

(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 1 月 9 日(09.01.2014)



(10) 国際公開番号
WO 2014/006735 A1

- (51) 国際特許分類:
G06F 13/00 (2006.01)
- (21) 国際出願番号: PCT/JP2012/067270
- (22) 国際出願日: 2012 年 7 月 6 日(06.07.2012)
- (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 (米国を除く全ての指定国について): 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; および
- (75) 発明者/出願人 (米国についてのみ): 高田 雅士 (TAKADA Masashi) [JP/JP]; 〒1858601 東京都国分寺市東恋ヶ窪一丁目 2 8 0 番地 株式会社日立製作所 中央研究所内 Tokyo (JP). 工藤 泰幸 (KUDO Yasuyuki) [JP/JP]; 〒1858601 東京都国分寺

市東恋ヶ窪一丁目 2 8 0 番地 株式会社日立製作所 中央研究所内 Tokyo (JP). 加藤 猛(KATO Takeshi) [JP/JP]; 〒1858601 東京都国分寺市東恋ヶ窪一丁目 2 8 0 番地 株式会社日立製作所 中央研究所内 Tokyo (JP).

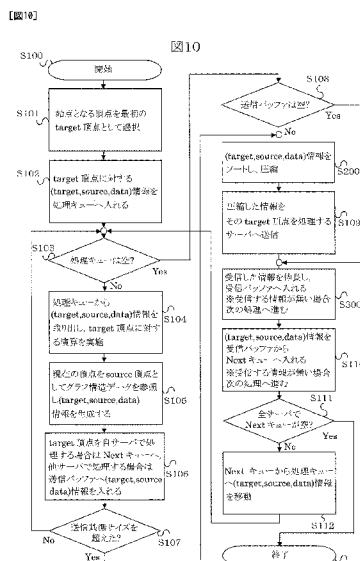
(74) 代理人: 井上 学, 外(INOUE Manabu et al.); 〒1008220 東京都千代田区丸の内一丁目 6 番 1 号 株式会社日立製作所内 Tokyo (JP).

(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, 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, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST,

[続葉有]

(54) Title: TRANSFER METHOD AND GRAPH PROCESSING SYSTEM

(54) 発明の名称: 転送方法およびグラフ処理システム



S100 Start
S101 Select vertex to be starting point as first target vertex
S102 Insert (target, source, data) information for target vertex into processing queue
S103 Processing queue is empty?
S104 Dequeue (target, source, data) information from processing queue, and perform computation for target vertex
S105 Reference graph structure data using current vertex as source vertex, and generate (target, source, data) information
S106 Insert (target, source, data) information into Next queue if target vertex is to be processed at own server, and into transmission buffer (target, source, data) if to be processed at another server
S107 Exceeded transmission reference size?
S108 Transmission buffer is empty?
S109 Sort and compress (target, source, data) information
S110 Transmit compressed information to server which processes target vertex of information
S111 Decompress received information, and insert into reception buffer
S112 If there is no information to receive, proceed to next processing
S113 Insert (target, source, data) information into Next queue from reception buffer
S114 Next queue is empty at all servers?
S115 Move (target, source, data) information from Next queue to processing queue
S116 End

(57) Abstract: In the case of graph processing using a plurality of compute nodes, there is the problem that, as the scale of a graph to be processed increases, the amount of intermediate data to be transferred between compute nodes increases, and therefore, the amount of time required for transfer of the intermediate data becomes longer, and subsequently, the speed of the entirety of the graph processing becomes slower. The present invention resolves the above problem, which is for transfer of intermediate data having sets of information of a graph vertex of a transmission source and information of a graph vertex of a transmission destination in graph processing by way of a plurality of compute nodes, by providing a transmission buffer which accumulates the intermediate data in each compute node, and on the basis of the information of the graph vertex of the transmission source or the information of the graph vertex of the transmission destination, sorting the order of the intermediate data within the group of the accumulated intermediate data and transferring the group of the intermediate data.

(57) 要約: 複数の計算ノードを用いたグラフ処理の場合に、処理対象となるグラフの規模が大きくなるほど、計算ノード間で転送される中間データ量が多くなるために、中間データの転送に要する時間が長くなり、ひいてはグラフ処理全体の速度が遅くなるという問題がある。本発明は、複数の計算ノードによるグラフ処理での、送信元のグラフ頂点の情報と送信先のグラフ頂点の情報の組を有する中間データの転送を、各計算ノードに中間データを蓄積する送信バッファを設け、送信元のグラフ頂点の情報または送信先のグラフ頂点の情報に基づいて、蓄積された中間データの群の中の間データの並びをソートし、中間データの群を転送することで、前述の課題を解決する。



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, 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： 転送方法およびグラフ処理システム

技術分野

[0001] 本発明は、複数の計算ノードを用いたグラフ処理に関し、特にグラフ情報の計算ノード間での転送に関するものである。

背景技術

[0002] グラフ処理における情報圧縮技術として、非特許文献1に記載されている技術がある。非特許文献1に記載の技術は、非ゼロの値を格納したベクトルA、非ゼロの箇所の列番号を格納したベクトルB、各行の先頭の非ゼロの箇所がベクトルAの何番目の値に対応するかを格納したベクトルCでグラフ構造データを表現することで情報を圧縮するものである。

先行技術文献

非特許文献

[0003] 非特許文献1：リチャード・バレット (Richard Barrett)、外9名著、「線形システムのソリューション向けテンプレート：反復法のためのビルディングブロック (Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods)」、ソサイエティ・フォー・インダストリアル・アンド・アプライド・マスマティクス (Society for Industrial and Applied Mathematics)、(米国)、1994年、pp. 64-65

発明の概要

発明が解決しようとする課題

[0004] 非特許文献1に記載の技術では、グラフ構造データを圧縮することは可能であるが、グラフ処理時の中間データであるグラフの頂点間を伝搬する情報は圧縮できない。本願発明者らは、複数の計算ノードを用いたグラフ処理の場合に、処理対象となるグラフの規模が大きくなるほど、計算ノード間で転

送される中間データ量が多くなるために、中間データの転送に要する時間が長くなり、ひいてはグラフ処理全体の速度が遅くなるという問題があることを見出した。

[0005] そこで本発明は、効率的なグラフ処理の中間データの転送を目的とする。

課題を解決するための手段

[0006] 本発明は、複数の計算ノードによるグラフ処理での、送信元のグラフ頂点の情報と送信先のグラフ頂点の情報の組を有する中間データの転送を、各計算ノードに中間データを蓄積する送信バッファを設け、送信元のグラフ頂点の情報または送信先のグラフ頂点の情報に基づいて、蓄積された中間データの群の中の間接データの並びをソートし、中間データの群を転送することで、前述の課題を解決する。

発明の効果

[0007] 本発明によれば、複数計算ノードでグラフ処理を行う際に、効率的な中間データの転送を実現できる。

図面の簡単な説明

[0008] [図1]本発明の実施例に係るグラフ処理システムを示す図である。

[図2]図1のApサーバのシステム構成を示す図である。

[図3]図2のサーバ装置の内部構成を示すブロック図である。

[図4]図3のメモリ装置上に用意された送信バッファや受信バッファを示す図である。

[図5]処理対象のグラフの例を示す図である。

[図6]図5のグラフを行列形式で表現した図である。

[図7]図5のグラフをCSR形式で表現した図である。

[図8]頂点間の転送情報の構成を示した図である。

[図9]頂点割当て情報の例を示す図である。

[図10]グラフ処理の例を示すフローチャートである。

[図11]図10の圧縮処理の詳細を示すフローチャートである。

[図12]図10の伸長処理の詳細を示すフローチャートである。

[図13]図5のグラフの各頂点がどのサーバで処理されるかの例を示した説明図である。

[図14]図13に対し頂点間の情報の伝搬を示した図である。

[図15(a)]サーバ装置420からサーバ装置430へ転送される情報のソート前の例を示す説明図である。

[図15(b)]サーバ装置420からサーバ装置430へ転送される情報のソート後の例を示す説明図である。

[図15(c)]サーバ装置420からサーバ装置430へ転送される情報の圧縮後の例を示す説明図である。

発明を実施するための形態

実施例 1

[0009] 図1は、本発明の実施の形態に係るグラフ処理システム10の構成を概略的に示す図である。図1に示されるように、グラフ処理システム10は、インターネット200を介して通信可能なPCや携帯端末といったクライアント100、インターネット200を介してクライアント100からの要求を受け付けるウェブサーバ（Webサーバ）300、Webサーバ300からの要求によりグラフ解析を行うアプリケーションサーバ（Apサーバ）400、およびデータベースへのアクセスを行うデータベースサーバ（DBサーバ）500を有している。但し、クライアント100とWebサーバ300を結ぶネットワークはインターネット200に限定されず、例えばLANでも良い。また、本実施例では、Webサーバ300、Apサーバ400、およびDBサーバ500を備えるウェブ3層構成のシステムを記載しているが、本発明はこの構成に限定されるものではない。例えばグラフ解析も行うWebサーバ300とDBサーバ500の2層構成などでも良い。

