、組み合わせ（論理式）「Ａ１×Ａ２」中の「×」は、論理番号Ａ１、及びＡ２の部分条件式の論理積を行うことを示す論理演算子となっている。その論理式は、抽出条件１の内容を表している。同様に、論理番号Ｚ４、及びＺ５の各論理式はそれぞれ抽出条件３、及び２の内容を表している。抽出条件２はＺ５＝Ｚ２×Ｚ３になっている。ここで１９０ｂのテーブル内で、Ｚ２＝Ａ３×Ａ４によりＡ３＝／ｒｏｏｔ／Ｃｏｍｐａｎｙ／ｃｏｄｅ、Ａ４＝＜９９に対応する。

【００２９】

また、Ｚ３＝Ａ１×Ａ５により、Ａ１＝／ｒｏｏｔ／ｏｒｉｇｉｎ、Ａ５＝“ｇｔａｃ”に対応する。したがって、抽出条件２は、Ｚ論理番号Ｚ５と介して、Ａ論理番号Ａ３、Ａ４、Ａ１、Ａ５に対応し、図８で示す抽出条件２の論理積（ＡＮＤ）は、図１３で示す論理テーブルとその要素間のリンク状態で示される。図８の抽出条件３は図１３の抽出条件３、Ｚ論理番号４、Ａ論理番号Ａ１、Ａ６の論理テーブルとその要素間のリンクで示される。すなわち、抽出条件３はＺ４＝Ａ１×Ａ６（Ａ１＝／ｒｏｏｔ／ｏｒｉｇｉｎ、Ａ６＝“ａａｃｇ”）としてＡ論理番号に対応している。すなわち、このような論理番号によって各抽出条件で形成される論理テーブルを使って抽出条件毎のデータ判別が可能となる。

【００３０】

図１３に示す検索結果判定情報１９５は、抽出条件毎に、その抽出条件を表現する論理番号の組み合わせに対して付された論理番号、その抽出条件を満たすデータ２１１を格納すべき出力バッファ１５０を示す番号（図中「出力バッファＮｏ．」と表記）、及びファイルディスクリプタ（対応付けられた出力条件）がまとめられたものである。それにより、何れかの抽出条件を満たすデータ２１１は、検索結果判定情報１９５を参照して出力すべき出力バッファ１５０に出力された後、出力すべきファイルに出力される。

【００３１】

上記オートマトン（タグＤＦＡ１７０、階層照合ＮＦＡ１７１、キーワードＤＦＡ１８０、ＣＳＶ解析ＤＦＡ１７２）は検索条件中の文字列をデータ２１１と照合するための状態遷移テーブルである。状態間は遷移の方向を示す矢印で結んで表現される。先頭を初期状態とし、この初期状態からデータ２１１中の文字列に応じて順次、状態を遷移させる。遷移させる状態には、検索条件中の文字列の最後に位置する文字に相当する受理状態が１つ以上、含まれている。それによりオートマトンは、データ２１１中に検出すべき文字列が存在していれば、何れかの受理状態に遷移するように生成される。受理状態に遷移した場合、その受理状態に応じたヒット情報を出力するようになっている。そのヒット情報は

、遷移した受理状態に応じた特有のものであり、オートマトンの生成時に併せて生成される。

【００３２】

上記タグＤＦＡ１７０は、キーワードと照合すべき文字列（要素内容）が存在する要素までの検索パスを検出するためのものである。抽出条件群２２０が図８に示す内容であった場合、抽出条件入力部１１０によって図９に示すようなタグＤＦＡ１７０が最終的に生成される。図８に示す抽出条件群２２０では、検索パスとして「／ｒｏｏｔ／ｏｒｉｇｉｎ」及び「／ｒｏｏｔ／Ｃｏｍｐａｎｙ／ｃｏｄｅ」が存在することから、それぞれがタグ名である文字列「ｒｏｏｔ」「ｏｒｉｇｉｎ」「Ｃｏｍｐａｎｙ」及び「ｃｏｄｅ」をそれぞれ検出できるように生成されている。それらの文字列の最後に位置する文字「ｔ」「ｎ」「ｙ」及び「ｅ」の何れかに相当する受理状態まで遷移することで、その文字に対応する文字列が検出されたことを示すヒット情報１７０ａ～ｄの何れかが出力される。

【００３３】

階層照合ＮＦＡ１７１は、現在、対象とする検索パスを管理するためのものである。抽出条件群２２０が図８に示す内容であった場合、抽出条件入力部１１０によって図１０に示すような階層照合ＮＦＡ１７１が最終的に生成される。そのＮＦＡ１７１は、図１０に示すように、何れかの検索パスに記述されたタグ名を単位とした状態遷移が行われるように生成されている。このため、その状態遷移は開始タグ、及び終了タグによって発生する。ここでは、「４」、及び「２」を表記した状態が受理状態に相当する。

【００３４】

「４」を表記した受理状態に遷移したことは、検索パス「／ｒｏｏｔ／Ｃｏｍｐａｎｙ／ｃｏｄｅ」が検出されたことを意味する。それにより、その検索パスで指定されたノードでは、その値が９９未満か否か、つまり論理番号Ａ４の部分条件式（論理）が成立するか否かの照合を行うためのヒット情報１７１ａが出力される。そのヒット情報１７１ａは、照合の対象となる部分条件式を示す論理番号（ここではＡ４）、検索パスの階層の深さを示す階層情報、及びその部分条件式で関係を確認すべき内容を示す比較情報（ここでは＜９９）を含むものである。同様に「２」を表記した受理状態に遷移したことは、検索パス「／ｒｏｏｔ／ｏｒｉｇｉｎ」が検出されたことを意味するから、その検索パスで指定されたノード、つまりタグ名「ｏｒｉｇｉｎ」のタグでは、その文字列が「ａｔｃｇ」「ｇｔａｃ」或いは「ａａｃｇ」の何れと一致するか否かの照合を行うためのヒット情報１７１ｂ－ｄが出力される。それらのヒット情報１７１ｂ－ｄで比較情報を示していないのは、それらに表記した論理番号に対応する部分条件式の照合はキーワードＤＦＡ１８０により行うためである。

【００３５】

階層照合ＮＦＡ１７１における状態遷移は、図９に示すタグＤＦＡ１７０を用いて行われる。例えばタグ名である文字列「ｒｏｏｔ」をタグＤＦＡ１７０により検出すると、つまりタグＤＦＡ１７０によりヒット情報１７０ａを出力すると、ＮＦＡ１７１では「０」を表記した初期状態から「１」を表記した状態に遷移する。次にタグＤＦＡ１７０により文字列「ｏｒｉｇｉｎ」を検出すると、ＮＦＡ１７１では「１」を表記した状態から「２」を表記した状態に遷移する。このとき、タグＤＦＡ１７０により文字列「Ｃｏｍｐａｎｙ」を検出すると、ＮＦＡ１７１では「１」を表記した状態から「３」を表記した状態に遷移する。それらの何れの文字列もタグＤＦＡ１７０により検出できなければ、ＮＦＡ１７１では「１」を表記した状態から「０」を表記した初期状態に遷移する。そのように遷移させることにより、階層照合ＮＦＡ１７１を用いて検索パスに沿った階層の移動の有無を把握し、対象とする検索パスを管理する。

【００３６】

ＣＳＶ解析ＤＦＡ１７２は、キーワードと照合すべき文字列（要素内容）が存在する要素までの検索パスを検出するためのものである。その要素が２つのダブルコーテーション間に存在するＣＳＶデータ（図７）では、抽出条件入力部１１０によって図１１に示すようなＣＳＶ解析ＤＦＡ１７２が生成される。図１１中に表記した「０ｘ」はそれに続くシ

ンボルが１６進数表現であることを表している。
【００３７】
キーワードＤＦＡ１８０は、抽出条件により指定されたキーワードと一致する文字列をデータ２１１中から検出するためのものである。抽出条件群２２０が図８に示す内容であった場合、抽出条件入力部１１０によって図１２に示すようなキーワードＤＦＡ１８０が最終的に生成される。それに登録された何れかのキーワードの最後に位置する文字に相当する受理状態まで遷移した場合、つまり文字列「ａａｃｇ」「ａｃｇｔ」及び「ｇｔａｃ」の何れかを検出できた場合、検出された文字列に応じてヒット情報１８０ａ～ｃの何れかが出力される。
【００３８】
データ入力構造検索部１２０は、入力装置２１０から所定量ずつ連続的にデータ２１１を入力し、そのデータ２１１の種類に応じて、照合に用いるオートマトンを決定する。それにより、データ２１１がＸＭＬデータであれば、タグＤＦＡ１７０、及び階層照合ＮＦＡ１７１を用いて抽出条件の何れかに記述された検索パスの検出を行う。データ２１１がＣＳＶデータであれば、ＣＳＶ解析ＤＦＡ１７２を用いて抽出条件の何れかに記述された項目名の検出を行う。検索パス、或いは項目名を検出すると、その検索パスによって指定されたノード、或いはその項目名のセルが開始する位置を示すデータ位置情報、及び検出された文字列を示すノード・セル情報を抽出条件判定部１３０に通知する。それらの情報は例えばヒット情報として生成するものか、或いはそれを含むものである。それらの情報の通知は、データ２１１の終端を検出するまで、検索パス、或いは項目名を検出する度に行う。その終端の検出は、ＸＭＬデータではルートタグと組になる終了タグの検出に相当し、ＣＳＶデータでは所定個数のセルの検出に相当する。データ入力構造検索部１２０による検索パス、或いは項目名の検出は、Ａ論理テーブル１９０ａに格納された部分条件式が成立することの確認に相当する。
【００３９】
抽出条件判定部１３０は、データ入力構造検索部１２０から通知されたデータ位置情報が示すデータ位置より、キーワードＤＦＡ１８０を用いた照合を行う。その照合の結果、そのデータ位置から何れかのキーワードと一致する文字列、或いは関係演算子が示す関係を満たす値（図８に示す抽出条件群２２０では９９未満の値）が存在することを確認すると、Ｚ論理テーブル１９０ｂの該当論理番号の箇所にそのことを示す符号（以降「真符号」と表記し、それと異なる符号を「偽符号」と表記する）を格納する。その確認ができる前にデータ２１１の終端を検出した場合には、その終端の位置を示すデータ位置情報をデータ入力構造検索部１２０に通知する。それにより、構造検索部１２０は、データ２１１の終端を自身が検出したか否かに係わらず、その終端まで走査が終了したことをデータ判定部１４０に通知する。
【００４０】
抽出条件判定部１３０は、上記通知を行うか、或いは構造検索部１２０が終端を検出するまで、構造検索部１２０から情報が通知される度にキーワードＤＦＡ１８０を用いた照合を行う。この結果、データ２１１が抽出条件２を満たしている場合には、論理番号Ｚ２、及びＺ３の符号として真符号が順次、格納され、最後に論理番号Ｚ５の符号として真符号が格納されることになる。そのようにして、対象とするデータ２１１が論理式を満たす論理番号の箇所にのみ真符号が格納されることから、Ｚ論理テーブル１９０ｂを参照することにより、データ２１１が満たす抽出条件を確認できるようになっている。
【００４１】
このようにして本実施の形態では、抽出条件を構成する条件式をそれが表現する論理により細分化し、その細分化によって得られた部分条件式（細分化論理）単位での照合を行うようにしている。それにより、一致する文字列、或いは検索パスの検出、関係演算子で表す関係の確認、及びそのようなことを行うべき箇所の特定、などをそれぞれ個別に実施している。そのようにすると、より柔軟に対応することが可能となり、データ２１１の種類やその構造などの情報がたとえ不足していたとしても、ユーザは得られている情報から

所望のデータ２１１が満たす内容を抽出条件としてより容易に定義できるようになる。このため、ユーザにとっての高い利便性が実現される。

【００４２】

部分条件式（細分化論理）は、同じ、或いは他の抽出条件で別に存在する場合がある。図８に示す例では、部分条件式「／ｒｏｏｔ／ｏｒｉｇｉｎ」は抽出条件１～３の何れにも記述されている。しかし、そのような複数の同じ記述は、条件式を細分化することにより、一つの部分条件式として残せば済むようになる。それにより、抽出条件の数や内容に係わらず、成立するか否か確認すべき部分条件式は必要最小限に抑えることができる。条件式、或いは抽出条件は複数の部分条件式の組み合わせで表現される。このため、それらが成立するか否かはより迅速に行えることとなる。

【００４３】

データ判定部１４０は、Ｚ論理テーブル１９０ｂを参照して、データ２１１が満たす抽出条件を確認する。その確認により、何れかの抽出条件を満たしていることが判明すると、検索結果判定情報１９５（図１３）を参照して、出力すべき出力バッファ１５０にデータ２１１を出力して格納する。

