(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第5535230号
(P5535230)

(45)発行日　平成26年7月2日(2014.7.2)　(24)登録日　平成26年5月9日(2014.5.9)

(51)Int.Cl.		F I		
G06N 3/00	(2006.01)	G06N	3/00	560A
G06N 7/02	(2006.01)	G06N	7/02	554D
G06F 17/30	(2006.01)	G06F	17/30	210D

請求項の数 43　(全 54 頁)

(21)出願番号	特願2011-533380 (P2011-533380)
(86)(22)出願日	平成21年10月23日 (2009.10.23)
(65)公表番号	特表2012-507085 (P2012-507085A)
(43)公表日	平成24年3月22日 (2012.3.22)
(86)国際出願番号	PCT/US2009/061899
(87)国際公開番号	WO2010/048538
(87)国際公開日	平成22年4月29日 (2010.4.29)
審査請求日	平成24年10月23日 (2012.10.23)
(31)優先権主張番号	61/107,971
(32)優先日	平成20年10月23日 (2008.10.23)
(33)優先権主張国	米国(US)

(73)特許権者　509123208
アビニシオ　テクノロジー　エルエルシー
アメリカ合衆国　０２４２１　マサチューセッツ州　レキシントン　スプリング　ストリート　２０１
(74)代理人　100079108
弁理士　稲葉　良幸
(74)代理人　100109346
弁理士　大貫　敏史
(72)発明者　アンダーソン　アーレン
イギリス　オーエックスファイブ　ツーエスジー　キドリントン　イスリップ　ロウワー　ストリート　３

審査官　塚田　肇

最終頁に続く

(54)【発明の名称】ファジーなデータ操作

(57)【特許請求の範囲】

【請求項１】

データ記憶システムに格納されたデータ要素をクラスタ化する方法であって、

上記データ記憶システムからデータ要素を読み取るステップ、

データ要素のクラスタを形成するステップであって、各データ要素が少なくとも１つのクラスタのメンバであり、且つ、所定のデータ要素における１つ以上のオブジェクトに少なくとも部分的に基づいて、上記所定のデータ要素が所定のクラスタのメンバとして割り付けられる、ステップ、

少なくとも１つのデータ要素を２つ以上のクラスタと関連付けるステップであって、上記２つ以上のクラスタのそれぞれに属する上記データ要素のメンバシップが、所定のクラスタにおける所定のデータ要素の内容と少なくとも１つの他のデータ要素の内容との間での比較に少なくとも部分的に基づいて、上記所定のクラスタにおける上記所定のデータ要素の部分的メンバシップを定量化する曖昧さの尺度によって表現され、上記所定のデータ要素の上記内容が、上記所定のデータ要素における１つ以上のオブジェクトにおける１つ以上の単語及び上記１つ以上の単語の変形の少なくとも１つを含む、ステップ、並びに

上記データ記憶システムに情報を格納して、上記形成されたクラスタを表現するステップ、

を含む方法。

【請求項２】

請求項１に記載の方法であって、

上記２つ以上のクラスタのそれぞれに属する上記データ要素のメンバシップを表現する上記曖昧さの尺度の各値が０と１との間にある、方法。
【請求項３】
請求項２に記載の方法であって、
上記メンバシップを表現する上記曖昧さの尺度の値が、上記２つ以上のクラスタのそれぞれに属する上記データ要素の確度に関連している、方法。
【請求項４】
請求項２に記載の方法であって、
上記２つ以上のクラスタのそれぞれに属する上記データ要素のメンバシップを表現する上記曖昧さの尺度の各値の合計が１である、方法。
【請求項５】
請求項４に記載の方法であって、
上記曖昧さの尺度の値を使用して会計の完全性を保つステップ、
を更に含む方法。
【請求項６】
請求項５に記載の方法であって、
所定の量についての会計の完全性を保つステップが、上記曖昧さの尺度の値によって上記量を重み付けするステップによって達成される、方法。
【請求項７】
請求項６に記載の方法であって、
上記メンバシップを表現する上記曖昧さの尺度の値を使用するデータ操作を実行するステップ、
を更に含む方法。
【請求項８】
請求項７に記載の方法であって、
上記データ操作が、上記１つ以上のクラスタの第１のクラスタ内の量の重み付けされた小計を計算するロールアップを含み、上記量が上記データ要素と関連付けられ、上記小計が、上記第１のクラスタ内で、上記第１のクラスタにおける上記データ要素の各々と関連付けられた上記量の値と上記第１のクラスタにおける上記データ要素の上記メンバシップを表現する上記曖昧さの尺度のそれぞれの値との積を合計することによって計算される、方法。
【請求項９】
請求項８に記載の方法であって、
上記量の排他的小計及び上記量の包括的小計を計算するステップであって、上記排他的小計が、２つ以上のクラスタと関連付けられる上記第１のクラスタにおける上記データ要素を排除することによって計算され、上記包括的小計が、２つ以上のクラスタと関連付けられる上記第１のクラスタにおける上記データ要素を包含することによって計算されるステップ、
を更に含む方法。
【請求項１０】
請求項２に記載の方法であって、
上記メンバシップを表現する上記曖昧さの尺度の値が関数に基づいて定められ、上記関数が、上記データ要素と上記２つ以上のクラスタとの間の関係を表現する、方法。
【請求項１１】
請求項１０に記載の方法であって、
上記関数によって表現される上記関係が、上記２つ以上のクラスタのそれぞれに属する上記データ要素の確度に関連している、方法。
【請求項１２】
請求項１０に記載の方法であって、
上記関数によって表現される上記関係が、上記データ要素と上記２つ以上のクラスタの