[0010] 図2は、図1のApサーバ400のシステム構成を示す図である。図2に示されるようにApサーバ400はサーバ装置420、430、440、450と、それらを結ぶネットワーク装置410とを有している。サーバ装置420-450のそれぞれは、グラフ処理における計算ノードとして働く。

サーバ装置の台数はサーバ装置の性能やタスクの負荷によって決められるものであるため、4台の構成に制限されず他の台数であっても良い。サーバ装置420-450は、並列処理を行い互いに通信しながら、最短経路問題を解く等のグラフ処理を実行する。サーバ装置420-450には、それぞれサーバ識別情報（サーバID）が与えられる。本実施例では、3台のサーバ装置をグラフ処理に用い、サーバ装置420のサーバIDは「1」、サーバ装置430のサーバIDは「2」、サーバ装置440のサーバIDは「3」である。

[0011] 図3は、図2に示されるサーバ装置420の内部構成を示すブロック図である。サーバ装置430、440、450の内部構成は、サーバ装置420の内部構成と同等の機能を有していれば良く、例えば、サーバ毎に異なるメーカーや性能であってもかまわない。以降、サーバ装置420を代表として取り上げて説明する。

[0012] サーバ装置420は、中央処理装置（CPU）600、メモリ装置610、ストレージ装置620、入力装置630、出力装置640、ネットワークインタフェース（I/F）650、およびバス660を備える。サーバ装置420内の構成要素間のデータ転送は主にバス660を介して行われる。CPU600はメモリ装置610を使用してグラフ解析プログラムを実行するとともにサーバ装置420全体の動作を制御する。メモリ装置610は、SDRAMなどの1次記憶装置であり、CPU600がプログラムを実行する際に必要な命令やデータを保持する。ストレージ装置620は、HDDやSSDといった2次記憶装置であり、プログラムやデータを長期間保持する他、メモリ装置610のスワップ領域としても利用される。入力装置630は、マウスやキーボードなどであり、出力装置640は表示装置やスピーカーなどである。ネットワークインタフェース650は他のサーバ装置との通信に利用されるものであり、InfiniBandなどを用いることができる。

[0013] 図4は、メモリ装置610のメモリ領域に確保された送信バッファ611

と、受信バッファ 612 と、処理キュー 613 と、グラフ演算モジュール 614 と、頂点割当て情報 615 と、次の回の処理を蓄積する NEXT キュー 616 と、圧縮モジュール 617 と、伸長モジュール 618 とを示す図である。送信バッファ 611 は、サーバ装置間で通信を行う際に送信するデータを一時保管するためのメモリ領域である。受信バッファ 612 は、サーバ装置間で通信を行う際に受信するデータを一時保管するためのメモリ領域である。サーバ装置間で転送されるグラフ処理の中間データは、送信バッファ 611 や受信バッファ 612 に一時保管される。送信バッファ 611 に蓄えられたデータは、CPU 600 の送信指示によりネットワークインタフェース 650 へ転送された後、ネットワーク装置 410 を介して別のサーバ装置へ送信される。送信データは、別のサーバ装置のネットワークインタフェース 650 で受信された後、別のサーバ装置の受信バッファ 612 に蓄えられる。処理キュー 613 は、グラフ処理のあるフェーズにおいて演算可能なデータを一時保管するためのメモリ領域であり、NEXT キュー 616 は、その次のフェーズにおいて演算可能なデータを一時保管するためのメモリ領域である。グラフ演算モジュールは、最短経路探索などのグラフ処理の演算を行うプログラムモジュールである。圧縮モジュール 617 は、後述するが、サーバ装置間で転送されるグラフ処理の中間データを圧縮するプログラムモジュールである。伸長モジュール 618 は、後述するが、サーバ装置間で転送されるグラフ処理の中間データを伸長するプログラムモジュールである。頂点割当て情報 615 は、後述するが、複数のサーバ装置に対するグラフ処理の割当てを示す情報である。

[0014] 図 5 は、本実施例のグラフ処理システム 10 の動作を説明するために具体例としてあげたグラフの図である。図 5 中の円はグラフの頂点を表しており、円内には頂点の識別番号（以下、頂点番号）を示している。頂点間の線はグラフの辺である。グラフは様々な事柄の関係性を表現するのに適している。例えば、グラフの頂点を駅や交差点とみなすと辺は線路や道路を表し、頂点を人や企業とみなした場合には、辺は人や企業間の相互関係を示す。辺は

頂点間の関係を示す重みを持ち、前者の例では時間や距離、後者の例では結びつきの強さを示す。

[0015] 図6および図7は、図5に示したグラフをグラフ構造データとして表現したもので、図5のグラフと同じグラフ構造を示している。図6は図5のグラフを行列形式で表現したものであり、図7は図5のグラフをCSR (Compressed Sparse Row) 形式で表現したものである。図6の表中の値は辺の重みを示し、値が0 (最短経路問題など解くべき問題によっては ∞ と表現した方が都合の良いものもある) の箇所は辺が存在しないことを意味する。図7のCSR形式は、非ゼロの値を格納したvalues、非ゼロの箇所の列番号を格納したcolumns、各行の先頭の非ゼロの箇所がvaluesの何番目の値に対応するかを格納したrowindexから成る。CSR形式は本図のような疎行列を表現するのに適している。行列形式では $n \times m$ の行列を表現するのに $n \times m$ の記憶容量が必要になるが、CSR形式では、非ゼロの数を l とすると $2 \times l + n + 1$ の記憶容量で済む。例えば、ビッググラフに代表されるスケールフリー性を持っているグラフは、疎行列となるため、一般にCSR形式やCSC (Compressed Sparse Column) 形式 (CSR形式に対して行と列を入れ替えたもの) といった圧縮格納形式でグラフ構造データが保存される。

[0016] 図8は、頂点間で転送される情報を示したもので、各行は送信先の頂点 (target頂点) の頂点番号 (target)、送信元の頂点 (source頂点) の頂点番号 (source)、送信データ (data) の組で構成される。ここでは、例として頂点3から頂点7、8、9へ転送される情報を示している。targetと送信先のサーバ装置、およびsourceと送信元サーバ装置の関係は、頂点割当て情報615として各サーバ装置のメモリ装置610に保存されている。

[0017] 図9は、本実施例の頂点割当て情報615を示す図である。頂点割当て情報615は、頂点番号と該頂点番号の頂点の演算が割り当てられているサーバ装置のサーバIDとの組をエントリとする。頂点割当て情報615は、各

サーバ装置への頂点の割当ての際に生成することができる。本実施例では、サーバIDの昇順に、頂点番号が若い頂点の群を配置しているが、頂点の配置の構成はこれに限定されず、頂点が各サーバに分散配置されていれば本発明を適用できる。

[0018] 図10に本実施例のグラフ処理システム10によるグラフ処理の動作例をフローチャートで示す。本実施例では、図7のグラフ構造データが処理対象としてストレージ装置620に格納されており、各サーバ装置のメモリ装置610に図7のグラフ構造データが転送された状態で、図9に示した割当てで各サーバ装置にそれぞれが演算する頂点が割当てられているとして説明する。各サーバ装置で演算する頂点の割当ては、各サーバ装置のメモリ装置610に頂点割当て情報615として、保存されている。また、グラフ解析の始点となる頂点は、サーバ装置420に割当てられるとして説明する。

[0019] まず、サーバ装置420のグラフ演算モジュール614が、グラフ解析の始点となる頂点を`target`頂点として選択する（ステップS100、S101）。次に、サーバ装置420のグラフ演算モジュール614が、選択した`target`頂点に対する`source`頂点と`data`の組を作り、処理キュー613へ入れる（ステップS102）。ここでは、`target`頂点が始点であるため、`source`頂点、`data`はダミーデータ（例えば、`source=target`、`data=Z`）とする。

[0020] ステップS103では、各サーバ装置のグラフ演算モジュール614が、各サーバ装置の処理キュー613が空であるか否かの判定を行い、各サーバ装置の動作は、空であればステップS108へ進み、空でなければステップS104へ進む。ステップS104では、各サーバ装置のグラフ演算モジュール614が、各サーバ装置の処理キュー613から（`target`、`source`、`data`）情報の組を取り出し、`target`頂点に対する演算を実施する。

[0021] 次に、ステップS105で各サーバ装置のグラフ演算モジュール614が、処理キュー613にあった`target`頂点を次の`source`頂点とし

て選択し、メモリ装置610にある図7のグラフ構造データを参照して、グラフ処理の中間データとなる、`source`頂点と、グラフ構造データから読み出した`target`頂点と、`data`の組を生成する。ステップ106で、各サーバ装置のグラフ演算モジュール614が、ステップ105で生成した組の`target`頂点を自サーバ装置が処理する場合は生成した組の情報を`Next`キュー616に入れ、他サーバが処理する場合は生成した組の情報を送信バッファ611へ入れる。送信バッファ611は、`target`頂点を処理するサーバ装置毎に分けておくことが望ましい。