【００４４】

図１４は、出力バッファの管理方法を説明する図である。

データ２１１を対応する出力バッファ１５０への出力は、出力バッファ情報１５１、及びバッファ情報１５２により管理している。出力バッファ情報１５１は、抽出条件群２２０により確保した出力バッファ１５０の数を示す取得バッファ数情報、及びバッファ情報１５２にアクセスするためのポインタ情報を備えている。そのバッファ情報１５２は、取得バッファ数情報が示す数のレコードを備えたものであり、各レコードには、対応する出力バッファ１５０（ここでは出力バッファ１５０ａ～ｃのうちの一つ）に関する複数の情報を有する個別バッファ情報１５３（ここでは個別バッファ情報１５３ａ～ｃのうちの一つ）がそれぞれ格納されている。それら出力バッファ情報１５１、及びバッファ情報１５２を格納するエリアは出力バッファ１５０と共に、データ抽出装置１００に搭載、或いは接続された記憶装置１４０１上に確保されている。タグＤＦＡ１７０、階層照合ＮＦＡ１７１、ＣＳＶ解析ＤＦＡ１７２、キーワードＤＦＡ１８０、及び論理テーブル１９０も例えばその記憶装置１４０１に格納される。

【００４５】

その個別バッファ情報１５３は、対応する出力バッファ１５０にアクセスするためのポインタ情報、そのデータ２１１を格納可能な全サイズを表す全バッファサイズ、そのサイズのなかでデータ２１１を格納可能な残りのサイズを表す残バッファサイズ、確保した出力バッファ１５０自体のサイズを表す出力バッファサイズ、を有している。各レコードに付した番号の大小関係は抽出条件の番号のそれと同じとさせている。つまり、レコード番号０のレコードは抽出条件１に対応している。それにより、データ２１１が満たす抽出条件に対応するレコードを特定できるようにさせている。

【００４６】

上述したようなことから、データ判定部１４０は、Ｚ論理テーブル１９０ｂを参照してデータ２１１が満たす抽出条件が存在していることを確認すると、検索結果判定情報１９５を参照してその抽出条件を確認し、出力バッファ情報１５１、及びバッファ情報１５２を参照する。それにより、確認した抽出条件に対応するレコードをバッファ情報１５２から取り出し、そのレコードに格納された個別バッファ情報１５３により指定される出力バッファ１５０にデータ２１１を出力する。残バッファサイズは、出力するデータ２１１のサイズにより更新する。

【００４７】

データ出力部１６０は、各出力バッファ１５０の例えば残バッファサイズを監視し、そのサイズが所定値以下になるか、或いは入力装置２１０から入力して処理するデータ２１１が無くなった場合に、検索結果判定情報１９５を参照して、出力バッファ１５０に格納されているデータ２１１を対応するファイルに出力する。それにより、出力条件で指定さ

れた出力先ファイル名のファイルに、これまでに抽出したデータ２１１を保存する。ここでは、３つのファイル２３１～２３３は共に同じ出力装置２３０上に保存させている。

【００４８】

図５は、データ抽出装置１００を実現できるコンピュータのハードウェア構成の一例を示す図である。抽出装置１００は複数のコンピュータ（データ処理装置）により実現させても良いが、ここでは図５に構成を示す１台のコンピュータによって実現されていることを前提として説明することとする。

【００４９】

図５に示すコンピュータは、ＣＰＵ５１、メモリ５２、入力装置５３、出力装置５４、外部記憶装置５５、媒体駆動装置５６、及びネットワーク接続装置５７を有し、これらがバス５８によって互いに接続された構成となっている。同図に示す構成は一例であり、これに限定されるものではない。

【００５０】

メモリ５２は、データを一時的に格納するＲＡＭ等のメモリである。外部記憶装置５５、若しくは媒体駆動装置５６がアクセスする可搬記録媒体ＭＤに記憶されているプログラム、あるいはデータが一時的に格納される。ＣＰＵ５１は、プログラムをメモリ５２に読み出して実行することにより、全体の制御を行う。そのプログラムは、ネットワーク接続装置５７によりネットワークを介して取得したものであっても良い。

【００５１】

入力装置５３は、例えば、キーボード、マウス等の入力機器と接続されているか、或いはそれらを有するものである。そのような入力機器に対するユーザの操作を検出し、その検出結果をＣＰＵ５１に通知する。

【００５２】

出力装置５４は、例えばディスプレイと接続されているか、或いはそれを有するものである。ＣＰＵ５１の制御によって送られてくるデータをディスプレイ上に出力させる。

ネットワーク接続装置５７は、例えばイントラネットやインターネット等のネットワークを介して、他の装置と通信を行うためのものである。外部記憶装置５５は、例えばハードディスク装置である。主に各種データやプログラムの保存に用いられる。

【００５３】

記憶媒体駆動装置５６は、フレキシブル・ディスク、光ディスク（ここではＣＤ－ＲＯＭ、ＣＤ－Ｒ、及びＤＶＤ等を含む）、或いは光磁気ディスク等の可搬型の記録媒体ＭＤにアクセスするものである。

【００５４】

図３に示す出力装置２３０は、図５に示す構成では外部記憶装置５５、記録媒体ＭＤが装着された媒体駆動装置５６、或いはネットワーク接続装置５７によりアクセス可能な外部装置に相当する。入力装置２１０は、記録媒体ＭＤが装着された媒体駆動装置５６、或いはネットワーク接続装置５７によりアクセス可能な外部装置に相当する。抽出条件群２２０の入力は、入力装置５３、記録媒体ＭＤが装着された媒体駆動装置５６、或いはネットワーク接続装置５７により行うことができる。図１４に示す記憶装置１４０１は、例えば外部記憶装置５５、及びメモリ５２の少なくとも一方に相当する。

【００５５】

検索条件入力部１１０は、例えば出力装置５４を除く各部５１～５３、及び５５～５８によって実現される。データ入力構造検索部１２０、及びデータ出力部１６０は共に、例えば入力装置５３、及び出力装置５４を除く各部５１、５２、及び５５～５７によって実現される。抽出条件判定部１３０、及びデータ判定部１４０は共に、例えば入力装置５３、出力装置５４、及びネットワーク接続装置５７を除く各部５１、５２、５５、５６、及び５８によって実現される。

【００５６】

次に、上述した各部１１０、１２０、１３０、及び１４０の動作について、図１５～図１８に示す各処理のフローチャートを参照して詳細に説明する。それらの処理は何れも、

例えばＣＰＵ５１が、外部記憶装置５５、若しくは媒体駆動装置５６に装着された可搬記録媒体ＭＤに記憶されているプログラムをメモリ５２に読み出して実行することにより実現される。

【００５７】

図１５は、抽出条件入力部１１０が実行する処理のフローチャートである。始めに図１５を参照して、その処理について詳細に説明する。その処理は、例えば抽出条件群２２０の入力をユーザが入力装置５３、或いはネットワークを介して指示することで起動される。その場合、抽出条件群２２０は入力装置５３、或いはネットワーク接続装置５７を介して入力される。

【００５８】

先ず、ステップ１１では、抽出条件群２２０を入力し、例えばメモリ５２に保存する。続くステップ１２では、保存した抽出条件群２２０のなかから１抽出条件を選択して読み出し、それを解析して対応するオートマトンの種類を特定する。その次に移行するステップ１３では、特定した種類のオートマトンを生成、或いは更新する。その生成、或いは更新により、抽出条件に記述された文字列が必要に応じてタグＤＦＡ１７０、階層照合ＮＦＡ１７１、或いはキーワードＤＦＡ１８０に登録される。

【００５９】

ステップ１３に続くステップ１４では、抽出条件群２２０のなかに選択していない他の抽出条件が有るか否か判定する。そのような抽出条件が残っていた場合、判定はＹＥＳとなって上記ステップ１２に戻り、他の選択条件を選択する。そうでない場合には、判定はＮＯとなり、ステップ１５で論理テーブル１９０の生成と併せて検索結果判定情報１９５（図１３）、出力バッファ情報１５１、及びバッファ情報１５２の生成を行い、抽出条件数に応じた出力バッファ１５０（図１４）の確保を行った後、一連の処理を終了する。このようにして、抽出条件群２２０の入力により、必要なオートマトンの生成に併せて、データ２１１を出力すべき出力先に出力するための準備が行われる。

【００６０】

図１６は、データ入力構造検索部１２０が実行する処理のフローチャートである。次に図１６を参照して、その処理について詳細に説明する。その処理は、例えばデータ２１１の入力装置２１０からの取り込みが指示されている間、実行される。

【００６１】

先ず、ステップ２１では、入力装置２１０から入力すべきデータ２１１が有るか否か判定する。そのようなデータ２１１が無かった場合、判定はＮＯとなり、再度、その判定を行う。それにより、そのデータ２１１が生じるのを待つ。一方、そうでない場合には、判定はＹＥＳとなってステップ２２に移行する。

【００６２】

ステップ２２では、入力装置２１０から所定量のデータ２１１を入力する。続くステップ２３では、入力したデータ２１１から一つを選択し、抽出条件入力部１１０によって決定したオートマトンを用いて、それに登録された文字列の何れかと一致する文字列の検索を行う。

【００６３】

その検索は１文字単位で行い、その検索が終了するとステップ２４に移行して、対象となる文字列（検索パス、項目名、など）を検出できたか否か判定する。そのような文字列を検出できなかった場合、判定はＮＯとなってステップ２７に移行する。そうでない場合には、判定はＹＥＳとなってステップ２５に移行する。

【００６４】

ステップ２５では、データ位置情報等を抽出条件判定部１３０に通知する。その通知により、抽出条件判定部１３はキーワードＤＦＡ１８０を用いた照合を行い、その照合によってデータ２１１の終端を検出すると、そのデータ位置情報を通知する。このことから、次のステップ２６では、その通知が有ったか否か判定する。その通知が有った場合、判定はＹＥＳとなってステップ２８に移行する。そうでない場合には、判定はＮＯとなって上

記ステップ２３に戻り、検索を続行する。

【００６５】

上記ステップ２４の判定がＮＯとなって移行するステップ２７では、検索によってデータ２１１の終端を検出したか否か判定する。その終端を検出した場合、判定はＹＥＳとなってステップ２８に移行する。そうでない場合には、判定はＮＯとなって上記ステップ２３に戻り、検索を続行する。

【００６６】

ステップ２８では、データ２１１の終端が検出されたことをデータ判定部１４０に通知する。続くステップ２９では、入力したデータ２１１のなかで未選択のデータ２１１が有るか否か判定する。未選択のデータ２１１が存在する場合、判定はＹＥＳとなって上記ステップ２３に戻り、未選択のデータ２１１を選択して検索を開始する。そうでない場合には、判定はＮＯとなって上記ステップ２１に戻る。それにより、入力装置２１０に入力すべきデータ２１１が有るか否かの確認を行う。

【００６７】

図１７は、抽出条件判定部１３０が実行する処理のフローチャートである。次に図１７を参照して、その処理について詳細に説明する。

先ず、ステップ４１では、レコードの終了通知が通知されるのを待つ。その通知を受け取ると、判定がＮＯとなってステップ４２に移行し、通知されたデータ位置情報、及びキーワードＤＦＡ１８０を用いた照合を行う。その次に移行するステップ４３では、キーワードＤＦＡ１８０に登録されたキーワードの何れかと一致する文字列をデータ２１１から検出できたか否か判定する。そのような文字列を検出できた場合、判定はＹＥＳとなり、ステップ４４で論理テーブル１９０（Ｚ論理テーブル１９０ｂ）の該当論理番号の箇所に真符号を設定した後、上記ステップ４１に戻り、通知待ちの状態に移行する。そうでない場合には、判定はＮＯとなってステップ４５に移行する。

【００６８】

ステップ４５では、データ２１１の終端を検出したか否か判定する。照合によってその終端を検出した場合、判定はＹＥＳとなり、そのことを通知するためにデータ位置情報をデータ入力構造検索部１２０にステップ４６で通知した後、上記ステップ４１に戻る。そうでない場合には、判定はＮＯとなって上記ステップ４２に戻り、照合を続行する。

【００６９】

上述したようにして、データ入力構造検索部１２０と抽出条件判定部１３０の間では必要な情報のやりとりが随時、行われ、その情報によってそれぞれ処理を進行させる。それにより、１データ２１１毎に、それが成立する抽出条件を確認し、その確認結果に応じた処理を行うようになっている。

【００７０】

図１８は、データ判定部１４０が実行する処理のフローチャートである。最後に図１８を参照して、その処理について詳細に説明する。

先ず、ステップ５１では、データ入力構造検索部１２０からデータ２１１の終端が通知されるのを待つ。その通知を受け取ると、判定がＮＯとなってステップ５２に移行し、論理テーブル１９０を参照して、現在、対象としているデータ２１１が満たす抽出条件を判定する。その後はステップ５３に移行する。