それぞれを表現する要素との間の定量化された類似性に基づく、方法。
【請求項１３】
請求項１２に記載の方法であって、
上記２つ以上のクラスタのそれぞれを表現する上記要素が、それぞれのクラスタのキーである、方法。
【請求項１４】
請求項１に記載の方法であって、
上記２つ以上のクラスタの各クラスタに属する上記データ要素の上記曖昧さの尺度の値が、各クラスタについて等しい、方法。
【請求項１５】
請求項１に記載の方法であって、
上記２つ以上のクラスタの各クラスタに属する上記データ要素の上記曖昧さの尺度の値が、参照集合における上記データ要素の観測度数に基づく、方法。
【請求項１６】
請求項１に記載の方法であって、
上記２つ以上のクラスタの各クラスタが、上記データ要素における異なる潜在的エラーを表現し、上記２つ以上のクラスタの各クラスタに属する上記データ要素の上記曖昧さの尺度の値が、各クラスタによって表現される上記データ要素における上記潜在的エラーの確度に基づく、方法。
【請求項１７】
請求項１に記載の方法であって、
データ・クラスタを形成するステップが、
データ要素の複数のスーパークラスタを形成するステップ、及び、
各スーパークラスタについて、上記スーパークラスタ内にデータ要素のクラスタを形成するステップ、
を含む、方法。
【請求項１８】
請求項１７に記載の方法であって、
各スーパークラスタを形成するステップが、
異なるデータ要素におけるオブジェクトの間での変形関係に基づいて、上記異なるデータ要素における上記オブジェクト間での整合を特定するステップ、
を含む、方法。
【請求項１９】
請求項１８に記載の方法であって、
第１のオブジェクトと第２のオブジェクトとの間の変形関係が、予め定められた閾値未満にある上記第１のオブジェクトと上記第２のオブジェクトとの間の距離を表現する関数の値に対応する、方法。
【請求項２０】
請求項１９に記載の方法であって、
上記変形関係が、同値関係ではない、方法。
【請求項２１】
請求項１７に記載の方法であって、
少なくとも１つのデータ要素が、１つより多くのスーパークラスタ中にある、方法。
【請求項２２】
データ記憶システムに格納されたデータ要素をクラスタ化するシステムであって、
上記データ記憶システムからデータ要素を読み取る手段、
データ要素のクラスタを形成する手段であって、各データ要素が少なくとも１つのクラスタのメンバであり、且つ、所定のデータ要素における１つ以上のオブジェクトに少なくとも部分的に基づいて、上記所定のデータ要素が所定のクラスタのメンバとして割り付けられる、手段、

少なくとも１つのデータ要素を２つ以上のクラスタと関連付ける手段であって、上記２つ以上のクラスタのそれぞれに属する上記データ要素のメンバシップが、所定のクラスタにおける所定のデータ要素の内容と少なくとも１つの他のデータ要素の内容との間での比較に少なくとも部分的に基づいて、上記所定のクラスタにおける上記所定のデータ要素の部分的メンバシップを定量化する曖昧さの尺度によって表現され、上記所定のデータ要素の上記内容が、上記所定のデータ要素における１つ以上のオブジェクトにおける１つ以上の単語及び上記１つ以上の単語の変形の少なくとも１つを含む、手段、及び

上記データ記憶システムに情報を格納して、上記形成されたクラスタを表現する手段、

を備えるシステム。

【請求項２３】

データ記憶システムに格納されたデータ要素をクラスタ化するためのコンピュータ・プログラムを格納するコンピュータ可読媒体であって、上記コンピュータ・プログラムが、

上記データ記憶システムからデータ要素を読み取るステップ、

データ要素のクラスタを形成するステップであって、各データ要素が少なくとも１つのクラスタのメンバであり、且つ、所定のデータ要素における１つ以上のオブジェクトに少なくとも部分的に基づいて、上記所定のデータ要素が所定のクラスタのメンバとして割り付けられる、ステップ、

少なくとも１つのデータ要素を２つ以上のクラスタと関連付けるステップであって、上記２つ以上のクラスタのそれぞれに属する上記データ要素のメンバシップが、所定のクラスタにおける所定のデータ要素の内容と少なくとも１つの他のデータ要素の内容との間での比較に少なくとも部分的に基づいて、上記所定のクラスタにおける上記所定のデータ要素の部分的メンバシップを定量化する曖昧さの尺度によって表現され、上記所定のデータ要素の上記内容が、上記所定のデータ要素における１つ以上のオブジェクトにおける１つ以上の単語及び上記１つ以上の単語の変形の少なくとも１つを含む、ステップ、及び

上記データ記憶システムに情報を格納して、上記形成されたクラスタを表現するステップ、

をコンピュータに実行させるための命令を含む、コンピュータ可読媒体。

【請求項２４】

請求項２２に記載のシステムであって、

上記２つ以上のクラスタのそれぞれに属する上記データ要素のメンバシップを表現する上記曖昧さの尺度の各値が０と１との間にある、システム。

【請求項２５】

請求項２４に記載のシステムであって、

上記メンバシップを表現する上記曖昧さの尺度の値が、上記２つ以上のクラスタのそれぞれに属する上記データ要素の確度に関連している、システム。

【請求項２６】

請求項２４に記載のシステムであって、

上記２つ以上のクラスタのそれぞれに属する上記データ要素のメンバシップを表現する上記曖昧さの尺度の各値の合計が１である、システム。

【請求項２７】

請求項２６に記載のシステムであって、

上記曖昧さの尺度の値を使用して会計の完全性を保つ手段、

を更に含むシステム。

【請求項２８】

請求項２７に記載のシステムであって、

所定の量についての会計の完全性を保つ手段が、上記曖昧さの尺度の値によって上記量を重み付けする手段によって達成される、システム。

【請求項２９】

請求項２８に記載のシステムであって、

上記メンバシップを表現する上記曖昧さの尺度の値を使用するデータ操作を実行する手

段、
を更に含むシステム。
【請求項３０】
請求項２９に記載のシステムであって、
上記データ操作が、上記１つ以上のクラスタの第１のクラスタ内の量の重み付けされた小計を計算するロールアップを含み、上記量が上記データ要素と関連付けられ、上記小計が、上記第１のクラスタ内で、上記第１のクラスタにおける上記データ要素の各々と関連付けられた上記量の値と上記第１のクラスタにおける上記データ要素の上記メンバシップを表現する上記曖昧さの尺度のそれぞれの値との積を合計することによって計算される、システム。
【請求項３１】
請求項３０に記載のシステムであって、
上記量の排他的小計及び上記量の包括的小計を計算する手段であって、上記排他的小計が、２つ以上のクラスタと関連付けられる上記第１のクラスタにおける上記データ要素を排除することによって計算され、上記包括的小計が、２つ以上のクラスタと関連付けられる上記第１のクラスタにおける上記データ要素を包含することによって計算される手段、
を更に含むシステム。
【請求項３２】
請求項２４に記載のシステムであって、
上記メンバシップを表現する上記曖昧さの尺度の値が関数に基づいて定められ、上記関数が、上記データ要素と上記２つ以上のクラスタとの間の関係を表現する、システム。
【請求項３３】
請求項３２に記載のシステムであって、
上記関数によって表現される上記関係が、上記２つ以上のクラスタのそれぞれに属する上記データ要素の確度に関連している、システム。
【請求項３４】
請求項３２に記載のシステムであって、
上記関数によって表現される上記関係が、上記データ要素と上記２つ以上のクラスタのそれぞれを表現する要素との間の定量化された類似性に基づく、システム。
【請求項３５】
請求項３４に記載のシステムであって、
上記２つ以上のクラスタのそれぞれを表現する上記要素が、それぞれのクラスタのキーである、システム。
【請求項３６】
請求項２２に記載のシステムであって、
上記２つ以上のクラスタの各クラスタに属する上記データ要素の上記曖昧さの尺度の値が、各クラスタについて等しい、システム。
【請求項３７】
請求項２２に記載のシステムであって、
上記２つ以上のクラスタの各クラスタに属する上記データ要素の上記曖昧さの尺度の値が、参照集合における上記データ要素の観測度数に基づく、システム。
【請求項３８】
請求項２２に記載のシステムであって、
上記２つ以上のクラスタの各クラスタが、上記データ要素における異なる潜在的エラーを表現し、上記２つ以上のクラスタの各クラスタに属する上記データ要素の上記曖昧さの尺度の値が、各クラスタによって表現される上記データ要素における上記潜在的エラーの確度に基づく、システム。
【請求項３９】
請求項２２に記載のシステムであって、
データ要素のクラスタを形成する手段が、