[0022] ステップS107では、送信情報をまとめて効率よく転送するために、各サーバ装置で、送信バッファ611に格納された情報の量が予め設定された送信基準サイズを超えたか否かの判定が行われる。基準のサイズを超えた場合は送信を開始するために、各サーバ装置の動作はステップS200へ進み、基準のサイズを超えていない場合はステップS103へ戻る。ステップ108では、各サーバ装置で、処理キュー613が空になった状態で未送信の情報が送信バッファ611に残っているか否かの判定が行われる。送信バッファ611が空でない場合は未送信情報を送信するために、各サーバ装置の動作はステップS200へ進み、送信バッファが空の場合は未送信情報がないため、受信を開始するためにステップS300へ進む。

[0023] 図11に、ステップS200の詳細なステップであるステップS201ーS205を示す。まず、ステップS200では、各サーバ装置の圧縮モジュール617が、送信バッファ611内の`target`頂点、`source`頂点、`data`のカラム毎に共通な値の数をカウントアップする（ステップS201、ステップS202）。その後、各サーバ装置の圧縮モジュール617は、各カラムで共通な値は連続するものとしてランレングス符号化を行った場合に最も圧縮率の高いカラムをキーとして送信バッファ611内のデータのソートを行う（ステップS203）。この際、`target`頂点、`source`頂点、`data`の組の対応は保持した形でソートが行われる。このように、送信バッファ611に蓄積されている中間データである`target`

t 頂点、source 頂点、data の組の群の中で、組の並びのソートが行われることで、効率的な送信データの圧縮が可能となり、効率的なデータ転送を実現でき、ひいてはグラフ処理を高速化できる。さらに、target 頂点、source 頂点、data の内で最も圧縮率が高くなるカラムをキーとすることで、さらに効率的な送信データの圧縮が可能となる。なお、target 頂点、source 頂点のカラム毎に共通な値の数をカウントアップし、target 頂点またはsource 頂点のいずれか圧縮率の高い方のカラムをキーとしてソートしてもよい。例えば、辺に重みが無いグラフ処理の場合には、data の送信は不要である。

[0024] 各サーバ装置の圧縮モジュール617は、ソートを行った後、キーとしたカラムに対しランレングス符号化を行う（ステップS204）。予めソートがおこなわれることで、共通の数値が連続するようになり、ランレングス符号化での圧縮率を高めることができる。また、伸長モジュール618による復号時にどのカラムをキーとして圧縮を行ったかが分かるように、各サーバ装置の圧縮モジュール617は、送信情報の先頭にキーとしたカラムを示す情報を付加する。ステップS200に続くステップS109では、各サーバ装置は、ステップS200で処理された送信データを、頂点割当て情報615を参照し、Target 頂点の演算を割当てられているサーバ装置へ送信する。

[0025] 図12に、ステップS300の詳細なステップであるステップS301ーS306を示す。伸長処理ステップS300では、各サーバ装置の伸長モジュール618が、まず受信バッファ612から情報を取り出し、その先頭の情報からどのカラムに復号を行えば良いかの特定を行う（ステップS301、S302）。伸長モジュール618は、特定したカラムに対し、ランレングス複合化を行い、もとの情報に伸長する（ステップS303）。ステップS304では、伸長モジュール618は、特定したカラムがtarget 頂点か否かの判定を行う。特定したカラムがtarget 頂点でない場合は、伸長モジュール618は、target 頂点をキーとしてソートを行う（ス

トップS305)。この際も、target頂点、source頂点、dataの組の対応は保持した形でソートが行われる。このように、受信バッファ612に到着した中間データの群の中で、中間データの並びのソートが行われる。特定したカラムがtarget頂点の場合は、送信前に既にソート済みであるためソートは行われず処理が終了する(S306)。

[0026] ステップS110では、各サーバ装置は、別のサーバ装置から送信された情報を受信バッファ612から取り出してNextキュー616へ入れる。各サーバ装置は、受信する情報が無い場合はステップS110で何もしない。ここには記載しないが、受信すべき情報があるか否かは別途サーバ装置間で通信し合う。その後ステップS111では、各サーバ装置は、グラフ処理を行っている全サーバ装置でNextキュー616が空か否かの判定を行う。グラフ処理を行っている全サーバ装置でNextキュー616が空か否かは、サーバ装置のそれぞれが、自身のNextキュー616が空になった場合に、他のサーバ装置に自身のNextキュー616が空になったことを通知することで、各サーバ装置が判定できる。Nextキュー616が空でない場合は、各サーバ装置は、Nextキュー616から処理キュー613へ情報を移動させ(ステップS112)、各サーバ装置の動作はステップS103へ戻る。全サーバ装置のNextキュー616が空の場合は、全ての頂点に対する処理が完了する(ステップS113)。

[0027] ステップS200とステップS300の圧縮・伸長の処理により、受信した情報はtarget頂点をキーとしてソートされた状態で処理キュー613へ入れられる。処理キュー613への格納順は頂点に対する各サーバ装置の演算順となるため、target頂点順としておくことで、同じtarget頂点に対する処理が連続に行われる。連続に処理が行われることで、中間データがCPU600のキャッシュからメモリ装置610へリプレイスされたり、メモリ装置610からストレージ装置620へのスワップが発生したりする確率を下げることができる。以降、具体例を用いて圧縮・伸長動作について説明する。

[0028] 図13は、図5のグラフの各頂点がどのサーバ装置で処理されるかの例を示した図である。この図では頂点1～10をグループ700、頂点11～15をグループ710、頂点16～25をグループ720に分割している。本実施例では、グループ700はサーバ装置420で、グループ710はサーバ装置430で、グループ720はサーバ装置440で処理する場合を考える。

[0029] 図14は、図13に対し、頂点3を始点としてグラフ解析を行った場合の頂点間の情報の流れを矢印で示した図である。以下、頂点3を始点としたグラフ解析について処理の流れを説明する。まず、サーバ装置420のグラフ演算モジュール614が、頂点3をtarget頂点として選択し、(target, source, data) = (3, 3, Z) (sourceおよびdataはダミーデータ) を処理キュー613へ入れる。サーバ装置420のグラフ演算モジュール614は、処理キュー613から(target, source, data) = (3, 3, Z) を取り出し、頂点3に対する演算を実行する。サーバ装置420のグラフ演算モジュール614は、頂点3をsource頂点としてグラフ構造データを参照し、次に処理すべきtarget頂点が頂点7、8、9であることを得る。頂点7、8、9は全てグループ700に属しているため、サーバ装置420のグラフ演算モジュール614は、自身のNextキュー616に(target, source, data) = (7, 3, c)、(8, 3, d)、(9, 3, e)を入れる。送受信される情報は存在しないため、サーバ装置420のグラフ演算モジュール614は、Nextキュー616の内容を処理キュー613へ移動させ、頂点7、8、9について演算を行い、グラフ構造データを参照する。各頂点が所属するグループの分類により、サーバ装置420のグラフ演算モジュール614は、頂点7ではNextキュー616へ(target, source, data) = (1, 7, a)、(2, 7, b)、送信バッファ611へ(target, source, data) = (11, 7, i)、頂点8では送信バッファ611へ(target, source, data)

= (11, 8, j)、頂点9ではNextキュー616へ(target, source, data) = (4, 9, f)、送信バッファ611へ(target, source, data) = (12, 9, l)、(11, 9, k)を入れる。サーバ装置420の圧縮モジュール617は、処理キュー613が空になったため、送信バッファ611の内容を圧縮して転送を行う。

[0030] 図15(a)に送信バッファ611の内容のソート前の状態、図15(b)に送信バッファ611の内容をソート後の状態、図15(c)に送信バッファ611の内容をソートし圧縮した後の状態をそれぞれ示す。送信開始に先立ち、サーバ装置420の圧縮モジュール617は、(target, source, data)の各カラムをソートしてランレングス符号化を行った場合に最も圧縮率の高いものを選択する。ここでは、targetが頂点11に対し3個共通、sourceが頂点9に対し2個共通、dataは共通なしであり、targetをロウ方向に圧縮することが選択される。その後、サーバ装置420の圧縮モジュール617は、ロウ方向のソートを行い(図15(b))、targetに対してランレングス符号化を行う(図15(c))。図15(c)では頂点11が3個共通ということを示す11x3と表現しているが、これを例えば、最上位ビットが1となっている数を繰り返し数として表現することができる。target、source、data、繰り返し数をそれぞれ4バイトの変数で表現した場合、圧縮前のサイズが48バイトに対し、圧縮後のサイズは44バイトとなる。さらに転送情報の先頭にtargetを圧縮したという情報が1バイト付加され、45バイトの情報がサーバ装置430へ送信される。サーバ装置430では、受理した情報の伸長が行われ、伸長されたデータがNextキュー616へ入れられる。サーバ装置420、430では、Nextキュー616から処理キュー613へ(target, source, data)情報が移動され、別の頂点についても処理が進められる。以後、各頂点に対する上述の処理の繰り返しなので、説明は省略する。

[0031] なお、各頂点に対する演算は上述のように処理キュー613から取り出さ

れた `target` 頂点の順に行われるが、同じ `target` 頂点に対する演算が連続に行われない場合、他の `target` 頂点への演算により、中間データがキャッシュからメモリ装置にリプレイスされたり、メモリ装置からストレージ装置へのスワップが発生したりすることで処理性能が低下してしまう。そこで、`source` または `data` で圧縮が行われた場合は、受信側で伸長を行った後に、`target` でソートが行われるフローとしている。