【００７１】

ステップ５３では、データ２１１が満たす抽出条件が有るか否か判定する。そのような抽出条件が存在した場合、判定はＹＥＳとなってステップ５４に移行し、検索結果判定情報１９５（図１３）、出力バッファ情報１５１、及びバッファ情報１５２（図１４）を参照してデータ２１１を出力すべき出力バッファ１５０に出力し、対応する個別バッファ情報１５３を更新した後、上記ステップ５１に戻る。それにより、通知待ちの状態に移行する。一方、そうでない場合には、判定はＮＯとなってそのステップ５１に戻る。

【００７２】

図１９～図２４は、上記データ抽出装置の適用例を説明する図である。以降は、図１９

～図２４を参照して、その適用可能な利用法について具体的に説明する。図１９～図２４において、データ抽出装置は「抽出器」と表記している。

【００７３】

図１９は、複数のデータ抽出装置１００を多段階で使用する場合の例を示している。データ１９０３を入力するデータ抽出装置１００は、そのデータ１９０３を２つの連結器１９１０に振り分けている。その二つの連結器１９１０の一方は、マスタファイル１９０１のデータをデータ１９０３と連結させて別のデータ抽出装置１００に出力し、そのデータ抽出装置１００は連結結果を２つの集計器１９２０に振り分けている。その２つの集計器１９２０はそれぞれ異なるデータ抽出装置１００に集計結果を出力し、その集計結果を入力するデータ抽出装置１００はそのデータをそれぞれ３つのファイルに振り分けて出力している。これらは、二つの連結器１９１０の他方側でも同様である。

【００７４】

図２０は、入力データの振り分けにデータ抽出装置１００を使用する場合の例を示している。その入力データは、ジャーナルファイル２０００に格納された各レコードのデータである。データ抽出装置１００は、抽出条件を満たすデータをジャーナルファイル２００１～３のうちの何れかに振り分けて出力するために用いられている。そのように振り分けるのは、例えばマスタＸ～Ｚとの連結条件がそれぞれ異なることに対応するためである。そのように振り分けると、データを３系統で並行して処理することが可能となることから、処理の高速化を実現できる。

【００７５】

図２１は、連結結果のデータの振り分けにデータ抽出装置１００を使用する場合の例を示している。その連結結果は、マスタとジャーナルのデータを連結させたものである。データ抽出装置１００は、抽出条件１～３の何れかを満たすデータを、その抽出条件に応じてファイル２１０１～３のうちの何れかに出力するために用いられている。

【００７６】

図２２は、集計結果のデータの振り分けにデータ抽出装置１００を使用する場合の例を示している。その集計結果は、マスタとジャーナルのデータの連結結果に対して集計操作を行ったものである。データ抽出装置１００は、抽出条件１～３の何れかを満たす集計結果のデータを、その抽出条件に応じてファイル２２０１～３のうちの何れかに出力するために用いられている。

【００７７】

図２３は、新聞社等で実施されるクリッピングサービスの提供用にデータ抽出装置１００を使用する場合の例を示している。その場合、データ抽出装置１００にはサービス登録者毎に、その登録者に送るべき記事データが満たす抽出条件を定義する。その抽出装置１００には随時、記事データが入力され、その記事データが満たす抽出条件に応じて対応するファイルに出力される。そのファイルに出力された記事データは、定期的にサービス登録者に配信される。サービス登録者の追加、削除、或いは要求の変更などは、抽出条件の追加、削除、或いは内容の変更によって対応することができる。

【００７８】

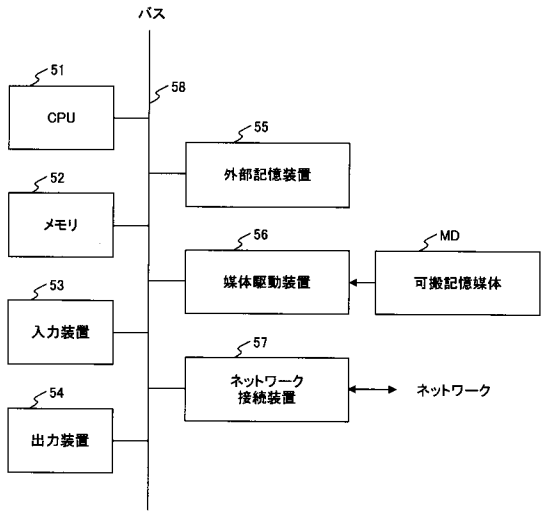
図２４は、ハイウェイ利用調査システムにデータ抽出装置１００を使用する場合の例を示している。その場合、ハイウェイのモニタシステムから随時、データがデータ抽出装置１００に入力される。その抽出装置１００には、必要なデータのみを抽出するための抽出条件を定義する。それにより、抽出装置１００は、抽出条件に従ってデータを選別する（フィルタリングする）。選別されたデータは、連結器によりマスタデータと照合され、より詳細なデータに展開される。例では、自動車の番号が「ｋ　２１０４」のデータに対して会社名「○○通運」が付加されている。マスタデータと照合されたデータは集計器により、例えば会社毎に集計されて出力される。

【００７９】

なお、本実施の形態では、抽出条件によって出力先を振り分けるデータそのものを外部から入力しているが、そのデータは実際に振り分けるデータの生成用、或いは特定用のも

のであっても良い。つまり符号化された圧縮データのようなものであっても良い。そのようなデータの入力は、記録媒体ＭＤに記録して行うようにしても良い。

【図５】



【図６】

211

```
<root>
    <Company>
        <Code>000001</Code>
        <Name>ABCカンパニー</Name>
    </Company>
    <Office>
        <ZipCode>211-0040</ZipCode>
        <Address>神奈川県川崎市上小田中...</Address>
        <Tel>444-999-0001</Tel>
        <Tel>444-999-0002</Tel>
        <Tel>444-999-0003</Tel>
        <Fax>444-999-9999</Fax>
    </Office>
</root>
```

【図７】

211

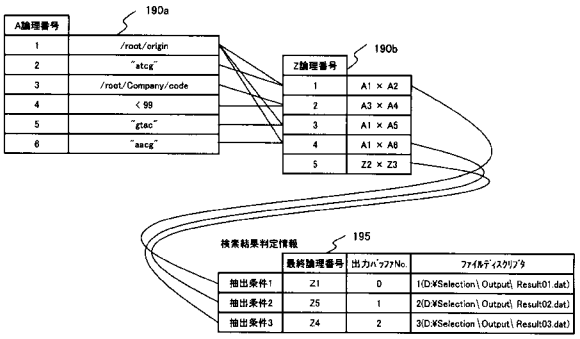
```
CompanyCode, CompanyName,ZipCode,Address,Tel,Fax
000001,"AAA株式会社","211-0000","神奈川県川崎市...","444-999-0001","444-999-9999"
000002,"BBB商会","164-0000","東京都中野区...","111-999-0001","111-999-9999"
000003,"有限会社CCC","230-0000","神奈川県横浜市...","222-999-0001","222-999-9999"
000004,"劇団DDD","811-0000","福岡県福岡市...","333-999-0001","333-999-9999"
000005,"EEE社","410-0000","静岡県沼津市...","555-999-0001","555-999-9999"
000006,"FFFカンパニー","211-0000","神奈川県川崎市...","444-999-0001","444-999-9990"
000007,"GGG放送","105-0000","東京都港区...","666-999-0001","666-999-9999"
.
.
.
.
.
.
.
.
.
```

【図８】

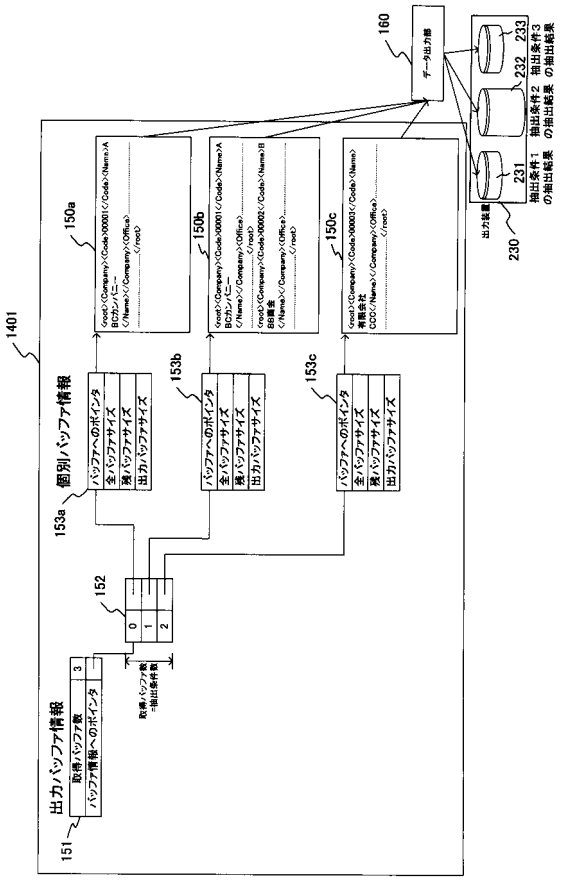
220

```
(1)抽出条件： /root/origin = "atcg"
  出力条件： "D:¥Selection\ Output\ Result01.dat"

(2)抽出条件： /root/Company/code < 99  AND /root/origin = "gtac"
  出力条件： "D:¥Selection\ Output\ Result02.dat"

(3)抽出条件： /root/origin = "aacg"
  出力条件： "D:¥Selection\ Output\ Result03.dat"
```