データ要素の複数のスーパークラスタを形成する手段、及び、
各スーパークラスタについて、上記スーパークラスタ内にデータ要素のクラスタを形成する手段、
を含む、システム。
【請求項４０】
請求項３９に記載のシステムであって、
各スーパークラスタを形成する手段が、
異なるデータ要素におけるオブジェクトの間での変形関係に基づいて、上記異なるデータ要素における上記オブジェクト間での整合を特定する手段、
を含む、システム。
【請求項４１】
請求項４０に記載のシステムであって、
第１のオブジェクトと第２のオブジェクトとの間の変形関係が、予め定められた閾値未満にある上記第１のオブジェクトと上記第２のオブジェクトとの間の距離を表現する関数の値に対応する、システム。
【請求項４２】
請求項４１に記載のシステムであって、
上記変形関係が、同値関係ではない、システム。
【請求項４３】
請求項２２に記載のシステムであって、
少なくとも１つのデータ要素が、１つより多くのスーパークラスタ中にある、システム。
【発明の詳細な説明】
【技術分野】
【０００１】
関連出願との相互参照
本願は、「ファジーなデータ操作（ＦＵＺＺＹ　ＤＡＴＡ　ＯＰＥＲＡＴＩＯＮＳ）」と題された米国特許出願第６１／１０７，９７１号（２００８年１０月２３日出願。引用により本明細書に組み込まれる）に基づく優先権を主張する。
【０００２】
本明細書は、データ管理の分野におけるファジーなデータ操作に関する。
【背景技術】
【０００３】
クラスタ化、結合、検索、ロールアップ、及びソート等のデータ操作は、データ管理においてデータを取り扱うのに用いられる。クラスタ化は、種々のグループにデータを分類する操作である。結合は、２つのデータを１つに結合する。キーによる検索は、そのキーに整合するデータ入力を見付け出す。ロールアップは、データのグループ全体の小計の１つ以上のレベル（又は他の組み合わせ）を計算する操作である。
【０００４】
データ管理においてデータ品質は重要である。データ操作の結果として生ずる誤りや不正確がデータ品質を低下させる。例えば、ＡＢＣ社の従業員（Ｊｏｈｎ　Ｓｍｉｔｈ）を臨時雇用者として分類するか、又は常勤者として分類するかによって、異なるレベルの待遇がＪｏｈｎ　Ｓｍｉｔｈに与えられる。Ｊｏｈｎ　Ｓｍｉｔｈの雇用形態の誤った分類（例えば、クラスタ化のデータ操作における誤り）は、ＡＢＣ社の人材データの品質に影響を及ぼす。
【０００５】
データ操作の実装によっては、整合するレコードを識別したり、関連するレコードのグループを定義したり、又はレコードをリンクさせたりするのに、フィールド値（「キー」）の厳密な比較を必要とする。データが曖昧であったり、不完全であったり、不十分であったり、又は不確実であったりする場合、フィールド値の厳密な比較に基づく方法は破綻

する場合がある。

【０００６】

データ操作（例えば、クラスタ化）に関連する固有の曖昧さが存在する場合、当該曖昧さを解消するための１つのアプローチは、単に、当該曖昧さを無視することや、データを特定のグループに強制的に分類することであってもよい。例えば、ＡＢＣ社の上記従業員（Ｊｏｈｎ　Ｓｍｉｔｈ）は、マーケティング部門及びＲ＆Ｄ部門の両方のために働いているとする。ＡＢＣ社の人材データベースにおいては、Ｊｏｈｎ　Ｓｍｉｔｈは、マーケティング部門又はＲ＆Ｄ部門の何れかと関連付けられているかもしれないが、多くの場合、一方の部門だけと関連付けられている。そのデータを特定のグループに強制的に分類することにより、当該固有の曖昧さを覆い隠して、データ品質に悪影響を及ぼす場合がある。

【０００７】

ある事象（例えば、ある資産の所有権に関する事業体Ａと事業体Ｂとの間での法的な争い）の結果が保留されているために、データ操作（例えば、クラスタ化）に関連する不確かさが存在する場合、データを特定のグループに強制的に分類することは、当該状況の流動性を扱う最善のアプローチではない場合がある。判決の前は、当該資産の所有権は不確かである。当該資産をＡ又はＢの何れかに与えることは不正確であると後に判明するかもしれない。

【０００８】

グループのメンバシップの識別が曖昧であるために、データ操作（例えば、ロールアップ）に関連する不確かさが存在する場合、幾つかの選択肢の中の１つのグループにメンバシップを割り付けて会計の完全性（ａｃｃｏｕｎｔｉｎｇ　ｉｎｔｅｇｒｉｔｙ）を保つことは、誤解を招く概念を生ずる場合がある。例えば、リスク評価や規制の目的のために取引先企業への貸出リスク（ｅｘｐｏｓｕｒｅ　ｏｎ　ｌｏａｎｓ）を特定することに銀行が興味を持つ場合がある。取引先企業の識別は会社名によってなされる場合が多いが、このことは、会社名の記録形式に幅広いバラツキがあるために、曖昧な識別に繋がる場合がある。次に、このことは、取引先企業への貸出リスクの割り付けが曖昧であることを意味する。正しくは１つの会社に関連付けられる貸出が、実は当該１つの会社の名前の単に変形形式である、見掛け上別個の幾つかの会社にされるようなことが起こる場合がある。このことは、何れの単一の取引先企業に対しても当該銀行がリスク（ｅｘｐｏｓｕｒｅ）を低く評価することに繋がる。あるいは、選択肢の中から任意に選ぶ場合、そのリスクが、正しくは別の取引先企業に属する際に、ある１つの取引先企業に誤って割り付けられる場合がある（一方にはリスクが高く評価され、他方にはリスクが低く評価される）。