[0032] 以上のように、本実施例の転送方法では、グラフ頂点間で転送される中間データの群の中で、中間データの並びのソートが行われることで、効率的な送信データの圧縮が可能となり、効率的なデータ転送を実現でき、ひいてはグラフ処理を高速化できる。さらに、各カラムをロウ方向ソートして圧縮した場合の圧縮率の比較を行うことで、サーバ間で転送される情報の量をさらに効率よく削減することができる。また、カラムのうち、`source` または `data` をキーとしてソートし、圧縮をかけて転送した場合には、受信側で伸長後に `target` で再ソートを行うことでメモリ装置やストレージ装置に待避する中間データ量を減らし、グラフ解析をさらに高速に行うことが可能になる。

符号の説明

[0033] 100 : クライアント、200 : インターネット、300 : Webサーバ、400 : Apサーバ、500 : DBサーバ、410 : ネットワーク装置、420 ~ 450 : サーバ装置、600 : CPU、610 : メモリ装置、611 : 送信バッファ、612 : 受信バッファ、613 : 処理キュー、614 : グラフ演算モジュール、615 : 頂点割当て情報、616 : NEXTキュー、617 : 圧縮モジュール、618 : 伸長モジュール、620 : ストレージ装置、630 : 入力装置、640 : 出力装置、650 : ネットワークインタフェース、660 : バス

請求の範囲

- [請求項1] 複数の計算ノードによるグラフ処理での中間データの転送方法であって、
- 各計算ノードは送信バッファを有し、
 - 各計算ノードには、処理対象のグラフ頂点が割当てられ、
 - 前記中間データには、送信元のグラフ頂点の情報と送信先のグラフ頂点の情報の組が含まれ、
 - 前記中間データを前記送信バッファに蓄積し、
 - 前記送信元のグラフ頂点の情報または前記送信先のグラフ頂点の情報に基づいて、蓄積された前記中間データの群の中の前記中間データの並びをソートし、
 - 前記中間データの群を転送することを特徴とするグラフ処理の中間データの転送方法。
- [請求項2] 請求項1に記載のグラフ処理の中間データの転送方法において、
- 各計算ノードは受信バッファを有し、
 - 各計算ノードは、
 - 転送された前記中間データの群が前記送信元のグラフ頂点の情報に基づいてソートされている場合には、
 - 前記送信先のグラフ頂点の情報に基づいて、前記受信バッファに到着した前記中間データの群の中の前記中間データの並びをソートすることを特徴とするグラフ処理の中間データの転送方法。
- [請求項3] 請求項1に記載のグラフ処理の中間データの転送方法において、
- 前記ソート後に、蓄積された前記中間データの群を圧縮することを特徴とするグラフ処理の中間データの転送方法。
- [請求項4] 請求項3に記載のグラフ処理の中間データの転送方法において、
- 前記送信元のグラフ頂点の情報または前記送信先のグラフ頂点の情報に基づいて、蓄積された前記中間データの群の中の前記中間データの並びをソートする際に、

前記送信元のグラフ頂点の情報または前記送信先のグラフ頂点の情報のいずれかソート後に圧縮率が高くなる方に基づいて、ソートを行うことを特徴とするグラフ処理の中間データの転送方法。

[請求項5] 請求項1に記載のグラフ処理の中間データの転送方法において、
前記計算ノードは、サーバ装置であることを特徴とするグラフ処理の中間データの転送方法。

[請求項6] 複数の計算ノードによるグラフ処理での中間データの転送方法であって、

各計算ノードは送信バッファを有し、
各計算ノードには、処理対象のグラフ頂点が割当てられ、
前記中間データには、送信元のグラフ頂点の情報、送信先のグラフ頂点の情報、および前記送信元のグラフ頂点と前記送信先のグラフ頂点の間の依存関係の情報の組が含まれ、
前記中間データを前記送信バッファに蓄積し、
前記送信元のグラフ頂点の情報、前記送信先のグラフ頂点の情報、または前記依存関係の情報に基づいて、蓄積された前記中間データの群の中の前記中間データの並びをソートし、
前記中間データの群を転送することを特徴とするグラフ処理の中間データの転送方法。

[請求項7] 請求項6に記載のグラフ処理の中間データの転送方法において、
各計算ノードは受信バッファを有し、
各計算ノードは、
転送された前記中間データの群が前記送信元のグラフ頂点の情報または前記依存関係の情報に基づいてソートされている場合には、
前記送信先のグラフ頂点の情報に基づいて、前記受信バッファに到着した前記中間データの群の中の前記中間データの並びをソートすることを特徴とするグラフ処理の中間データの転送方法。

[請求項8] 請求項6に記載のグラフ処理の中間データの転送方法において、

前記ソート後に、蓄積された前記中間データの群を圧縮することを特徴とするグラフ処理の中間データの転送方法。

[請求項9] 請求項8に記載のグラフ処理の中間データの転送方法において、
前記送信元のグラフ頂点の情報、前記送信先のグラフ頂点の情報、
または前記依存関係の情報に基づいて、蓄積された前記中間データの群の中の前記中間データの並びをソートする際に、

前記送信元のグラフ頂点の情報、前記送信先のグラフ頂点の情報、
または前記依存関係の情報のいずれかソート後に圧縮率が最も高くなる情報に基づいて、ソートを行うことを特徴とするグラフ処理の中間データの転送方法。

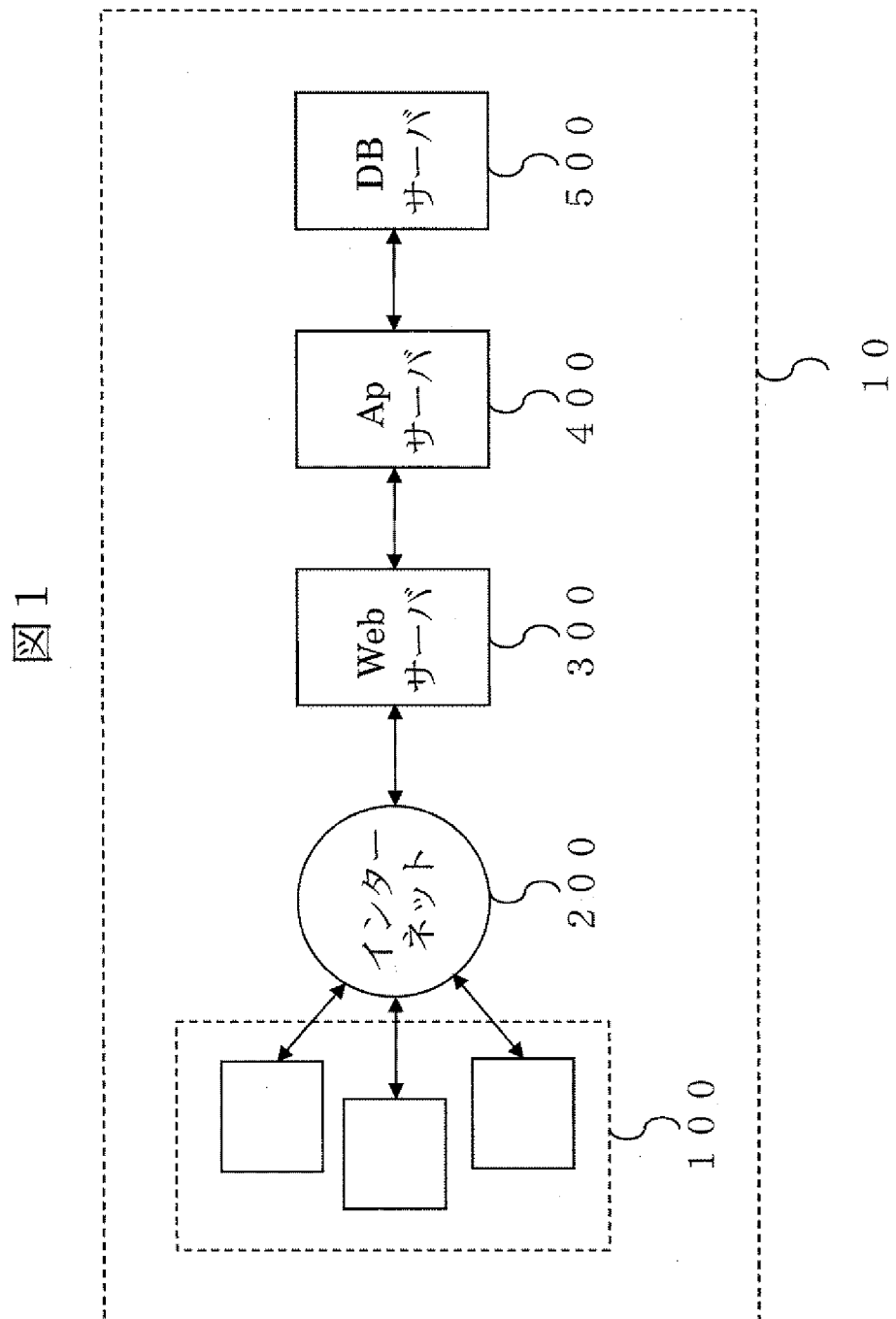
[請求項10] 請求項6に記載のグラフ処理の中間データの転送方法において、
前記計算ノードは、サーバ装置であることを特徴とするグラフ処理の中間データの転送方法。