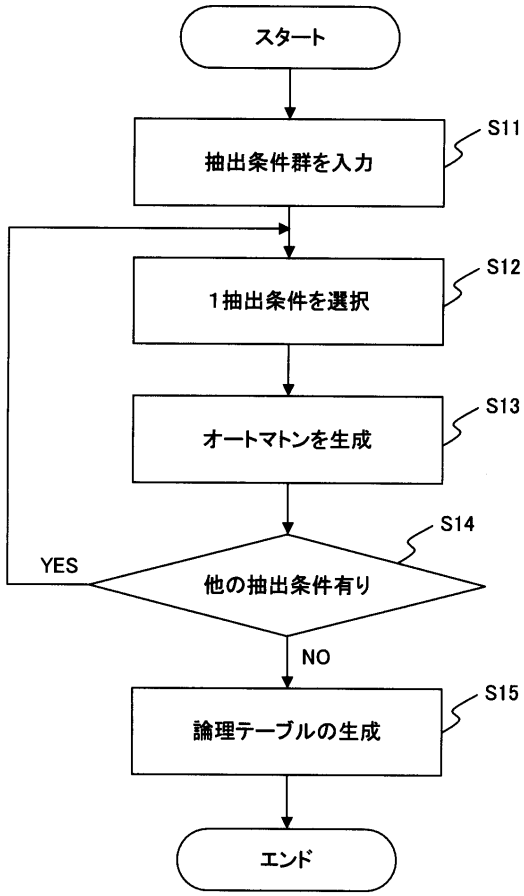
【図１３】



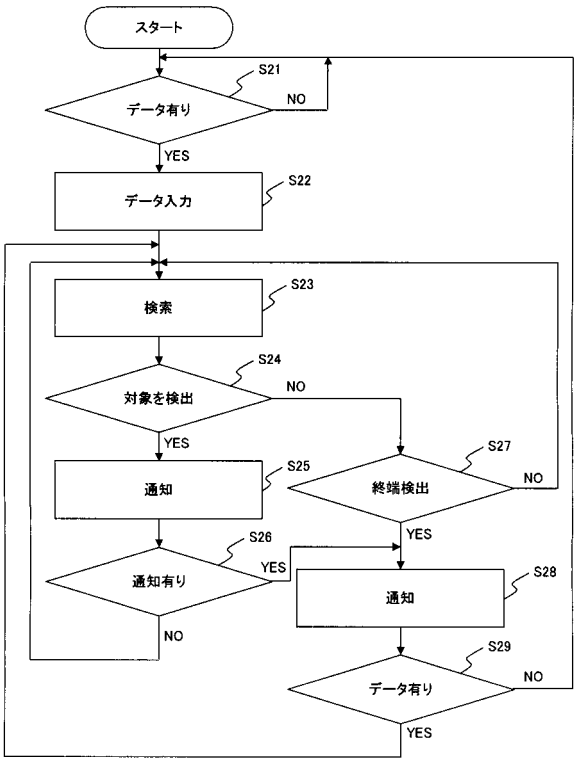
【図１４】



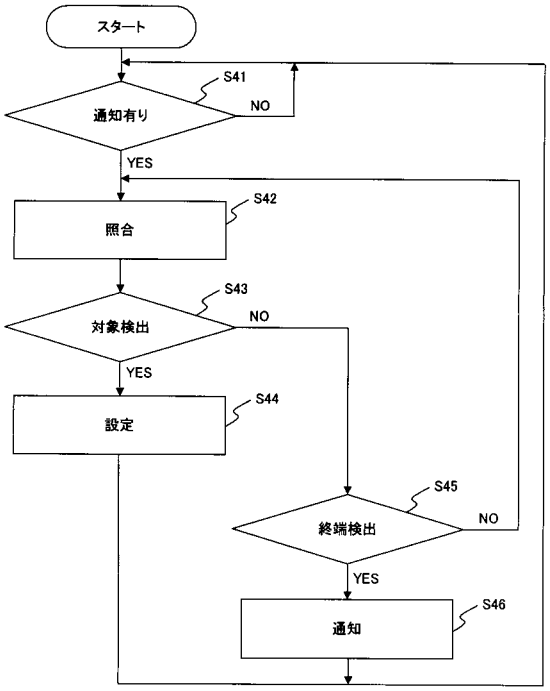
【図１５】



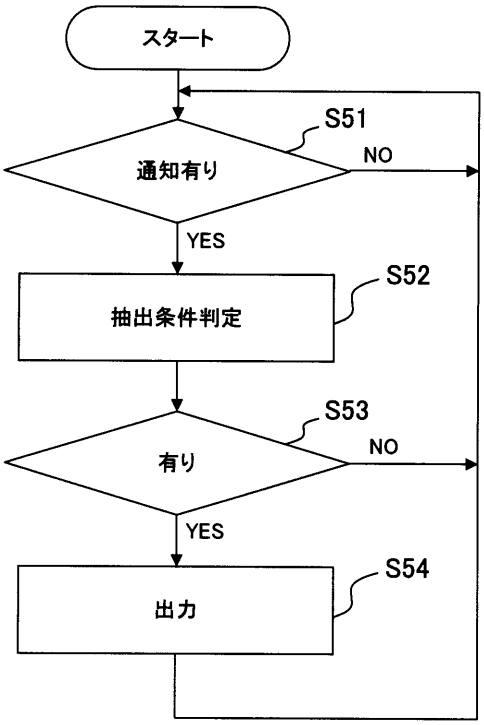
【図１６】



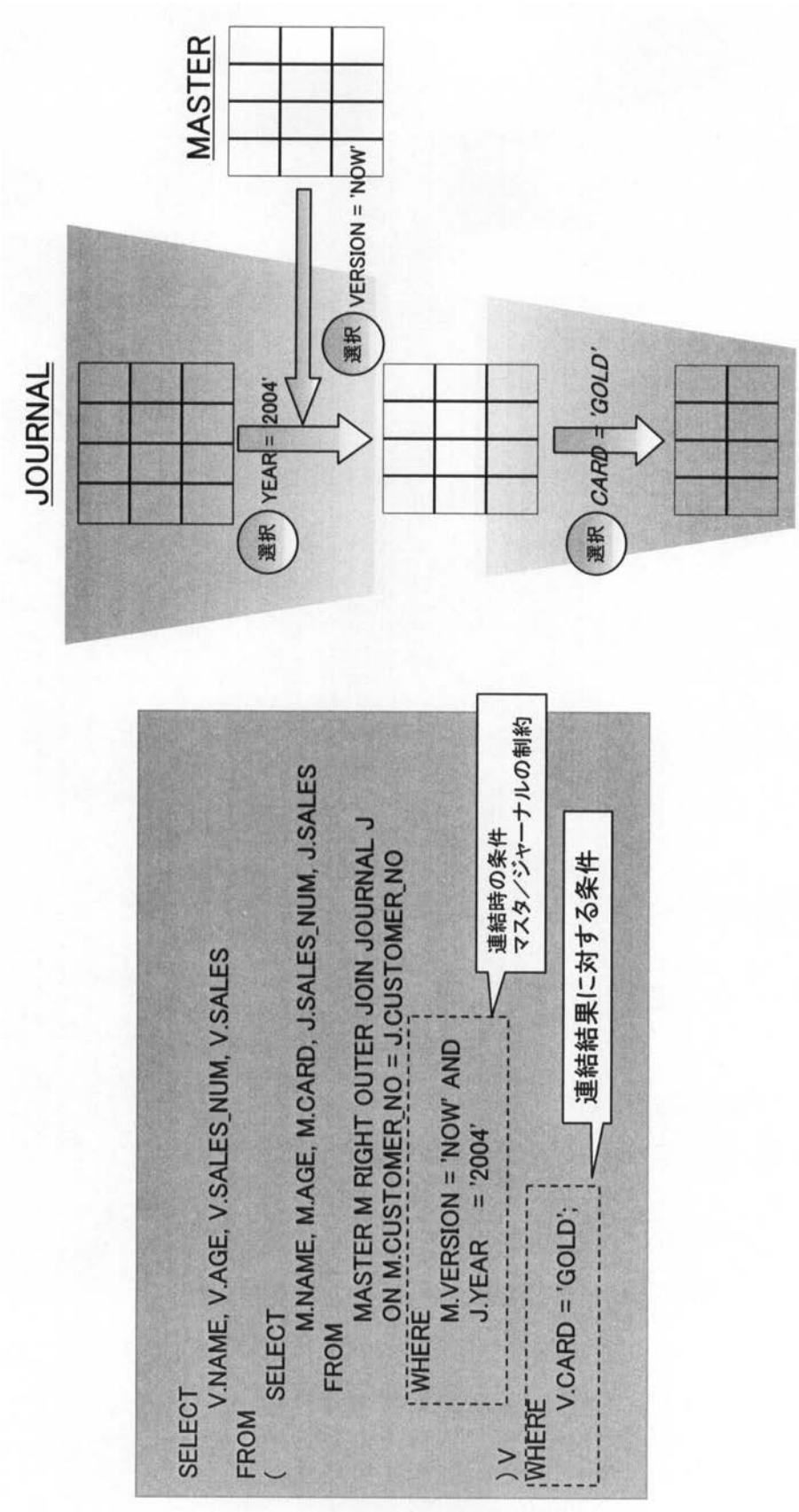
【図１７】


スタート
S41
通知有り
NO
YES
S42
照合
S43
対象検出
NO
YES
S44
設定
S45
終端検出
NO
YES
S46
通知

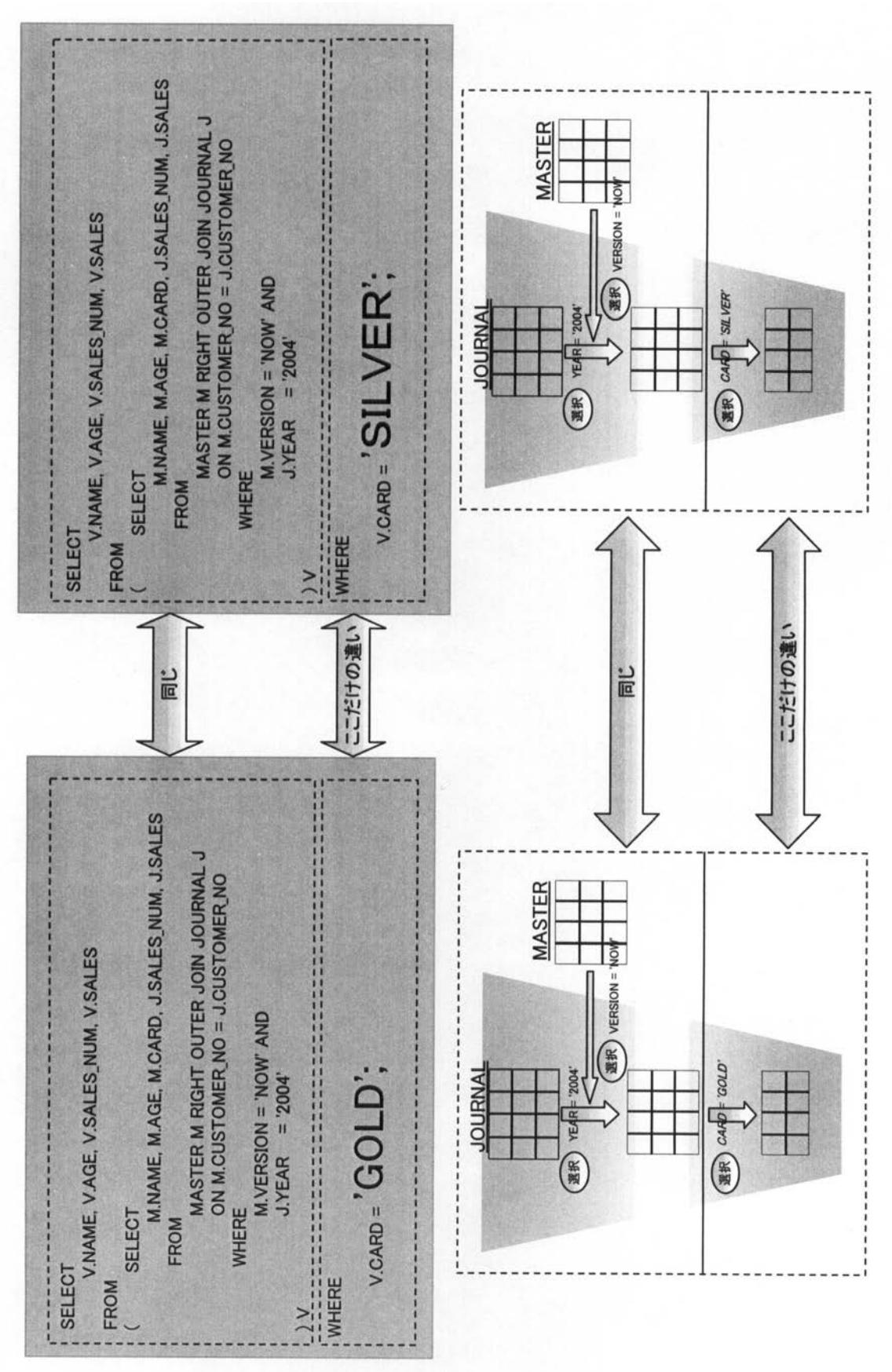
【図１８】


スタート
S51
通知有り
NO
YES
S52
抽出条件判定
S53
有り
NO
YES
S54
出力

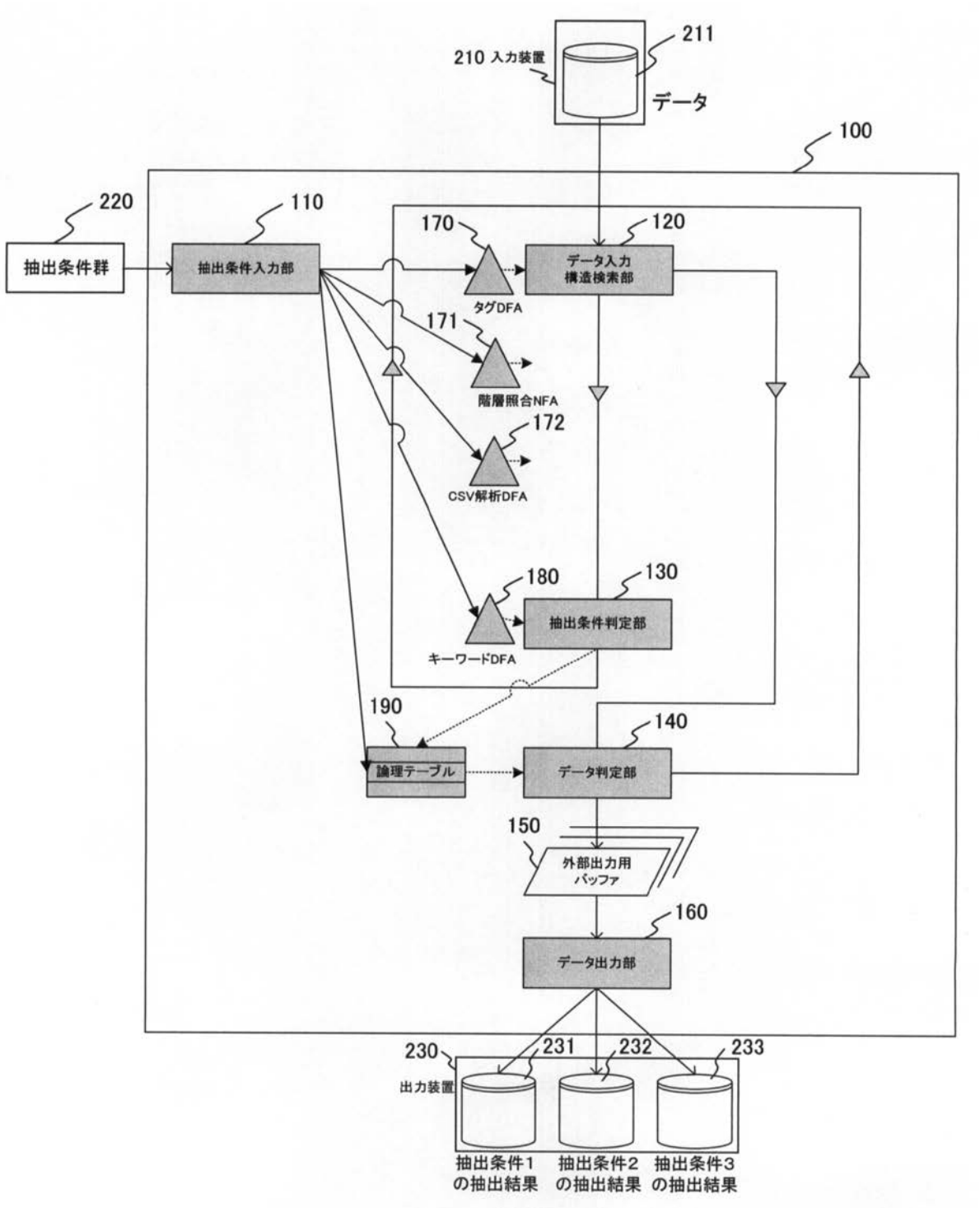
【図１】


JOURNAL
MASTER
選択
YEAR = '2004'
選択
VERSION = 'NOW'
選択
CARD = 'GOLD'
SELECT
V.NAME, V.AGE, V.SALES_NUM, V.SALES
FROM
(
SELECT
M.NAME, M.AGE, M.CARD, J.SALES_NUM, J.SALES
FROM
MASTER M RIGHT OUTER JOIN JOURNAL J
ON M.CUSTOMER_NO = J.CUSTOMER_NO
WHERE
M.VERSION = 'NOW' AND
J.YEAR = '2004'
) V
WHERE
V.CARD = 'GOLD';
連結時の条件
マスタ／ジャーナルの制約
連結結果に対する条件