【０００９】

正しくない情報又は欠落した情報のために、データ操作（例えば、結合）に関連する不確かさが存在する場合、データを特定のグループに強制的に分類したり、又は当該データを無視したりすることは、誤った関連付け又は情報の損失の何れかに繋がる場合がある。例えば、２つの異なるデータベースからのテーブルを結合しようとする際、これらのデータベースのテーブルによって共有される共通キーが全く無い場合が多い。このことを克服するために、これらのテーブル内のデータ（例えば、顧客の住所）を使用して、これら２つのデータベースにおけるレコード間の関係を推定する。しかしながら、住所の情報は正しくなかったり、又は不十分であったりする場合がある。ポスタル・アドレス・ファイル（Ｐｏｓｔａｌ　Ａｄｄｒｅｓｓ　Ｆｉｌｅ）等の信頼できる参照集合に照らした住所の妥当性確認により、テーブルＡにおけるレコード上の番地（ｈｏｕｓｅ　ｎｕｍｂｅｒ）が無効である（その番地を有する家は存在しない）一方、テーブルＢには、その住所の有効な代替の完成形であるかもしれない複数の住所があると示されたと仮定する。テーブルＡにおけるレコードにおける住所の完成形を任意に選択することは誤った関連付けに繋がる場合がある一方、当該レコードを無視することは情報の損失に繋がる。

【００１０】

不正確なデータ入力のために、データ操作（例えば、検索）に関連する不確かさが存在

する場合、１つのアプローチは、単一の選択肢又は代替の補正値の単純なリストを提案することである。このことが、オペレータによってデータベースに入力されているデータの妥当性確認プロセスの一部である場合、複数の選択肢が存在する際に単一の選択肢を提案することは、その補正値を許容することにおいて誤った安心感をオペレータに与えることに繋がる場合がある。選択肢の単純なリストを提供する場合、これらの選択肢の中から選択するための合理的根拠をオペレータが有していない場合がある。単一選択が求められ且つ間違った選択に起因するデータ品質のある程度の低下が許容される場合、起こり得るデータ品質の損失の最小化及び定量化が目的となる。

【発明の概要】

【００１１】

一般に、１つの態様において、データ記憶システムに格納されたデータ要素をクラスタ化する方法は、上記データ記憶システムからデータ要素を読み取るステップを含む。データ要素のクラスタが形成され、各データ要素は少なくとも１つのクラスタのメンバとなる。少なくとも１つのデータ要素が、２つ以上のクラスタと関連付けられる。上記２つ以上のクラスタのそれぞれに属するデータ要素のメンバシップは、曖昧さの尺度によって表現される。情報が上記データ記憶システムに格納され、形成されたクラスタを表現する。

【００１２】

以下の特徴の１つ以上を含む態様もあり得る。

【００１３】

上記２つ以上のクラスタのそれぞれに属するデータ要素のメンバシップを表現する曖昧さの尺度の各値がゼロ（０）と１との間にあってもよい。

【００１４】

上記メンバシップを表現する曖昧さの尺度の値が、上記２つ以上のクラスタのそれぞれに属するデータ要素の確度に関連していてもよい。

【００１５】

上記２つ以上のクラスタのそれぞれに属するデータ要素のメンバシップを表現する曖昧さの尺度の各値の合計が１であってもよい。

【００１６】

上記方法は、上記曖昧さの尺度の値を使用して会計の完全性を保つステップ、を含んでもよい。

【００１７】

所定の量についての会計の完全性が、上記曖昧さの尺度の値によって上記量を重み付けするステップによって達成されてもよい。

【００１８】

上記方法は、上記メンバシップを表現する曖昧さの尺度の値を使用するデータ操作を実行するステップ、を含んでもよい。

【００１９】

上記データ操作が、上記１つ以上のクラスタの第１のクラスタ内の量の重み付けされた小計を計算するロールアップを含み、上記量が上記データ要素と関連付けられ、そして上記小計が、上記第１のクラスタ内で、上記第１のクラスタにおけるデータ要素の各々と関連付けられた量の値と上記第１のクラスタにおけるデータ要素のメンバシップを表現する曖昧さの尺度のそれぞれの値との積を合計することによって計算されてもよい。

【００２０】

上記方法は、上記量の排他的小計及び上記量の包括的小計を計算するステップであって、上記排他的小計が、２つ以上のクラスタと関連付けられる第１のクラスタにおけるデータ要素を排除することによって計算され、上記包括的小計が、２つ以上のクラスタと関連付けられる第１のクラスタにおけるデータ要素を包含することによって計算されるステップ、を含んでもよい。

【００２１】

上記メンバシップを表現する曖昧さの尺度の値が関数に基づいて定められ、上記関数が

、上記データ要素と上記２つ以上のクラスタとの間の関係を表現してもよい。
【００２２】
上記関数によって表現される関係が、上記２つ以上のクラスタのそれぞれに属するデータ要素の確度に関連していてもよい。
【００２３】
上記関数によって表現される関係が、上記データ要素と上記２つ以上のクラスタのそれぞれを表現するデータ要素との間の定量化された類似性に基づく、方法。
【００２４】
上記２つ以上のクラスタのそれぞれを表現する要素が、それぞれのクラスタのキーであってもよい。
【００２５】
アレンジによっては、上記２つ以上のクラスタの各クラスタに属するデータ要素の曖昧さの尺度の値が、各クラスタについて等しくてもよい。
【００２６】
上記２つ以上のクラスタの各クラスタに属するデータ要素の曖昧さの尺度の値が、参照集合におけるデータ要素の観測度数に基づいてもよい。
【００２７】
上記２つ以上のクラスタの各クラスタが、上記データ要素における異なる潜在的エラーを表現してもよい。上記２つ以上のクラスタの各クラスタに属するデータ要素の曖昧さの尺度の値が、各クラスタによって表現されるデータ要素における潜在的エラーの確度に基づいてもよい。
【００２８】
データ・クラスタを形成するステップが、
データ要素の複数のスーパークラスタを形成するステップ、及び、
各スーパークラスタについて、上記スーパークラスタ内にデータ要素のクラスタを形成するステップ、
を含んでもよい。
【００２９】
各スーパークラスタを形成するステップが、
異なるデータ要素におけるオブジェクトの間での変形関係（ｖａｒｉａｎｔ　ｒｅｌａｔｉｏｎ）に基づいて、上記異なるデータ要素におけるオブジェクト間での整合を特定するステップ、
を含んでもよい。
【００３０】
第１のオブジェクトと第２のオブジェクトとの間の変形関係が、予め定められた閾値未満にある上記第１のオブジェクトと上記第２のオブジェクトとの間の距離を表現する関数の値に対応してもよい。
【００３１】
アレンジによっては、上記変形関係は、同値関係（ｅｑｕｉｖａｌｅｎｃｅ　ｒｅｌａｔｉｏｎ）ではなくてもよい。
【００３２】
少なくとも１つのデータ要素が、１つ以上のスーパークラスタ中にあってもよい。
【００３３】
もう１つの態様において、一般に、データ記憶システムに格納されたデータ要素をクラスタ化するシステムは、
上記データ記憶システムからデータ要素を読み取る手段、
データ要素のクラスタを形成する手段であって、各データ要素が少なくとも１つのクラスタのメンバである、手段、
少なくとも１つのデータ要素を２つ以上のクラスタと関連付ける手段であって、上記２つ以上のクラスタのそれぞれに属するデータ要素のメンバシップが、曖昧さの尺度によっ