[請求項11] 複数の計算ノードを有するグラフ処理システムであって、
各計算ノードは、
送信元のグラフ頂点の情報と送信先のグラフ頂点の情報の組を有するグラフ処理の中間データを保存する送信バッファと、
前記送信元のグラフ頂点の情報または前記送信先のグラフ頂点の情報に基づいて、前記送信バッファに蓄積された前記中間データの群の中の前記中間データの並びをソートするモジュールと、
前記中間データの群を転送するモジュールとを有することを特徴とするグラフ処理システム。

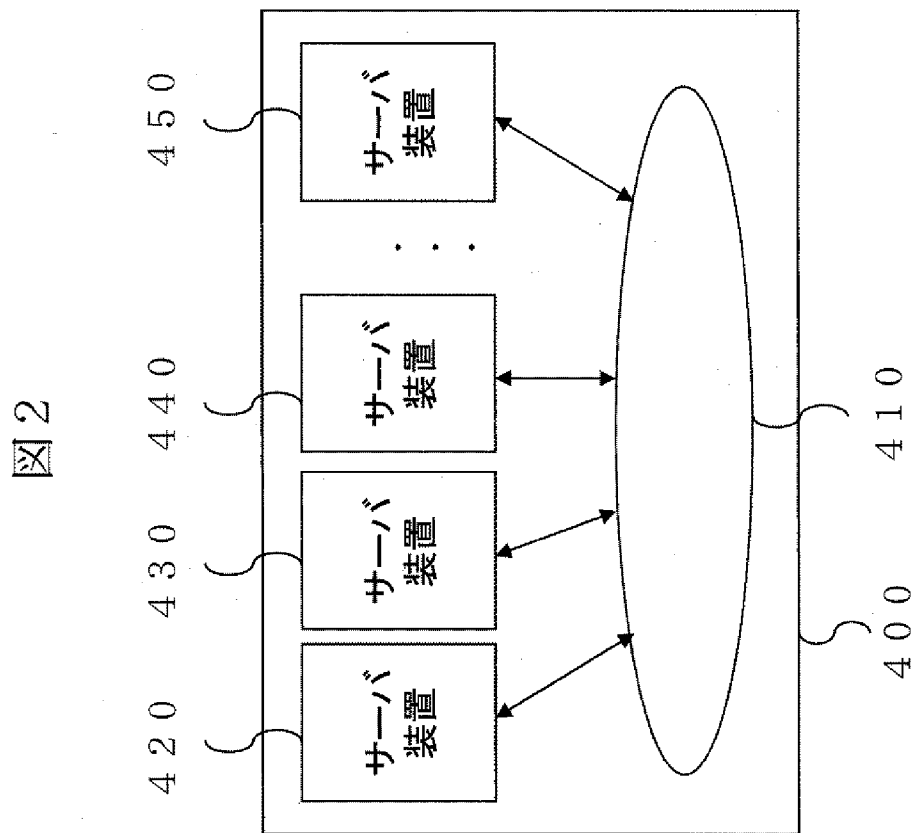
[請求項12] 請求項11に記載のグラフ処理システムにおいて、
各計算ノードは、
受信バッファと、
前記送信先のグラフ頂点の情報に基づいて、前記受信バッファに到着した前記中間データの群の中の前記中間データの並びをソートするモジュールとを有することを特徴とするグラフ処理システム。

[請求項13] 請求項 1 1 に記載のグラフ処理システムにおいて、
前記計算ノードは、サーバ装置であることを特徴とするグラフ処理
システム。

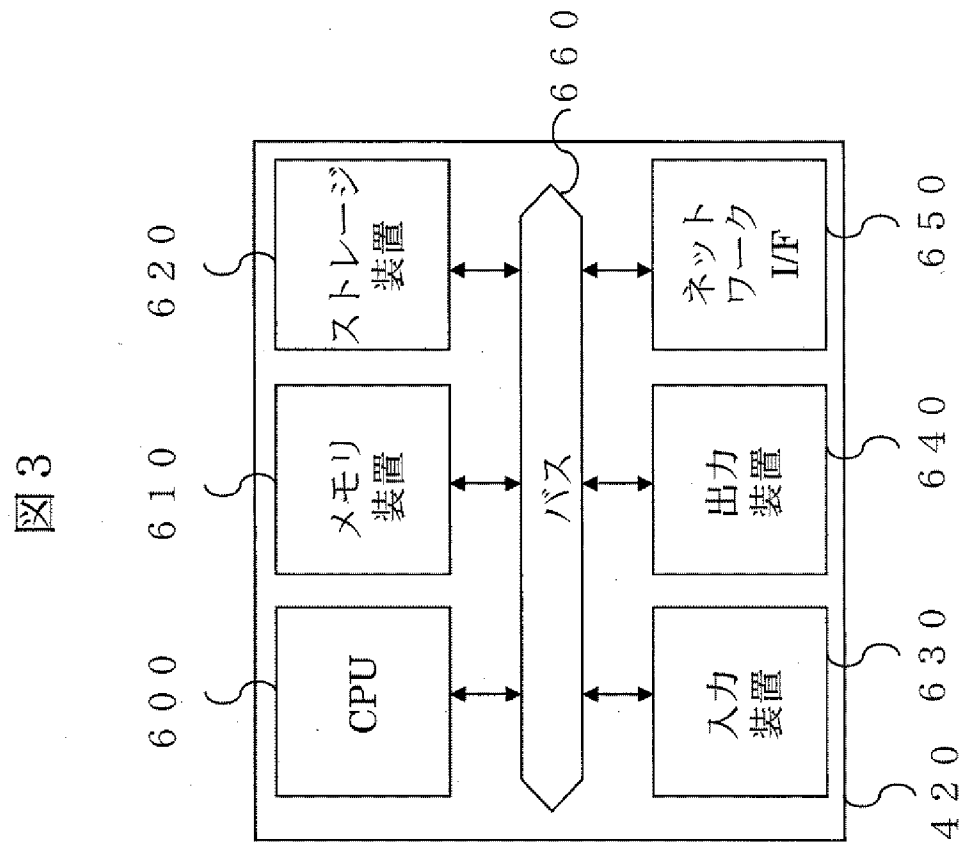
[図1]



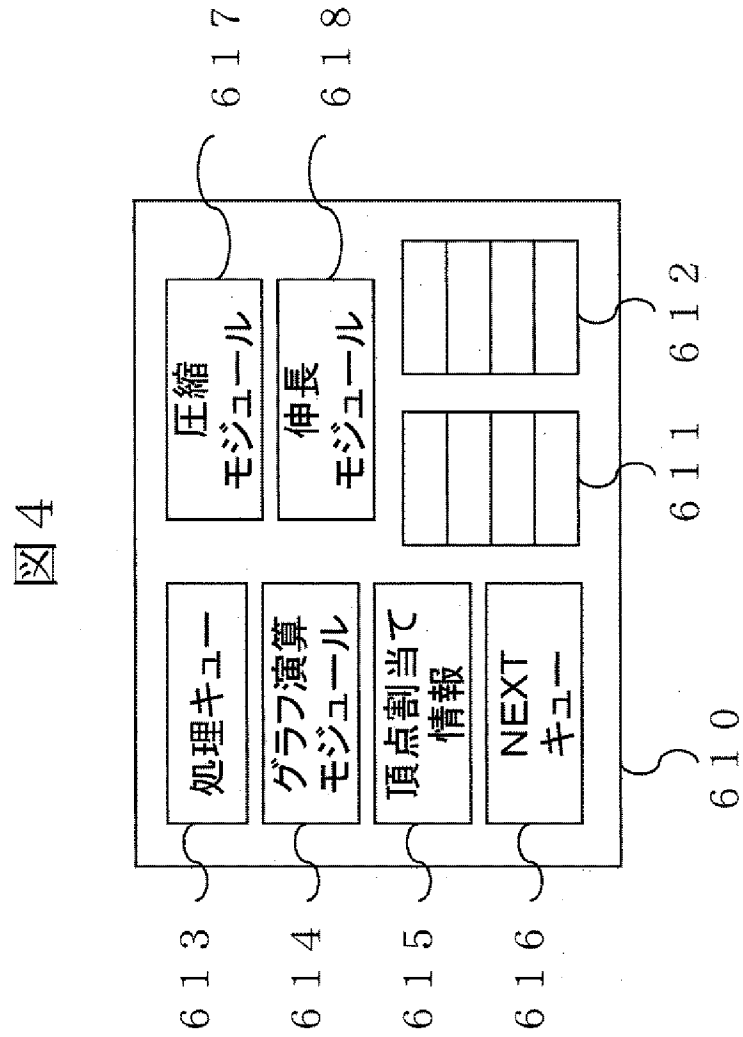
[図2]



[図3]