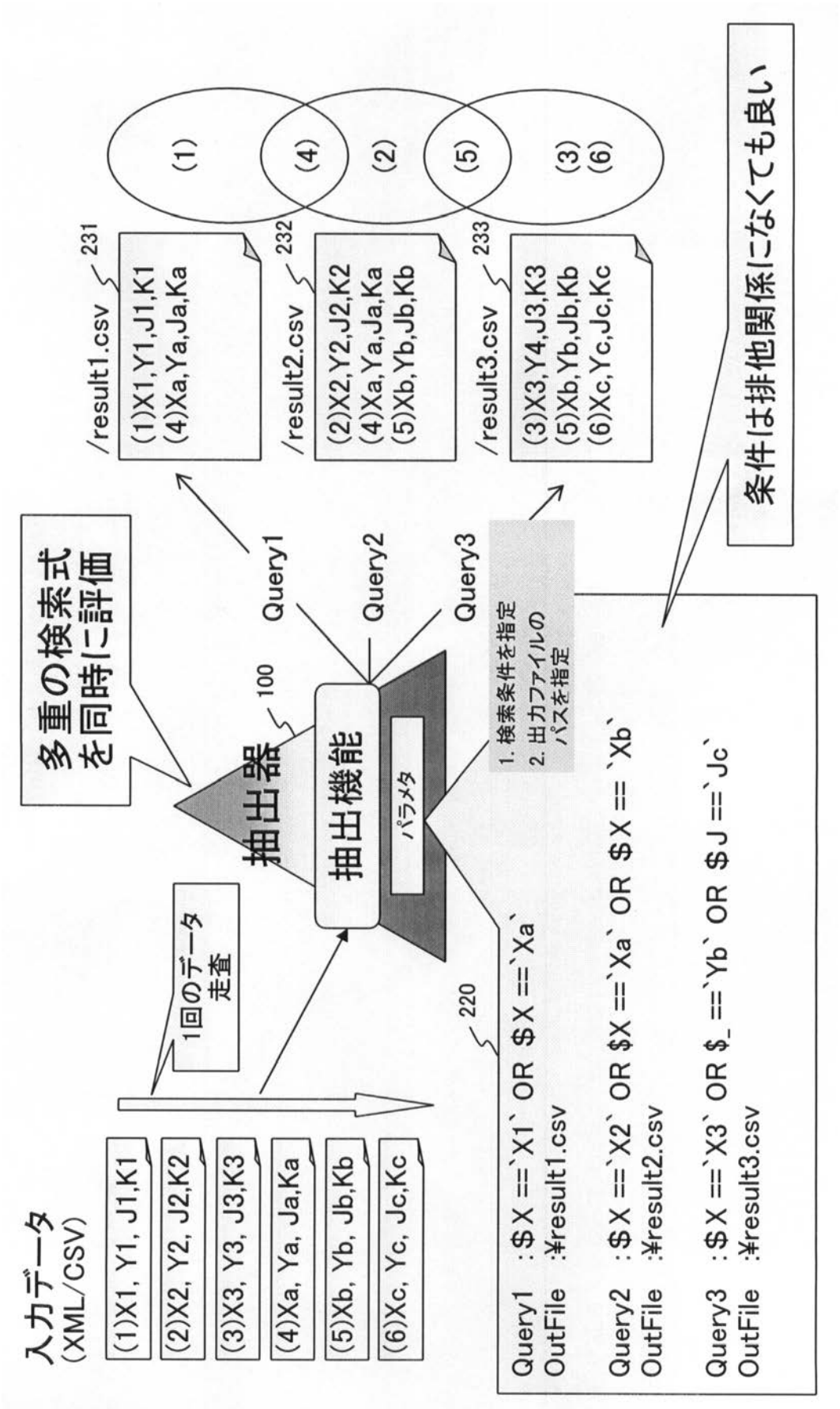
【図２】


SELECT
V.NAME, V.AGE, V.SALES_NUM, V.SALES
FROM
(
SELECT
M.NAME, M.AGE, M.CARD, J.SALES_NUM, J.SALES
FROM
MASTER M RIGHT OUTER JOIN JOURNAL J
ON M.CUSTOMER_NO = J.CUSTOMER_NO
WHERE
M.VERSION = 'NOW' AND
J.YEAR = '2004'
) V
WHERE
V.CARD = 'GOLD';
同じ
ここだけの違い
SELECT
V.NAME, V.AGE, V.SALES_NUM, V.SALES
FROM
(
SELECT
M.NAME, M.AGE, M.CARD, J.SALES_NUM, J.SALES
FROM
MASTER M RIGHT OUTER JOIN JOURNAL J
ON M.CUSTOMER_NO = J.CUSTOMER_NO
WHERE
M.VERSION = 'NOW' AND
J.YEAR = '2004'
) V
WHERE
V.CARD = 'SILVER';
JOURNAL
MASTER
選択
YEAR = '2004'
VERSION = 'NOW'
CARD = 'GOLD'
CARD = 'SILVER'
同じ
ここだけの違い

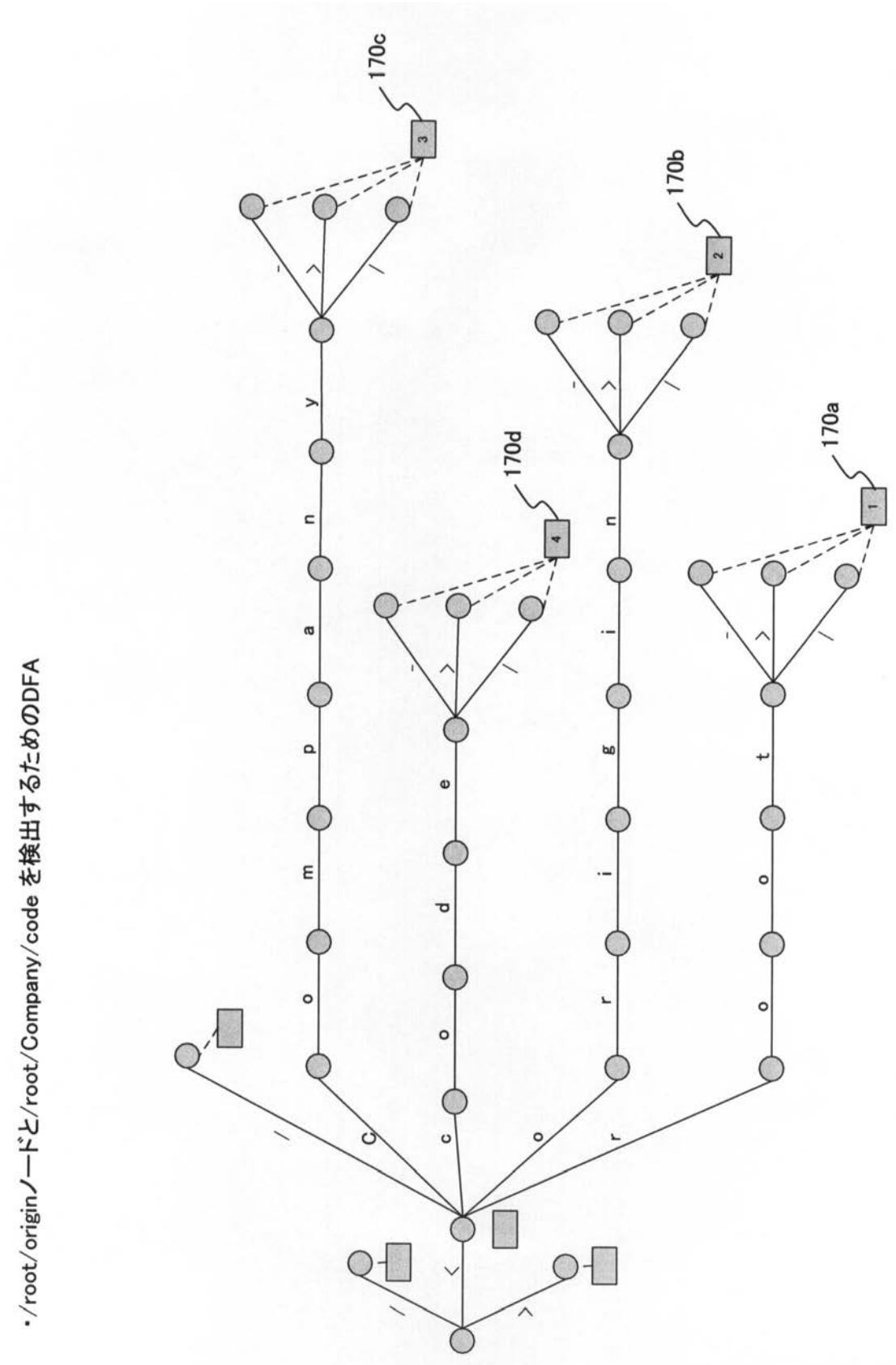
【図３】


210 入力装置
211
データ
100
220
抽出条件群
110
抽出条件入力部
170
タグDFA
120
データ入力
構造検索部
171
階層照合NFA
172
CSV解析DFA
180
キーワードDFA
130
抽出条件判定部
190
論理テーブル
140
データ判定部
150
外部出力用
バッファ
160
データ出力部
230
出力装置
231
232
233
抽出条件1
の抽出結果
抽出条件2
の抽出結果
抽出条件3
の抽出結果