て表現される、手段、及び
上記データ記憶システムに情報を格納して、上記形成されたクラスタを表現する手段、を備える。
【００３４】
もう１つの態様において、一般に、データ記憶システムに格納されたデータ要素をクラスタ化するためのコンピュータ・プログラムを格納するコンピュータ可読媒体について説明する。上記コンピュータ・プログラムは、
上記データ記憶システムからデータ要素を読み取るステップ、
データ要素のクラスタを形成するステップであって、各データ要素が少なくとも１つのクラスタのメンバである、ステップ、
少なくとも１つのデータ要素を２つ以上のクラスタと関連付けるステップであって、上記２つ以上のクラスタのそれぞれに属するデータ要素のメンバシップが、曖昧さの尺度によって表現される、ステップ、及び
上記データ記憶システムに情報を格納して、上記形成されたクラスタを表現するステップ、
をコンピュータに実行させるための命令を含む。
【００３５】
もう１つの態様において、一般に、キーを受け取り、データ記憶システムから１つ以上のデータ要素を返すデータ操作を実行する方法は、上記キーと上記データ要素の１つ以上の検索フィールドの値との間での候補整合（ｃａｎｄｉｄａｔｅ　ｍａｔｃｈｅｓ）に基づいて、複数の候補データ要素を特定するステップを含む。上記候補整合は、上記検索フィールドとは異なる上記候補データ要素の１つ以上の比較フィールドの値に基づいて確証される。
【００３６】
以下の特徴の１つ以上を含む態様もあり得る。
【００３７】
上記データ操作が、
データ要素のクラスタを形成するステップであって、各データ要素が少なくとも１つのクラスタのメンバである、ステップ、
を含んでもよい。
【００３８】
少なくとも１つのデータ要素が２つ以上のクラスタと関連付けられ、上記２つ以上のクラスタのそれぞれに属するデータ要素のメンバシップが、曖昧さの尺度によって表現されてもよい。
【００３９】
上記データ操作が、上記１つ以上のクラスタの第１のクラスタ内の量の重み付けされた小計を計算するロールアップを含み、上記量が上記データ要素と関連付けられ、そして上記小計が、上記第１のクラスタ内で、上記第１のクラスタにおけるデータ要素の各々と関連付けられた量の値と上記第１のクラスタにおけるデータ要素のメンバシップを表現する曖昧さの尺度のそれぞれの値との積を合計することによって計算されてもよい。
【００４０】
また、上記方法は、上記量の排他的小計及び上記量の包括的小計を計算するステップであって、上記排他的小計が、２つ以上のクラスタと関連付けられる第１のクラスタにおけるデータ要素を排除することによって計算され、上記包括的小計が、２つ以上のクラスタと関連付けられる第１のクラスタにおけるデータ要素を包含することによって計算されるステップ、を含んでもよい。
【００４１】
上記２つ以上のクラスタのそれぞれに属するデータ要素のメンバシップを表現する曖昧さの尺度の各値がゼロ（０）と１との間にあってもよい。
【００４２】

上記メンバシップを表現する曖昧さの尺度の値が、上記２つ以上のクラスタのそれぞれに属するデータ要素の確度に関連していてもよい。

【００４３】

上記メンバシップを表現する曖昧さの尺度の値が関数に基づいて定められ、上記関数が、上記データ要素と上記２つ以上のクラスタとの間の関係を表現してもよい。

【００４４】

上記関数によって表現される関係が、上記２つ以上のクラスタのそれぞれに属するデータ要素の確度に関連していてもよい。

【００４５】

上記方法は、所定のクラスタにおける所定のデータ要素の１つ以上の比較フィールドの値に基づいて、上記所定のデータ要素のメンバシップを特定するステップ、を含んでもよい。

【００４６】

もう１つの態様において、一般に、キーを受け取り、データ記憶システムから１つ以上のデータ要素を返すデータ操作を実行するシステムは、

上記キーと上記データ要素の１つ以上の検索フィールドの値との間での候補整合に基づいて、複数の候補データ要素を特定する手段、及び

上記検索フィールドとは異なる上記候補データ要素の１つ以上の比較フィールドの値に基づいて、上記候補整合を確証する手段、

を備える。

【００４７】

もう１つの態様において、一般に、キーを受け取り、データ記憶システムから１つ以上のデータ要素を返すデータ操作を実行するためのコンピュータプログラムを格納するコンピュータ可読媒体について説明する。上記コンピュータ・プログラムは、

上記キーと上記データ要素の１つ以上の検索フィールドの値との間での候補整合に基づいて、複数の候補データ要素を特定するステップ、及び

上記検索フィールドとは異なる上記候補データ要素の１つ以上の比較フィールドの値に基づいて、上記候補整合を確証するステップ、

をコンピュータに実行させるための命令を含む。

【００４８】

もう１つの態様において、一般に、データ記憶システムにおけるデータ要素のデータ品質を測定する方法は、上記データ記憶システムからデータ要素（１２０）を読み取るステップ、を含む。上記データ要素の１つ以上のフィールドにおける１つ以上の入力の各々について、上記入力についての曖昧さの尺度の値が演算される。上記曖昧さの尺度の値に基づいて、上記データ記憶システムにおけるデータ要素のデータ品質の表現が出力される。

【００４９】

以下の特徴の１つ以上を含む態様もあり得る。

【００５０】

上記曖昧さの尺度の値を演算するステップが、上記データ要素の１つ以上のフィールドにおける入力を参照値と比較するステップ、を含んでもよい。参照値との厳密な整合ではない少なくとも第１の入力について１つ以上の変形が識別されてもよい。上記第１の入力についての変形に基づいて、上記第１の入力についての曖昧さの尺度の値が演算されてもよい。

【００５１】

上記第１の入力についての曖昧さの尺度の値が、上記第１の入力についての変形の数に基づいてもよい。

【００５２】

上記データ記憶システムにおけるデータ要素のデータ品質の表現が、特定数の変形を有する入力の数のヒストグラム・プロットを含んでもよい。

【００５３】

上記特定数の変形が、ある範囲内にあるものとして規定されてもよい。

【００５４】

上記データ記憶システムにおけるデータ要素のデータ品質の表現が、予め定められた閾値よりも大きい数の変形を有する入力のリストを含んでもよい。

【００５５】

曖昧さの尺度の値を演算するステップが、１つ以上のフィールドにおける異なる入力のそれぞれの頻度を特定するステップ、を含んでもよい。他の入力の頻度と比較した第１の入力の相対頻度に基づいて、上記第１の入力についての曖昧さの尺度の値が演算されてもよい。