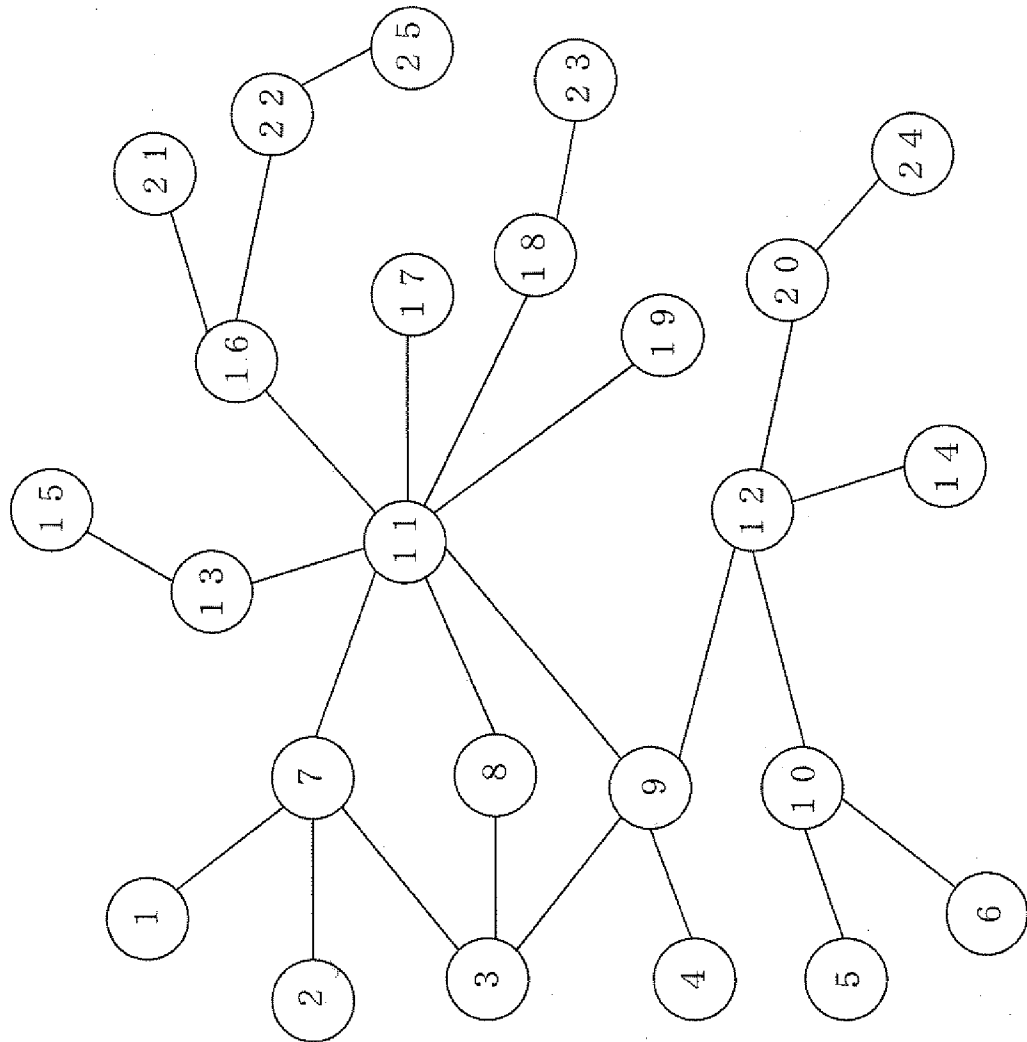


[図4]



[図5]

図5



[図6]

図6

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1		0	0	0	0	0	a	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0		0	0	0	0	b	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0		0	0	0	c	d	e	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0		0	0	0	0	f	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0		0	0	0	g	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0		0	0	h	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	a	b	c	0	0	0		0	0	0	i	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	d	0	0	0	0		0	0	j	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	e	f	0	0	0	0		0	k	l	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	g	h	0	0	0		m	n	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	i	j	k	0		0	0	s	0	0	p	q	r	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	l	m	0	0		0	0	0	0	0	0	t	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	n	0	0	0		u	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	s	0	0	0		0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	u	0	0		0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	v	w	0	0	0
17	0	0	0	0	0	0	0	0	0	0	p	0	0	0	0	0	0		0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0		0	0	x	0	0	0
19	0	0	0	0	0	0	0	0	0	0	r	0	0	0	0	0	0	0	0		0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	t	0	0	0	0	0	0	0	0		0	y	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	v	0	0	0	0	0	0		0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	w	0	0	0	0	0	0	0		z
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	x	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	y	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	z	0	0	0

[図7]

図7

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
values	a	b	c	d	e	f	g	h	a	b	c	i	d	j	e	f	k	l	g	h
columns	7	7	7	8	9	9	10	10	1	2	3	11	3	11	3	4	11	12	5	6

#	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
values	m	i	j	k	n	o	p	q	r	l	m	s	t	n	u	s	u	o	v	w
columns	12	7	8	9	13	16	17	18	19	9	10	14	20	11	15	12	13	11	21	22

#	41	42	43	44	45	46	47	48	49	50	51	52
values	p	q	x	r	t	y	v	w	z	x	y	z
columns	11	11	23	11	12	24	16	16	25	18	20	22

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
rowindex	1	2	3	6	7	8	9	13	15	19	22	30	34	36	37	38	41	42	44	45

#	21	22	23	24	25	26
rowindex	47	48	50	51	52	53

[図8]

図 8

#	target	source	data
1	7	3	c
2	8	3	d
3	9	3	e

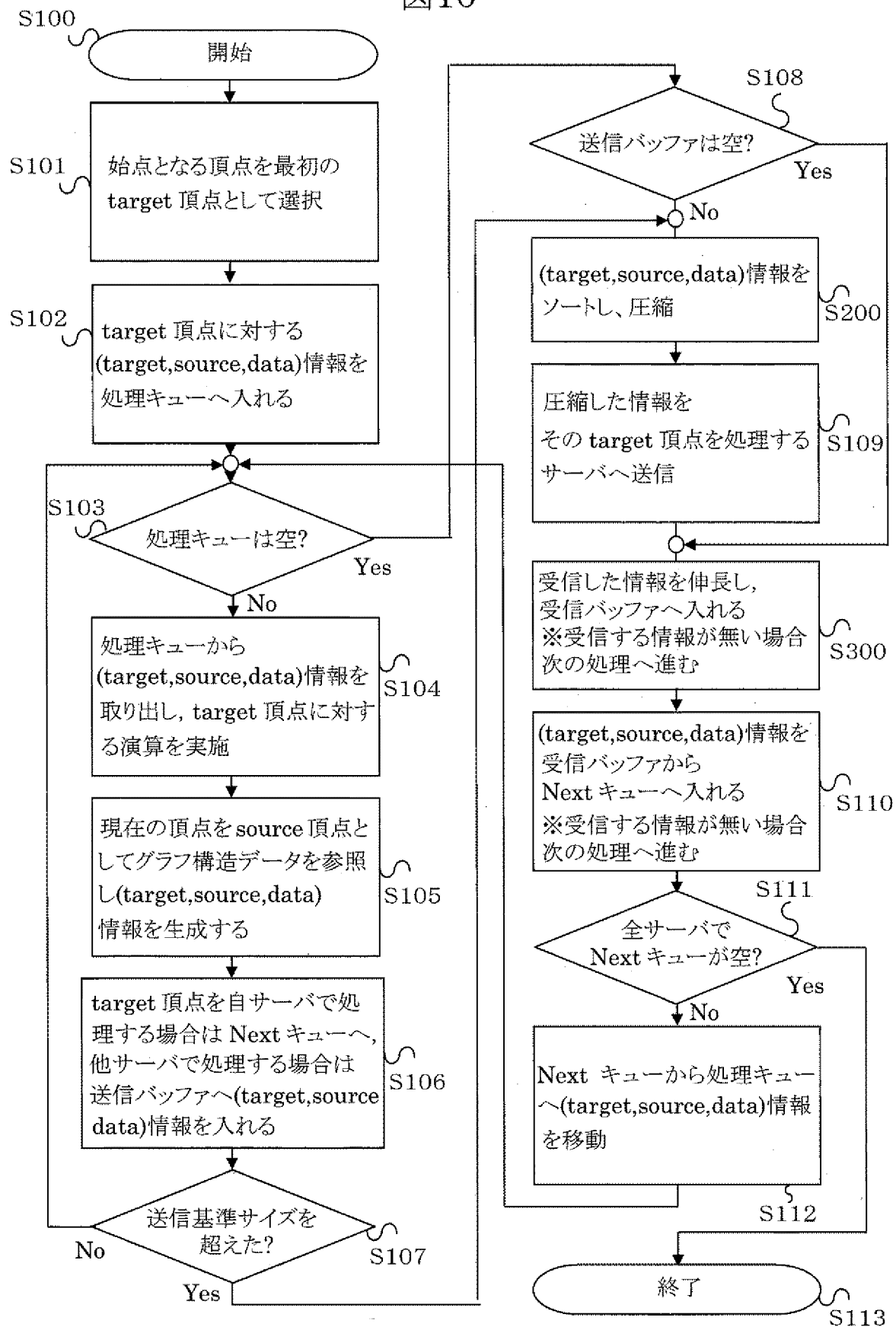
[図9]