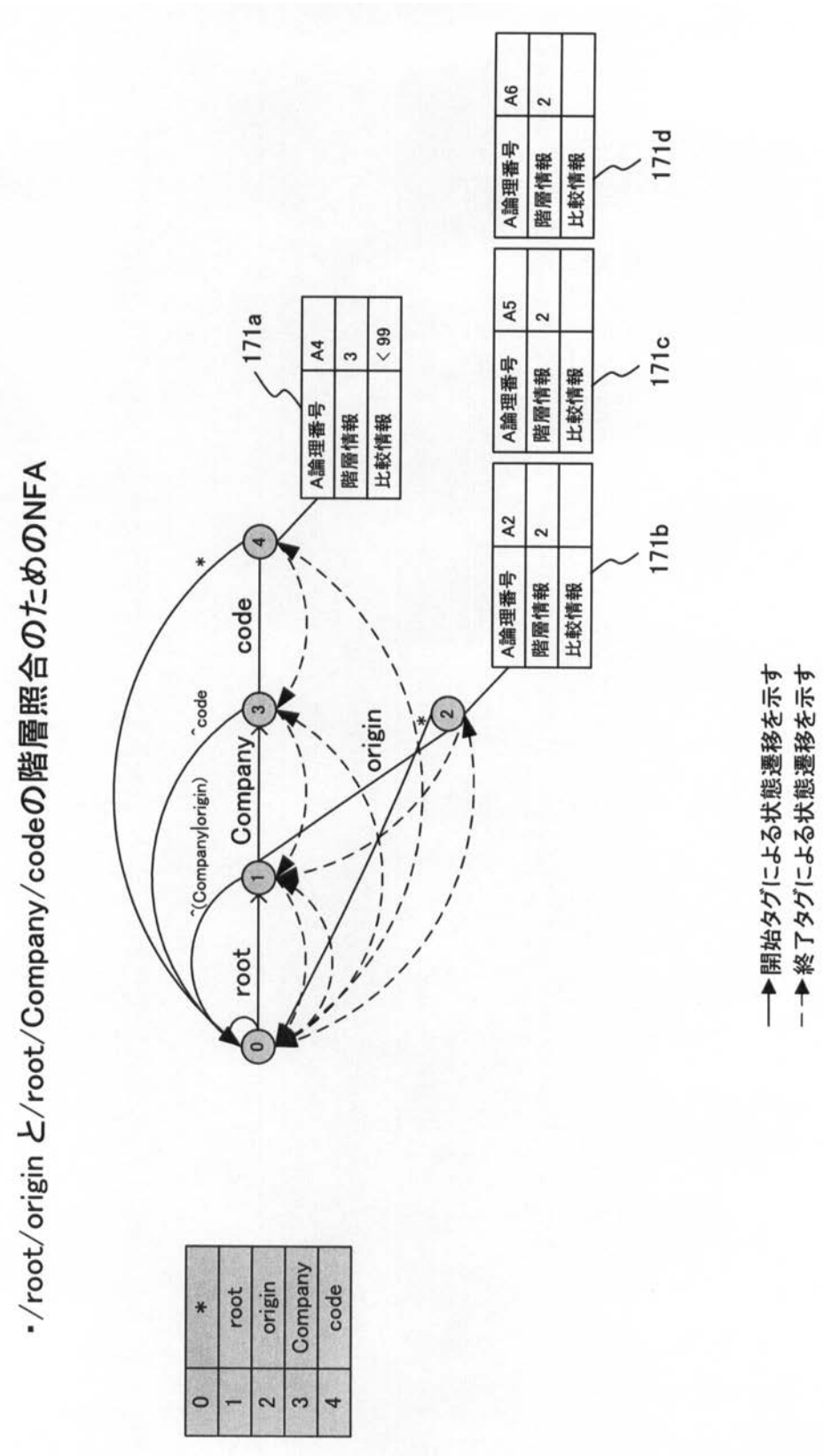
【図４】


入力データ
(XML/CSV)
(1)X1, Y1, J1,K1
(2)X2, Y2, J2,K2
(3)X3, Y3, J3,K3
(4)Xa, Ya, Ja,Ka
(5)Xb, Yb, Jb,Kb
(6)Xc, Yc, Jc,Kc
1回のデータ走査
抽出器
100
抽出機能
パラメタ
220
Query1
Query2
Query3
1. 検索条件を指定
2. 出力ファイルのパスを指定
多重の検索式を同時に評価
Query1 :$X == `X1` OR $X == `Xa`
OutFile :¥result1.csv
Query2 :$X == `X2` OR $X == `Xa` OR $X == `Xb`
OutFile :¥result2.csv
Query3 :$X == `X3` OR $_ == `Yb` OR $J == `Jc`
OutFile :¥result3.csv
/result1.csv
231
(1)X1,Y1,J1,K1
(4)Xa,Ya,Ja,Ka
/result2.csv
232
(2)X2,Y2,J2,K2
(4)Xa,Ya,Ja,Ka
(5)Xb,Yb,Jb,Kb
/result3.csv
233
(3)X3,Y4,J3,K3
(5)Xb,Yb,Jb,Kb
(6)Xc,Yc,Jc,Kc
(1)
(4)
(2)
(5)
(3)
(6)
条件は排他関係になくても良い

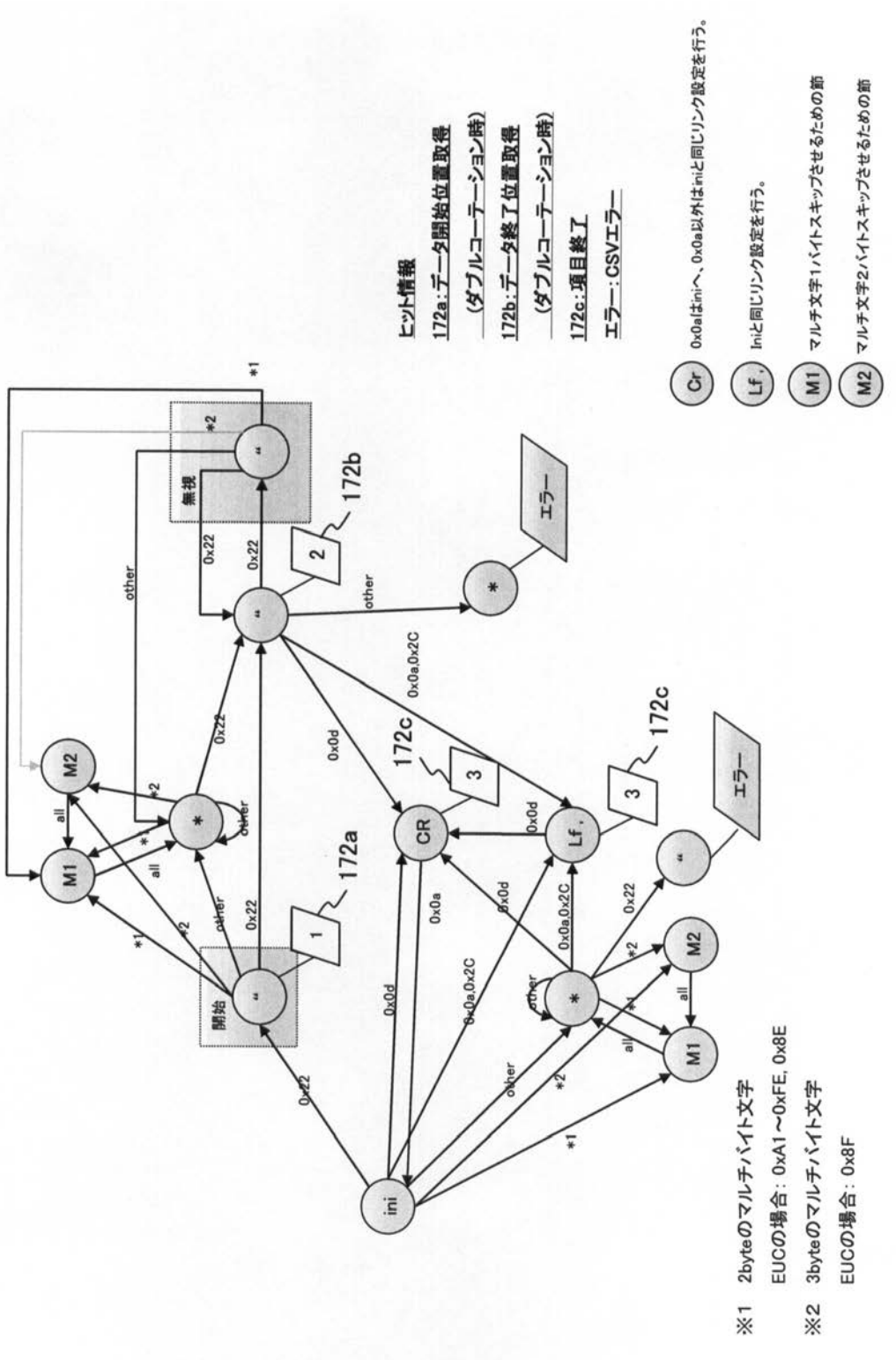
【図９】


・/root/originノードと/root/Company/code を検出するためのDFA
170c
170b
170d
170a
1
2
3
4

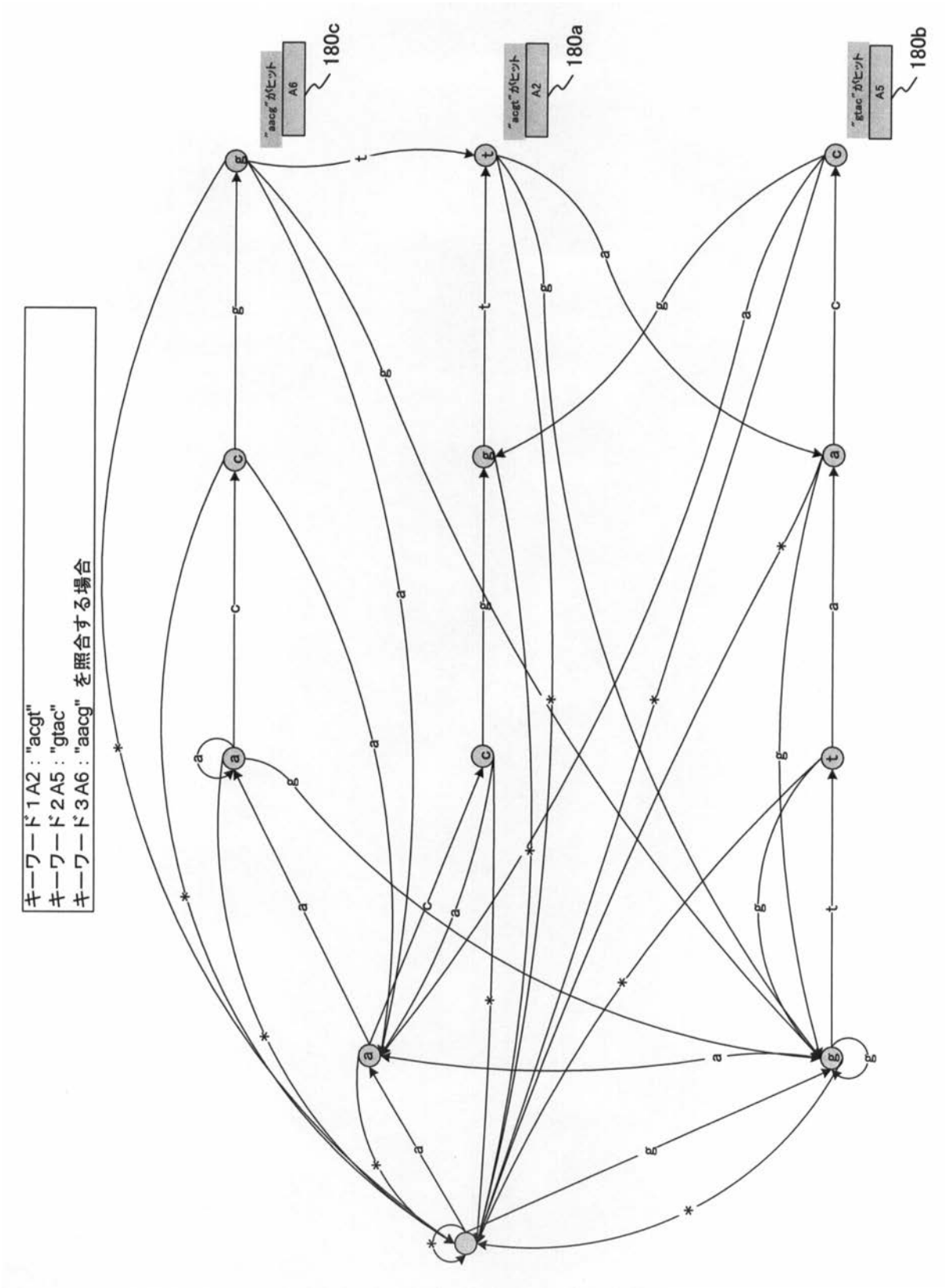
【図１０】


・/root/origin と/root/Company/codeの階層照合のためのNFA
0 *
1 root
2 origin
3 Company
4 code
root
Company
origin
code
*
^(Company|origin)
^code
A論理番号 A4
階層情報 3
比較情報 < 99
171a
A論理番号 A2
階層情報 2
比較情報
171b
A論理番号 A5
階層情報 2
比較情報
171c
A論理番号 A6
階層情報 2
比較情報
171d
開始タグによる状態遷移を示す
終了タグによる状態遷移を示す