【００５６】

もう１つの態様において、一般に、データ記憶システムにおけるデータ要素のデータ品質を測定するシステムは、

上記データ記憶システムからデータ要素を読み取る手段、

上記データ要素の１つ以上のフィールドにおける１つ以上の入力の各々について、上記入力についての曖昧さの尺度の値を演算する手段、及び

上記曖昧さの尺度の値に基づいて、上記データ記憶システムにおけるデータ要素のデータ品質の表現を出力する手段、

を備える。

【００５７】

もう１つの態様において、一般に、データ記憶システムにおけるデータ要素のデータ品質を測定するためのコンピュータ・プログラムを格納するコンピュータ可読媒体について説明する。上記コンピュータ・プログラムは、

上記データ記憶システムからデータ要素を読み取るステップ、

上記データ要素の１つ以上のフィールドにおける１つ以上の入力の各々について、上記入力についての曖昧さの尺度の値を演算するステップ、及び

上記曖昧さの尺度の値に基づいて、上記データ記憶システムにおけるデータ要素のデータ品質の表現を出力するステップ、

をコンピュータに実行させるための命令を含む。

【００５８】

もう１つの態様において、一般に、少なくとも１つのデータ記憶システムに格納された２つ以上のデータセットからのデータ要素を結合する方法は、第１のデータセットからのデータ要素におけるオブジェクトと第２のデータセットからのデータ要素におけるオブジェクトとの間の変形関係に基づいて、上記第１のデータセットからのデータ要素におけるオブジェクトと上記第２のデータセットからのデータ要素におけるオブジェクトとの間の整合を特定するステップ、を含む。それぞれのオブジェクトが整合として特定されたそれぞれのデータ要素が評価される。データ要素の上記評価に基づいて、上記第１のデータセットからのデータ要素が上記第２のデータセットからのデータ要素と結合される。

【００５９】

以下の特徴の１つ以上を含む態様もあり得る。

【００６０】

第１のオブジェクトと第２のオブジェクトとの間の変形関係が、予め定められた閾値未満である上記第１のオブジェクトと上記第２のオブジェクトとの間の距離を表現する関数の値に対応してもよい。

【００６１】

上記変形関係は、同値関係ではなくてもよい。

【００６２】

上記第１のデータセットからの第１のデータ要素におけるオブジェクトと上記第２のデータセットにおける第２のデータ要素におけるオブジェクトとの間の整合を特定するステップが、

上記変形関係が上記第１のデータ要素におけるオブジェクトと上記第２のデータ要素

におけるオブジェクトとの間に当てはまることを特定するステップ、
を含んでもよい。
【００６３】
上記第１のデータセットからの第１のデータ要素におけるオブジェクトと上記第２のデータセットにおける第２のデータ要素におけるオブジェクトとの間の整合を特定するステップが、
上記変形関係が上記第１のデータ要素におけるオブジェクトと上記第１のデータセットにおける第３のデータ要素におけるオブジェクトとの間に当てはまること、及び上記変形関係が上記第３のデータ要素におけるオブジェクトと上記第２のデータ要素におけるオブジェクトとの間に当てはまること、を特定するステップ、
を含んでもよい。
【００６４】
それぞれのオブジェクトが整合として特定された、それぞれのデータ要素を評価するステップが、
上記それぞれのデータ要素において整合として特定されたそれぞれのオブジェクト以外のオブジェクトを比較するステップ、
を含んでもよい。
【００６５】
もう１つの態様において、一般に、少なくとも１つのデータ記憶システムに格納された２つ以上のデータセットからのデータ要素を結合するシステムは、
第１のデータセットからのデータ要素におけるオブジェクトと第２のデータセットからのデータ要素におけるオブジェクトとの間の変形関係に基づいて、上記第１のデータセットからのデータ要素におけるオブジェクトと上記第２のデータセットからのデータ要素におけるオブジェクトとの間の整合を特定する手段、
それぞれのオブジェクトが整合として特定されたそれぞれのデータ要素を評価する手段、及び
データ要素の上記評価に基づいて、上記第１のデータセットからのデータ要素を上記第２のデータセットからのデータ要素と結合する手段、
を備える。
【００６６】
もう１つの態様において、一般に、少なくとも１つのデータ記憶システムに格納された２つ以上のデータセットからのデータ要素を結合するためのコンピュータ・プログラムを格納するコンピュータ可読媒体について説明する。上記コンピュータ・プログラムは、
第１のデータセットからのデータ要素におけるオブジェクトと第２のデータセットからのデータ要素におけるオブジェクトとの間の変形関係に基づいて、上記第１のデータセットからのデータ要素におけるオブジェクトと上記第２のデータセットからのデータ要素におけるオブジェクトとの間の整合を特定するステップ、
それぞれのオブジェクトが整合として特定されたそれぞれのデータ要素を評価するステップ、及び
データ要素の上記評価に基づいて、上記第１のデータセットからのデータ要素を上記第２のデータセットからのデータ要素と結合するステップ、
をコンピュータに実行させるための命令を含む。
【図面の簡単な説明】
【００６７】
【図１】グラフに基づく演算を実行するシステムのブロック図である。
【図２Ａ】複数のクラスタに属するデータ要素の例である。
【図２Ｂ】クラスタに対して実行される操作の例である。
【図２Ｃ】距離の計算例である。
【図２Ｄ】距離の計算例である。
【図３】ファジーなクラスタの図解である。