図9

頂点番号	サーバID
1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	1
11	2
12	2
13	2
14	2
15	2
16	3
17	3
18	3
19	3
20	3
21	3
22	3
23	3
24	3
25	3

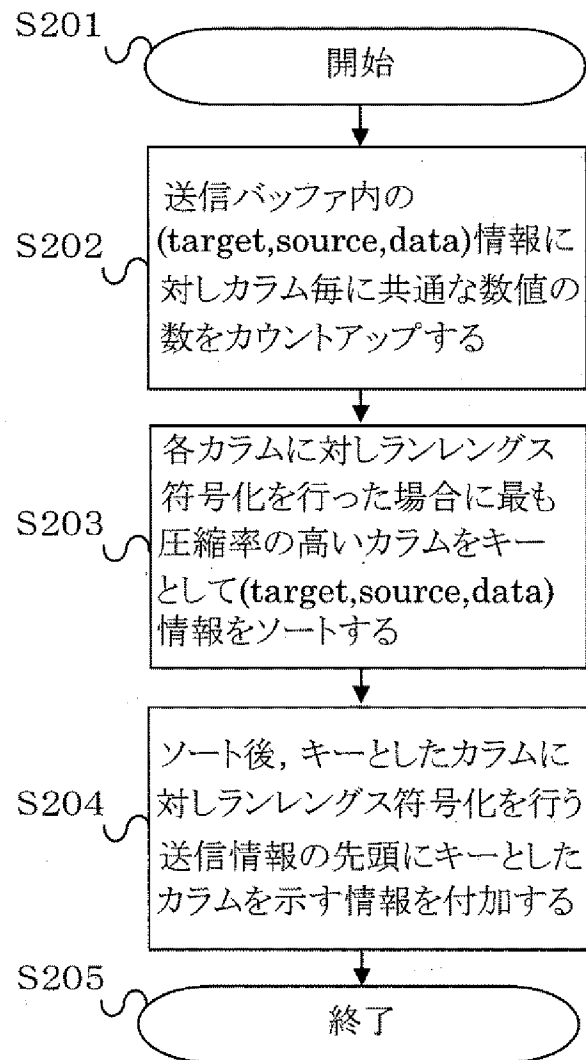
[図10]

図10



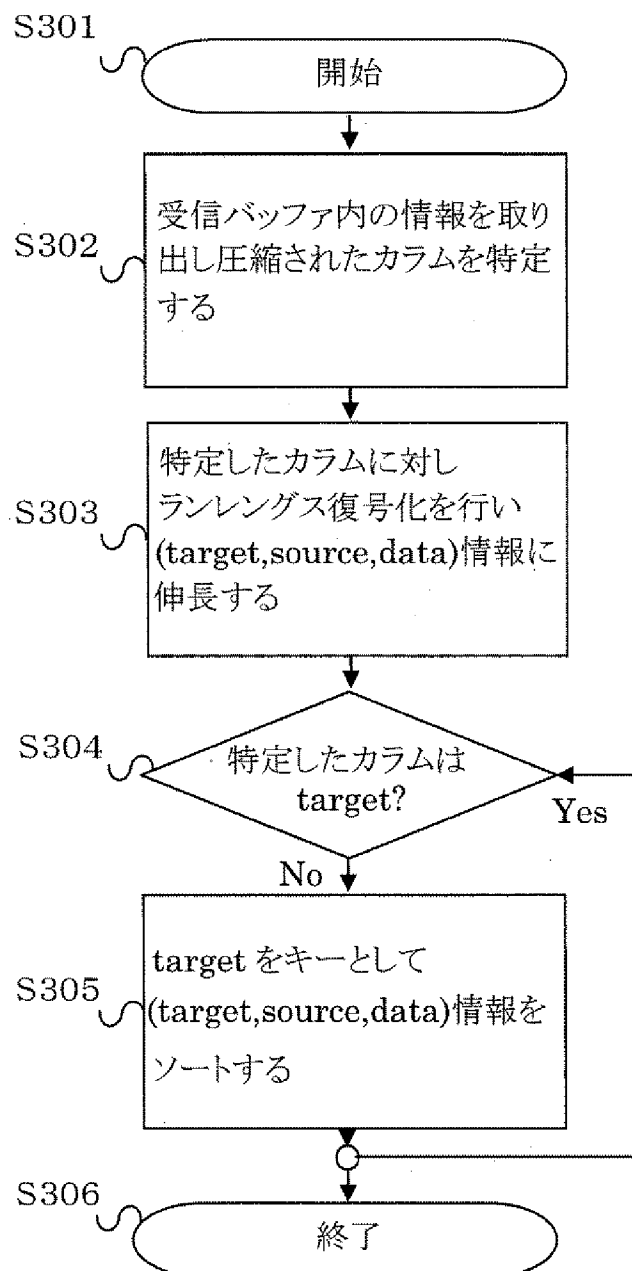
[図11]

図11



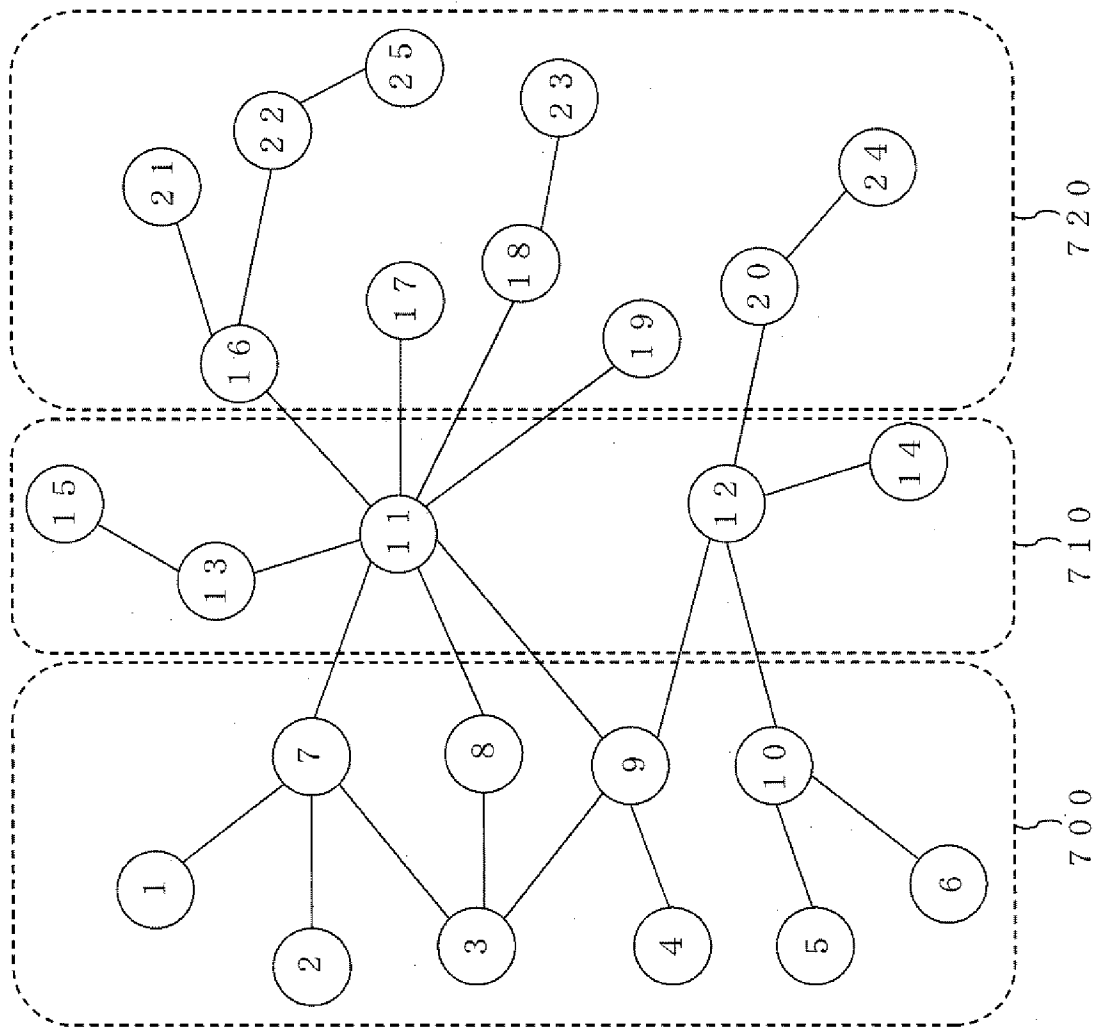
[図12]