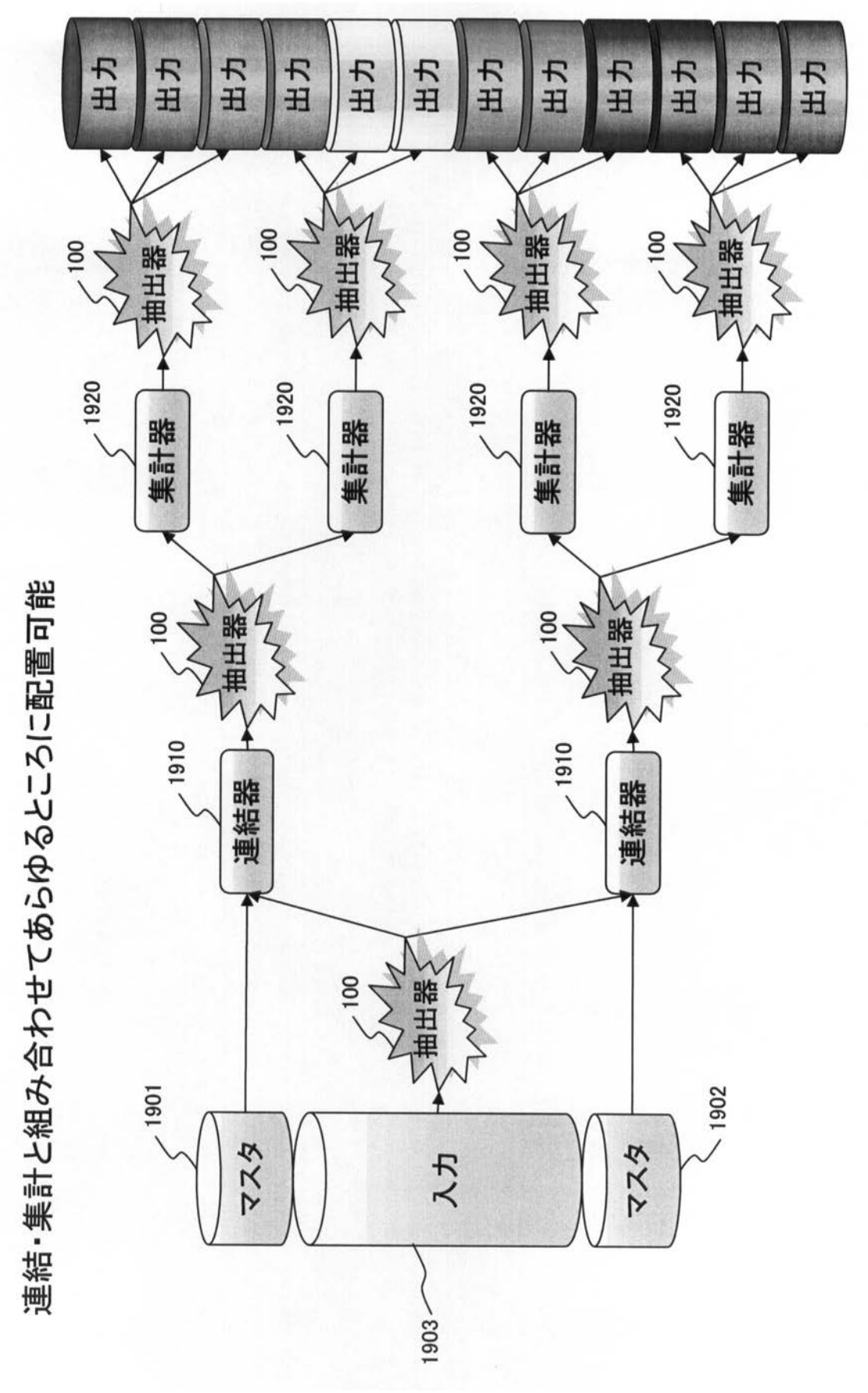
【図１１】


ヒット情報
172a:データ開始位置取得
(ダブルコーテーション時)
172b:データ終了位置取得
(ダブルコーテーション時)
172c:項目終了
エラー:CSVエラー
Cr 0x0aはini～、0x0a以外はiniと同じリンク設定を行う。
Lf, Iniと同じリンク設定を行う。
M1 マルチ文字1バイトスキップさせるための節
M2 マルチ文字2バイトスキップさせるための節
※1 2byteのマルチバイト文字
EUCの場合：0xA1～0xFE, 0x8E
※2 3byteのマルチバイト文字
EUCの場合：0x8F
開始
無視
ini
CR
Lf,
エラー
172a
172b
172c
other
all
0x0d
0x0a
0x0a,0x2C
0x22

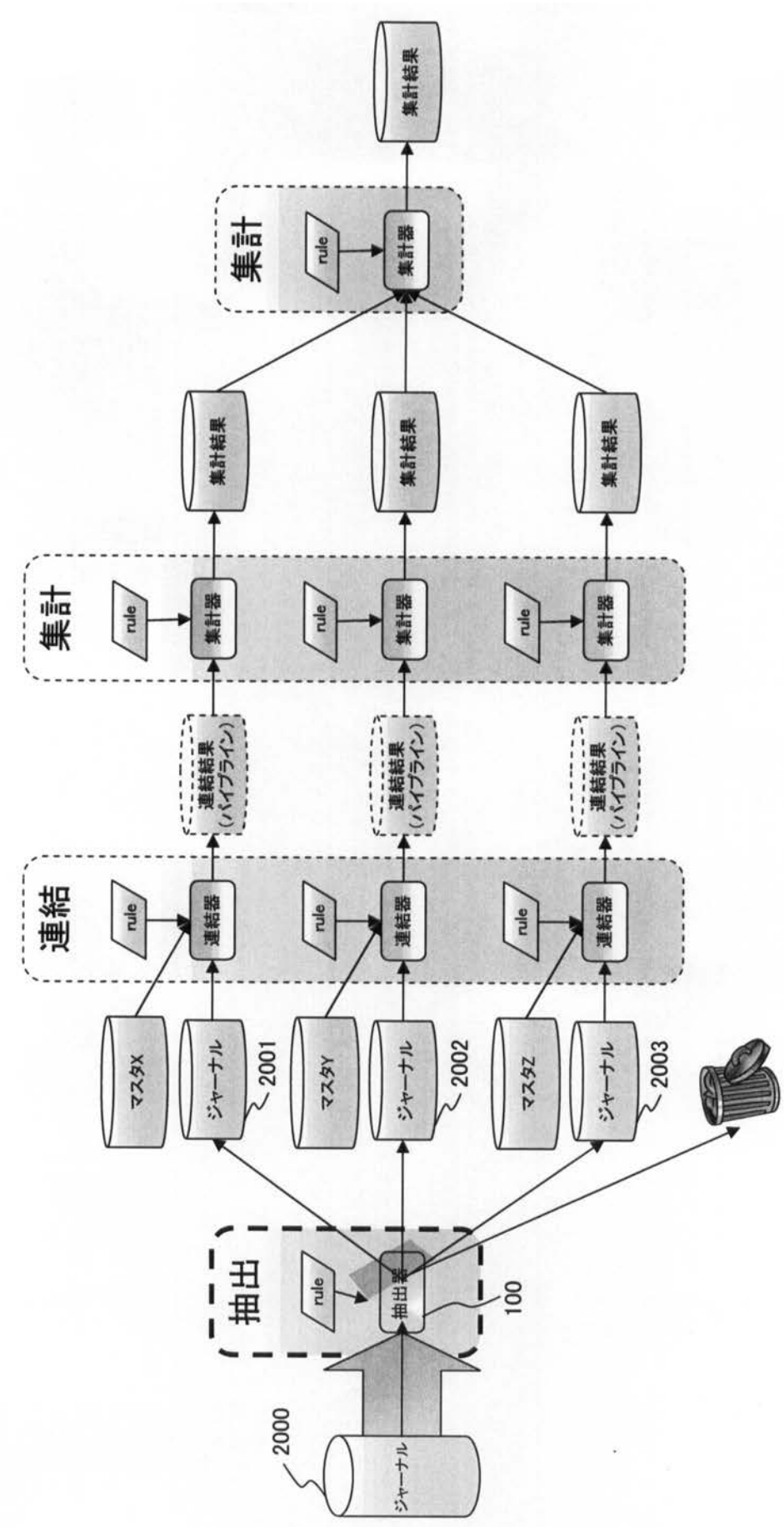
【図１２】


キーワード1 A2 : "acgt"
キーワード2 A5 : "gtac"
キーワード3 A6 : "aacg" を照合する場合
"aacg"がヒット
A6
180c
"acgt"がヒット
A2
180a
"gtac"がヒット
A5
180b