【図４】ファジーなクラスタのもう１つの図解である。
【図５】ファジーなクラスタを生成する方法のフローチャートである。
【図６】ファジーな検索の例の図解である。
【発明を実施するための形態】
【００６８】
ファジーなデータ操作を実行するための手法は、データセットを格納する種々の形式のデータベースを含む様々なタイプのシステムに適用することができる。本明細書において使用されているように、データセットは、それぞれのフィールド（「属性」又は「カラム」とも呼ばれる）についての値を有する記録としてデータの部分を編成することを可能とする何等かのデータの集まりを含む。上記データベース・システム及び格納されたデータセットは、最新式のデータベース管理システム又は単純な単層ファイル（ｆｌａｔ　ｆｉｌｅ）を格納するファイル・システム等の様々な形式の何れをとることもできる。様々なデータベース・システムの１つの態様は、データセット内のレコードに使用されるレコード構造のタイプである（各レコード内のフィールドに使用されるフィールド構造を含むこともできる）。システムによっては、データセットのレコード構造は、個々のテキスト文書をレコードとして単純に定義するものであってもよく、当該文書の内容は１つ以上のフィールドの値を表現する。システムによっては、単一のデータセット内の全てのレコードが同じ構造（例えば、フィールド構造）を有することを要件としない。
【００６９】
複雑な演算は、有向グラフを通るデータ・フローとして表すことができることが多く（データフロー・グラフと呼ばれる）、当該演算の構成要素は当該グラフの節点（ｖｅｒｔｉｃｅｓ）と関連付けられ、これらの構成要素の間のデータ・フローは当該グラフのリンク（アーク、エッジ）に対応する。このようなグラフに基づく演算を実装するシステムは、「グラフとして表された演算の実行（ＥＸＥＣＵＴＩＮＧ　ＣＯＭＰＵＴＡＴＩＯＮＳ　ＥＸＰＲＥＳＳＥＤ　ＡＳ　ＧＲＡＰＨＳ）」と題された米国特許第５，９６６，０７２号に記載されている（引用により、本明細書に組み込まれる）。グラフに基づく演算を実行するための１つのアプローチは、当該グラフの種々の節点と各々が関連付けられた多数のプロセスを実行すること、及び当該グラフのリンクに従って当該プロセス間の通信経路を確立することである。例えば、上記通信経路は、ＴＣＰ／ＩＰ若しくはＵＮＩＸ（登録商標）ドメイン・ソケットを使用するものであっても、プロセス間でデータを渡すのに共有メモリを使用するものであってもよい。
【００７０】
図１を参照すると、グラフに基づく演算を実行するシステム１０は、データ・ストア１２に連結された開発環境１３及びデータ・ストア１２に連結されたランタイム環境１８を備える。開発者１１は、開発環境１４を使用してアプリケーションを構築する。アプリケーションは、データ・ストア１２におけるデータ構造（開発者による開発環境１４の使用結果として当該データ・ストアに書き込まれたものであってもよい）によって規定される１つ以上のデータフロー・グラフと関連付けることができる。演算グラフ１５についてのデータ構造１３は、例えば、演算グラフの節点（構成要素又はデータセット）及び節点間のリンク（ワーク要素のフローを表現する）を規定する。また、上記データ構造は、データフロー・グラフの構成要素、データセット、及びフローの様々な特徴を含むこともできる。
【００７１】
ランタイム環境１８は、ＵＮＩＸ（登録商標）オペレーティング・システム等の、好適なオペレーティング・システムの制御下にある１つ以上の汎用コンピュータ上にホスティングすることができる。例えば、ランタイム環境１８は、ローカル型であっても（例えば、ＳＭＰコンピュータ等のマルチプロセッサ・システム）、又はローカル分散型であっても（例えば、クラスタ若しくはＭＰＰとして連結された複数のプロセッサ）、又はリモート型であっても、又はリモート分散型であっても（例えば、ＬＡＮ若しくはＷＡＮネットワークを介して連結された複数のプロセッサ）、あるいはそれらの何れの組み合わせであ

っても、複数の中央処理装置（ＣＰＵ）を使用するコンピュータ・システムの構成を備える多重ノード並列演算環境を含むことができる。

【００７２】

ランタイム環境１８は、データ・ストア１２及び／又はユーザ１７から演算を実行及び構成するための制御入力を受け取るように構成される。上記制御入力は、（格納されたグラフ・データ構造において規定される）対応するデータフロー・グラフを使用して特定のデータセットを処理するコマンドを含むことができる。ユーザ１７は、例えば、コマンド・ライン又はグラフィカル・インターフェースを使用して、ランタイム環境１８と双方向に働きかけることができる。

【００７３】

ランタイム環境１８は、前実行モジュール２０及び実行モジュール２２を含む。前実行モジュール２０は、あらゆる前処理プロシージャを実行し、様々なファジーな操作（例えば、米国特許出願公開第２００９／０１８２７２８号に記載されているような操作。引用により、本明細書に組み込まれる）に使用される辞書２１やアーカイブ２４等の演算グラフを実行するためのリソースを準備し、維持する。辞書２１は、データセットに現れる単語及び単語についての関連情報を格納する。アーカイブ２４は、データセットの単語、句、又はレコードに基づく前処理からの様々な結果を格納する。辞書２１及びアーカイブ２４は、様々なフォーマットの何れでも実装されてることができ、データの単一の集まりとして、又は複数の辞書及びアーカイブとして、編成することができる。実行モジュール２２は、構成要素の演算を実行するために演算グラフに割り付けられたプロセスの実行をスケジュールし、制御する。実行モジュール２２は、グラフの構成要素と関連付けられる処理中にアクセスされるシステム１０に連結された外部の演算リソース（例えば、データベース・システムからレコードを提供するデータ・ソース２６）と双方向に働きかけることができる。システム１０において実行されるファジーな操作は、データ品質を評定するためのデータの分析又はデータの編成及び／若しくは統合等の様々な目的に使用することができる。

【００７４】

何れの企業又は他の組織においても、核となる資産は、取引する製品、サービス及び顧客、個人、銀行及び他の企業との契約及び会計のリストに由来する、その事業を運営するためにその企業又は他の組織が保有しているデータである。このデータは、複数のフォーマットで、紙やスプレッドシートからリレーショナル・データベースやエンタープライズ・アプリケーションに至る複数のシステム（例えば、会計システム又はサプライ・チェーン管理システム）に、格納される。全ての組織の懸念の中心は、このデータの品質及び完全性である。

【００７５】

請求書が正しくない価格や偽装された商品を含む場合、間違った金額が課されたり、又は誤った品物が配送されたりする。顧客又はサプライヤの住所が間違っている場合、出荷又は発注が遅れたり又は無くなったり、請求書又は支払いが目的とする相手に届かなかったりする場合がある。あるシステム上での顧客を表現するキーが別のシステム上での異なる顧客の会計にリンクしている場合、その顧客の会計の状態についての報告は信頼できないであろうし、更に悪いことに、ある顧客が他の顧客の会計へのアクセスを有することになる場合もある。不十分なデータ品質は、事業のきちんとした運営を混乱させ、収益の損失、悪評又は機会損失に繋がる場合がある。

【００７６】

事業又は組織のデータの重要な部分集合は、時としてマスター・データと称される、非取引（ｎｏｎ－ｔｒａｎｓａｃｔｉｏｎａｌ）参照用データである。これは、製品、顧客、会計、サプライヤ、及び各データ項目の特定の属性を表現するのに使用される特定の有効な値（例えば、顧客が男性若しくは女性の性別を有すること、又は製品が列挙リストの中の１つの色を有すること）のリストを含むことができる。一般に、取引や価格等の、組織の短期的な運営データはマスター・データから排除される。マスター・データ管理は、