図12

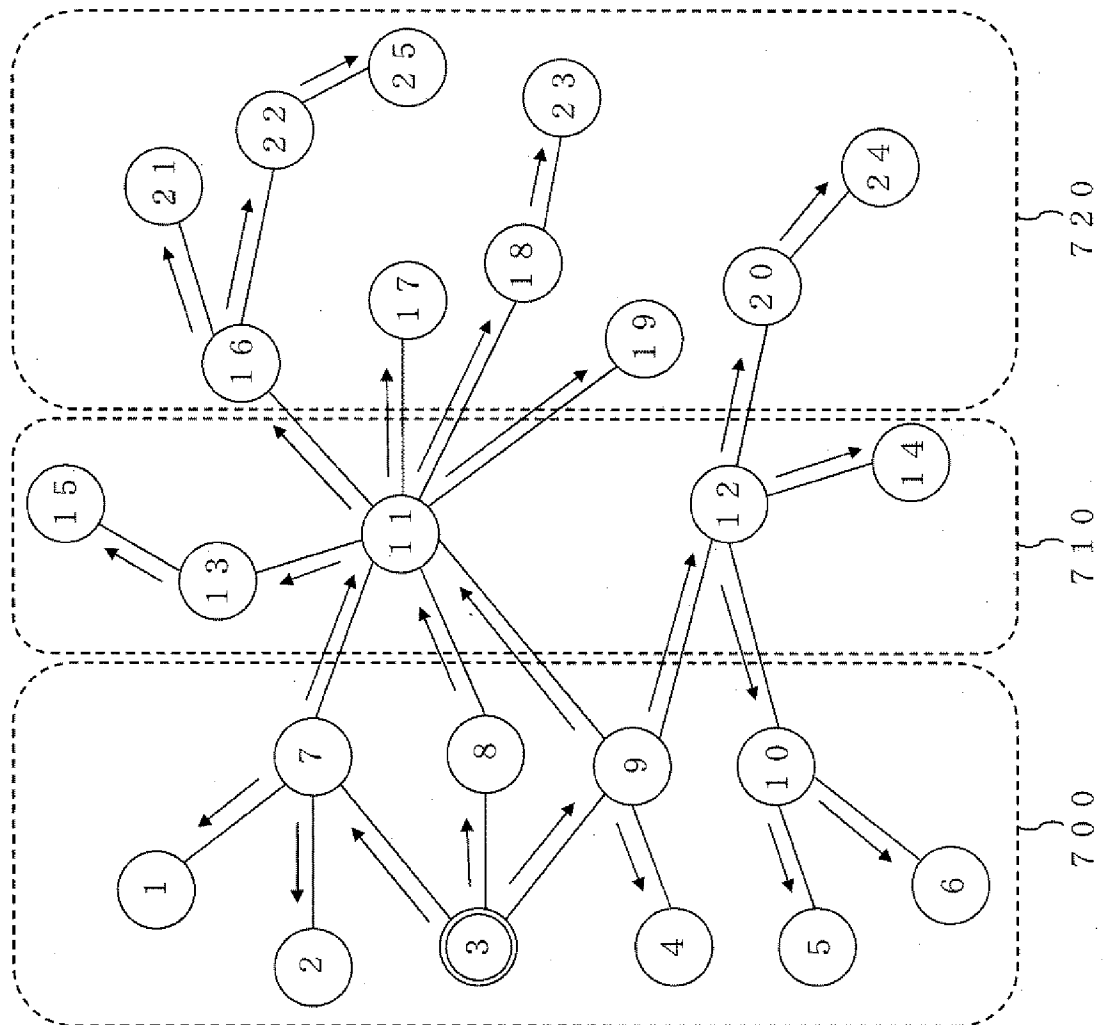


[図13]

図13



[図14]



[図15(a)]

図 1 5 (a)

(a)送信バッファ内情報

#	target	source	data
1	11	7	i
2	11	8	j
3	12	9	l
4	11	9	k

[図15(b)]

図 1 5 (b)

(b)ソート後の状態

#	target	source	data
1	11	7	i
2	11	8	j
3	11	9	k
4	12	9	l

[図15(c)]

図 1 5 (c)

(c)圧縮後の状態

#	target	source	data
1	11x3	7	i
2		8	j
3		9	k
4	12	9	l

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2012/067270

A. CLASSIFICATION OF SUBJECT MATTER

G06F13/00 (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)

G06F13/00, G06F9/46, G06F15/16

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 1996-2012

Kokai Jitsuyo Shinan Koho 1971-2012 Toroku Jitsuyo Shinan Koho 1994-2012

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y A	JP 5-504644 A (Thinking Machines Corp.), 15 July 1993 (15.07.1993), page 12, upper left column, line 12 to lower left column, line 13; page 19, lower right column, line 15 to page 21, lower left column, line 12; page 21, lower right column, line 23 to page 22, upper left column, line 6; page 25, upper left column, line 24 to lower right column, line 13; fig. 1, 7A, 7B	1, 3, 5, 6, 8, 10, 11, 13 2, 4, 7, 9, 12
Y A	JP 2003-338830 A (Matsushita Electric Industrial Co., Ltd.), 28 November 2003 (28.11.2003), paragraphs [0045] to [0051]	1, 3, 5, 6, 8, 10, 11, 13 2, 4, 7, 9, 12

☒ 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
06 November, 2012 (06.11.12)Date of mailing of the international search report
13 November, 2012 (13.11.12)Name and mailing address of the ISA/
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2012/067270

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 2010-244563 A (Ab Initio Software LLC), 28 October 2010 (28.10.2010), entire text; all drawings	1-13

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/JP2012/067270

JP 5-504644 A	1993.07.15	US 5247694 A1	1993.09.21
		WO 1991/020024 A1	1991.12.26
		EP 533835 A1	1993.03.31
		AU 8184791 A	1992.01.07
		AU 667797 B	1996.04.18
		CA 2083255 A1	1991.12.15
JP 2003-338830 A	2003.11.28	JP 2003-338817 A	2003.11.28
		US 2005/0018615 A1	2005.01.27
		WO 2003/077485 A2	2003.09.18
		EP 1491009 A2	2004.12.29
		AU 2003212650 A	2003.09.22
JP 2010-244563 A	2010.10.28	JP 2006-504160 A	2006.02.02
		US 2004/0073529 A1	2004.04.15
		US 2007/0022077 A1	2007.01.25
		US 2007/0179923 A1	2007.08.02
		WO 2004/034322 A1	2004.04.22
		EP 1559069 A1	2005.08.03
		EP 2284774 A1	2011.02.16
		CA 2501876 A	2004.04.22
		AU 2003279228 A	2004.05.04
		AU 2007202782 A	2007.07.05

A. 発明の属する分野の分類（国際特許分類（I P C））

Int.Cl. G06F13/00(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（I P C））

Int.Cl. G06F13/00, G06F9/46, G06F15/16

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	1 9 2 2 - 1 9 9 6 年
日本国公開実用新案公報	1 9 7 1 - 2 0 1 2 年
日本国実用新案登録公報	1 9 9 6 - 2 0 1 2 年
日本国登録実用新案公報	1 9 9 4 - 2 0 1 2 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）

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
Y A	JP 5-504644 A（スインキング マシンズ コーポレーション） 1993.07.15, 12 頁左上欄 12 行－左下欄 13 行, 19 頁右下欄 15 行－ 21 頁左下欄 12 行, 21 頁右下欄 23 行－22 頁左上欄 6 行, 25 頁左上 欄 24 行－右下欄 13 行, 図 1, 7A, 7B	1, 3, 5, 6, 8, 10 , 11, 13 2, 4, 7, 9, 12
Y A	JP 2003-338830 A（松下電器産業株式会社） 2003.11.28, 段落【0045】－【0051】	1, 3, 5, 6, 8, 10 , 11, 13 2, 4, 7, 9, 12

☒ 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

0 6 . 1 1 . 2 0 1 2

国際調査報告の発送日

1 3 . 1 1 . 2 0 1 2

国際調査機関の名称及びあて先

日本国特許庁（I S A / J P）

郵便番号 1 0 0 - 8 9 1 5

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）

寺谷 大亮

電話番号 0 3 - 3 5 8 1 - 1 1 0 1 内線 3 5 6 8

5 T

9 8 5 1

C (続き) . 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 2010-244563 A (アビニシオ ソフトウェア エルエルシー) 2010.10.28, 全文, 全図	1-13

JP 5-504644 A	1993. 07. 15	US 5247694 A1	1993. 09. 21
		WO 1991/020024 A1	1991. 12. 26
		EP 533835 A1	1993. 03. 31
		AU 8184791 A	1992. 01. 07
		AU 667797 B	1996. 04. 18
		CA 2083255 A1	1991. 12. 15
JP 2003-338830 A	2003. 11. 28	JP 2003-338817 A	2003. 11. 28
		US 2005/0018615 A1	2005. 01. 27
		WO 2003/077485 A2	2003. 09. 18
		EP 1491009 A2	2004. 12. 29
		AU 2003212650 A	2003. 09. 22
JP 2010-244563 A	2010. 10. 28	JP 2006-504160 A	2006. 02. 02
		US 2004/0073529 A1	2004. 04. 15
		US 2007/0022077 A1	2007. 01. 25
		US 2007/0179923 A1	2007. 08. 02
		WO 2004/034322 A1	2004. 04. 22
		EP 1559069 A1	2005. 08. 03
		EP 2284774 A1	2011. 02. 16
		CA 2501876 A	2004. 04. 22
		AU 2003279228 A	2004. 05. 04
		AU 2007202782 A	2007. 07. 05