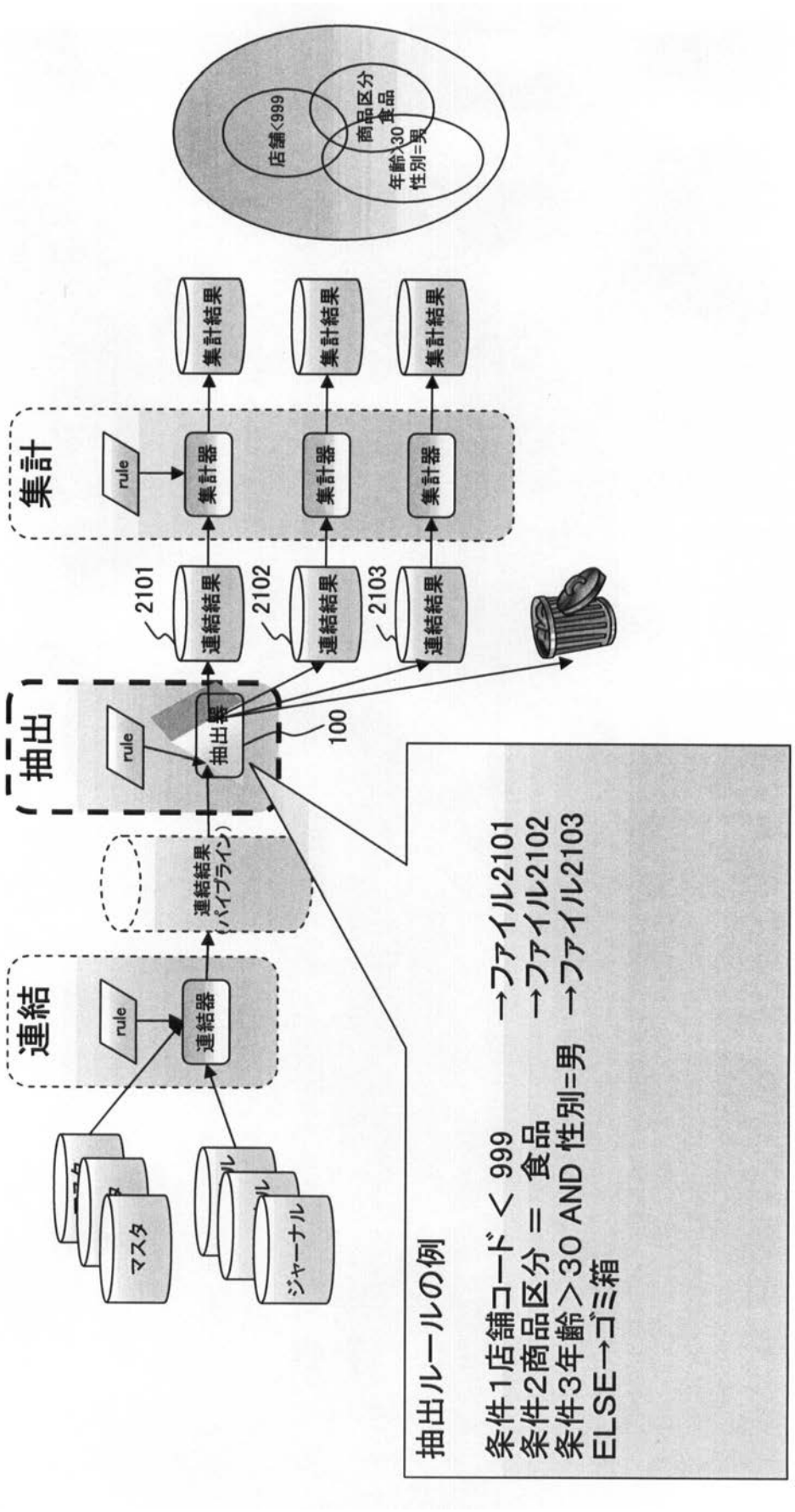
【図１９】


連結・集計と組み合わせてあらゆるところに配置可能
出力
抽出器
100
集計器
1920
連結器
1910
マスタ
1901
入力
1903
マスタ
1902

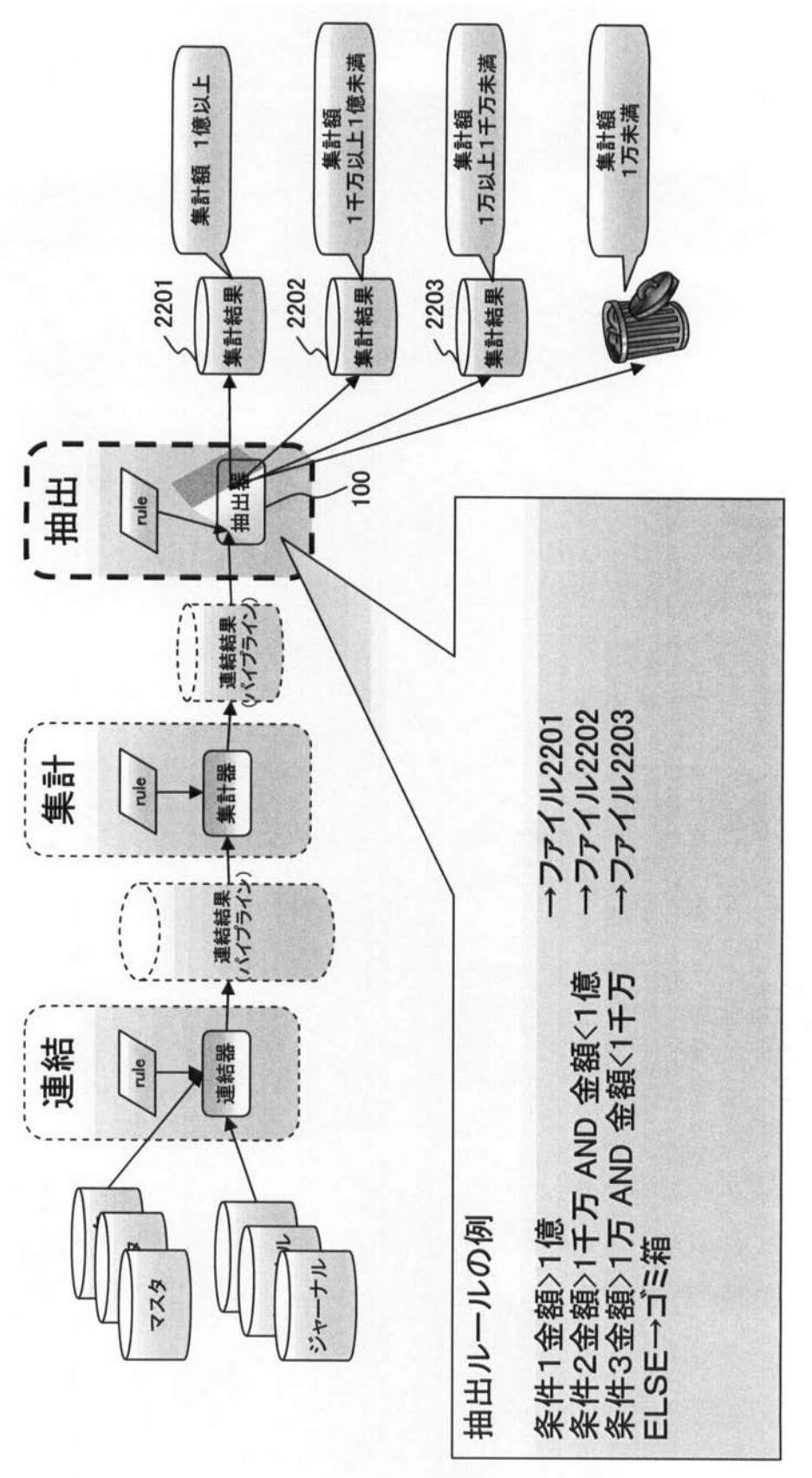
【図２０】


集計結果
集計
rule
集計器
集計結果
集計結果
集計結果
集計
rule
集計器
rule
集計器
rule
集計器
連結結果
（パイプライン）
連結結果
（パイプライン）
連結結果
（パイプライン）
連結
rule
連結器
rule
連結器
rule
連結器
マスタX
ジャーナル
2001
マスタY
ジャーナル
2002
マスタZ
ジャーナル
2003
抽出
rule
抽出器
100
2000
ジャーナル

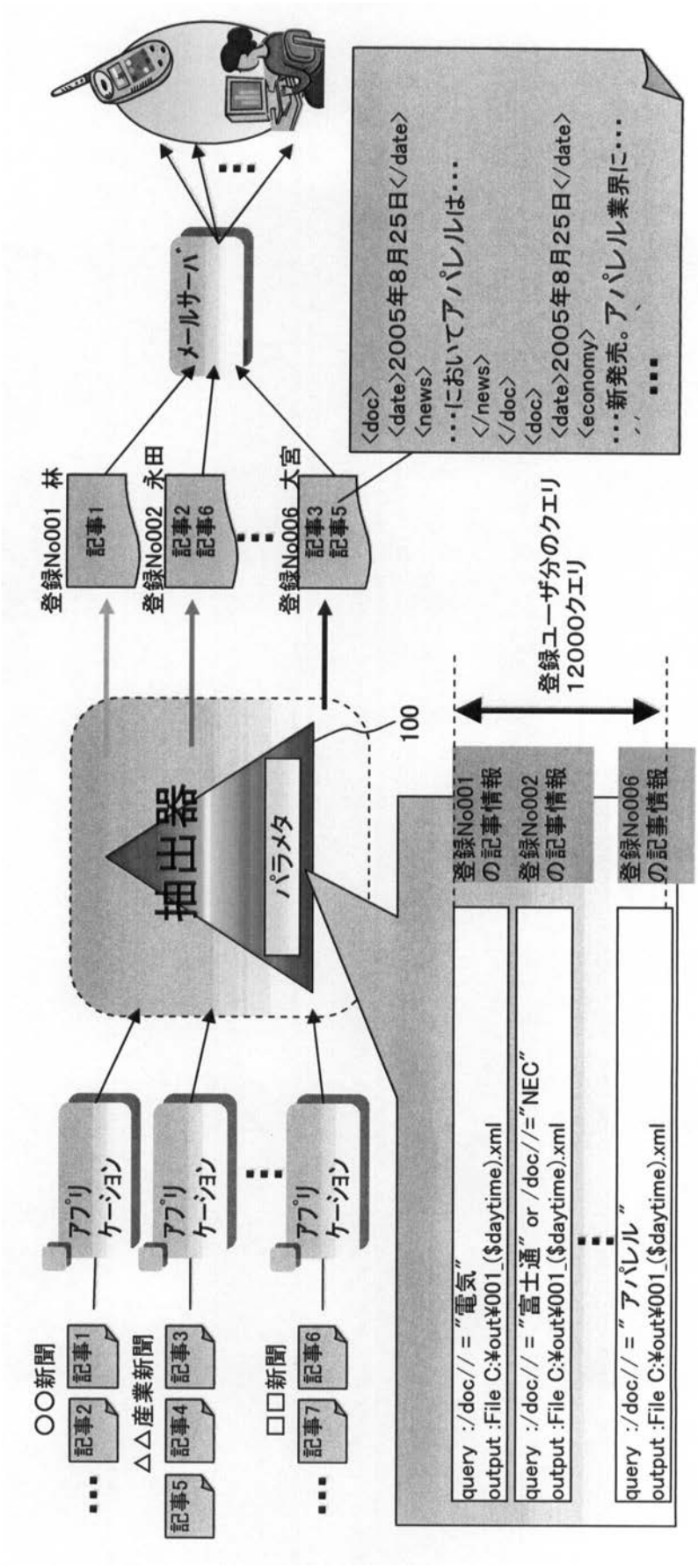
【図２１】


店舗<999
商品区分
食品
年齢>30
性別=男
集計結果
集計結果
集計結果
集計
rule
集計器
集計器
集計器
2101
2102
2103
連結結果
連結結果
連結結果
抽出
rule
抽出条
100
連結結果
(パイプライン)
連結
rule
連結器
マスタ
ジャーナル
抽出ルールの例
条件1店舗コード＜999 →ファイル2101
条件2商品区分＝食品 →ファイル2102
条件3年齢＞30 AND 性別=男 →ファイル2103
ELSE→ゴミ箱

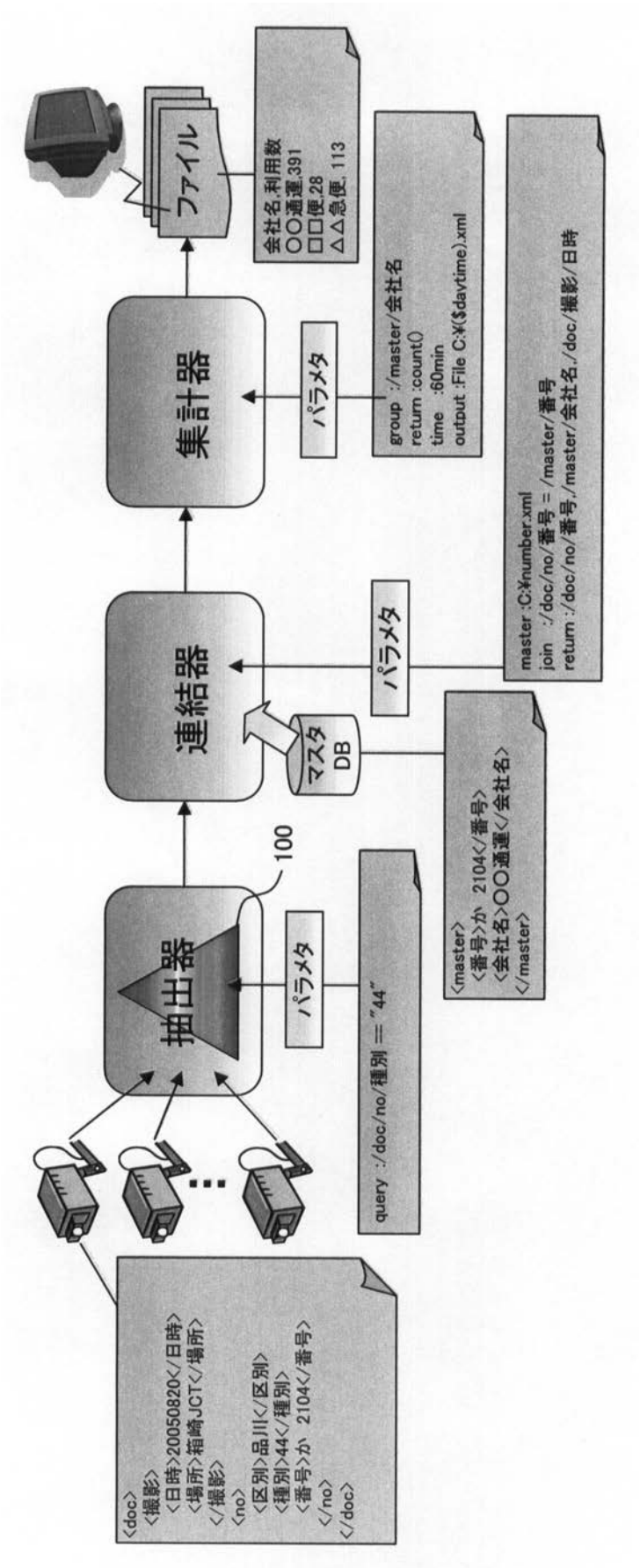
【図２２】


抽出
集計
連結
rule
連結器
集計器
抽出器
連結結果
(パイプライン)
マスタ
ジャーナル
100
2201
2202
2203
集計結果
集計額 1億以上
集計額
1千万以上1億未満
集計額
1万以上1千万未満
集計額
1万未満
抽出ルールの例
条件1金額>1億 →ファイル2201
条件2金額>1千万 AND 金額<1億 →ファイル2202
条件3金額>1万 AND 金額<1千万 →ファイル2203
ELSE→ゴミ箱

【図２３】


登録No001 林
登録No002 永田
登録No006 太宮
記事1
記事2
記事6
記事3
記事5
メールサーバ
<doc>
<date>2005年8月25日</date>
<news>
・・・においてアパレルは・・・
</news>
</doc>
<doc>
<date>2005年8月25日</date>
<economy>
・・・新発売。アパレル業界に・・・
</economy>
</doc>
抽出器
パラメタ
100
登録ユーザ分のクエリ
12000クエリ
query :/doc//＝"電気"
output :File C:¥out¥001_($daytime).xml
query :/doc//＝"富士通" or /doc//="NEC"
output :File C:¥out¥001_($daytime).xml
query :/doc//＝"アパレル"
output :File C:¥out¥001_($daytime).xml
登録No001の記事情報
登録No002の記事情報
登録No006の記事情報
アプリケーション
アプリケーション
アプリケーション
○○新聞
△△産業新聞
□□新聞
記事1
記事2
記事3
記事4
記事5
記事6
記事7

【図２４】


ファイル
会社名,利用数
〇〇通運,391
□□便,28
△△急便, 113
集計器
パラメタ
group :/master/会社名
return :count()
time :60min
output :File C:¥($daytime).xml
連結器
パラメタ
master :C:¥number.xml
join :/doc/no/番号 = /master/番号
return :/doc/no/番号,/master/会社名,/doc,/撮影/日時
マスタ
DB
100
抽出器
パラメタ
query :/doc/no/種別 = "44"
<master>
<番号>か 2104</番号>
<会社名>〇〇通運</会社名>
</master>
<doc>
<撮影>
<日時>20050820</日時>
<場所>箱崎JCT</場所>
</撮影>
<no>
<区別>品川</区別>
<種別>44</種別>
<番号>か 2104</番号>
</no>
</doc>

フロントページの続き

(72)発明者　永田　真彦
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　大宮　清英
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

合議体
審判長　長島　孝志
審判官　田中　秀人
審判官　仲間　晃

(56)参考文献　特開２０００－３３９３４６（ＪＰ，Ａ）
特開２００５－１０１９２８（ＪＰ，Ａ）
特開平６－１３９２９１（ＪＰ，Ａ）
特開２００１－３４４２８２（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F17/30,12/00