組織及び組織の参照用データのメンテナンスと関係する。

【００７７】

データ品質及び参照整合性に関する問題は多くの形式をとる。これらの問題は、一貫性を保つのが困難な場合がある、異なる種類のデータ・システムが複数存在することによって悪化する。潜在的な問題の非網羅的なリストは以下の通りである。

【００７８】

１）データが不正確に入力又は記録される場合がある：なされた入力が、意図したものではない。例えば、顧客の名前若しくは住所、製品のラベル若しくは説明書、又は列挙リストから選ばれるべき値における単語のスペル違いに繋がる誤植又は転記の誤りが入力にある場合がある。多くのデータ入力アプリケーションは、入力時にデータを確認して、このようなエラーを防ぐことを目的とするセーフガードを有するが、未だにエラーは起こっている。

【００７９】

２）データが不十分である場合がある：全てのフィールドが埋まっていない。顧客アプリケーションが特定のフィールドが欠落している情報を有していたかもしれない。フォームの完成が入力中に中断されたかもしれない。情報が入力時に無効とみなされ、破棄されたかもしれない。入力を完成するための情報が入力時には入手可能ではなかったかもしれない（何等かの他の作業の完了の保留等）。

【００８０】

３）データが無効である場合もある：フィールドは埋まっているが、無効な値で埋まっている。列挙リストから期待される値の何れにも入力が整合しない場合がある。期待されるデータ型に対して入力が有効ではないかもしれない（例えば、十進数のフィールドに英字が存在したり、日付における日が、その月の日数よりも大きかったり（例えば、６月３１日）する場合がある）。

【００８１】

４）誤ったフィールドにデータが入力される場合がある：住所の通りのフィールドに、都市名コードや郵便番号が現れる場合がある。期待されるものと異なるフォーマットを有する外国の住所を、期待されるフォーマットに強制的に適合させたかもしれない。請求書又は発注書のフォーム上の説明又はコメントのフィールドに製品ＩＤがあるかもしれない。名字が一般的な名である場合（例えば、Ｇｒｅｇｏｒｙ　Ｐａｕｌ）や、名前が珍しい名前若しくは外国の名前である場合、個人の姓（ラスト・ネーム）及び名（ファースト・ネーム）が入れ替わっているかもしれない。

【００８２】

５）データ入力についての標準が存在しない場合がある：データが一貫性無く入力される場合がある。住所の行の順序が標準化されておらず、同じデータセットにおいてさえ、常に同じように記録されている訳ではない場合がある。会社名の詳細なフォームが標準化されておらず、同じデータセットにおいてさえ、多くの変形形式が許容されている場合がある。顧客名が正式なミドル・ネームを含む場合があり、又はミドル・ネームが無い場合があり、又はミドル・イニシャルだけの場合もある。同様に、名（ファースト・ネーム）がイニシャルだけの場合もある。「二重姓」の姓（ラスト・ネーム）が、ハイフンを伴う若しくは伴わない姓（ラスト・ネーム）に存在する場合があり、又は、ミドル・ネームのフィールドと姓（ラスト・ネーム）のフィールドとの間で分けられている場合がある。

【００８３】

６）データがフリー・テキスト・フィールドに保持されている場合がある：請求書又は発注書のフォーム上のコメント・フィールドにおける注釈が、製品名又は記述的な属性等の、さもなければ欠落していたであろう、重要な情報を含んでいる場合がある。例えば、結婚により女性の姓（ラスト・ネーム）が変わる場合、データベース・テーブルにおける記述フィールドが、他のフィールドに対する変更についての説明を含む場合がある。

【００８４】

７）キー関係が壊れている場合がある：データベースはキーを使用して、異なるテーブ

ルや時には異なるデータベースにおけるテーブルに保持されている関連データをリンクする。キーが正しいレコードを正しくリンクしない場合、当該データベースの参照整合性が破綻している。ある顧客が正しくは別の顧客に属する会計にリンクされているような場合、キーがレコードを不正確にリンクする場合がある。存在しないレコードにキーがリンクする場合もある（例えば、ある会計レコードのける顧客キーが、実在する顧客レコードの何れにもリンクしない）。場合によっては、上記顧客レコードは存在するが、異なるキーを有する（このようなキーは時として「紛失キー」と記述される）。他の場合においては、対応する顧客レコードが全く存在しない場合、当該会計レコードは孤児であると言われる。

【００８５】

８）キー関係が存在しない場合がある：出所が異なるデータベースが類似のデータを保持するものの、それらによって共有されるデータをリンクするキーを保持していないことがある。ある事業分野が別の事業分野との顧客の共有を実現しないことは希である。企業又は組織が合併する場合、これら２つの事業体のマスター・データが組み合わされる場合がある。これら２つの事業体の異なる標準及び不等価な有効値により、マスター・データの一貫した集合の達成が困難となるが、顧客等の共有データの識別及びリンクの問題の方がより厳しいことが多い。

【００８６】

データ・クレンジングは、これらの問題点を識別し、訂正しようとするものである。レガシー・システムの数及び複雑さ、システム間のインターフェースの数、並びに新しいシステムの導入の速度の故に、真の課題は、データ品質の問題を如何に修復するかではなく、それらを如何に上手く対処するかであることが多い。

【００８７】

企業又は組織のシステムにおけるデータを見付け出し、アクセスし、そして操ることにおける主要な概念は、「キー」の概念である。主キーはフィールド、又はフィールドの組み合わせであり、その値は、データセットにおけるレコードを一意的に識別するのに役立つ。リレーショナル・データベース内では、各テーブルが、そのテーブル内のレコードを一意的に識別する主キーを有してもよい（主キーが一意的でない場合、それはデータ品質の別問題である）。テーブルにおける外部キーは、他のテーブルにおけるレコードにリンクするキーである。

【００８８】

多くのデータ操作を行うことができるが、データ操作は、データベースのテーブル又は他のデータセットのキーに依存する。よく知られているキーに基づくデータ操作は、キーによるルックアップ、結合、ロールアップ、スキャン、ソート、マージ及び、並列処理における、キーによる分割である。これらのデータ操作は、キーの厳密な一致（ａｇｒｅｅｍｅｎｔ）（本明細書においては「厳密な整合」と呼ぶ）に基づく。ルックアップというデータ操作においては、キーを使用して、厳密に整合するキーを有するルックアップ・データセットから１つ以上のレコードを読み出す。結合というデータ操作においては、１回に１レコードずつ、あるデータセットからのレコードの内容に、別のデータセットからの共通キーを共有するレコードの内容を付け加える（そして、ことによると、部分集合化する）ことによって、２つ（又は２つ以上）のデータセットを組み合わせる。２つ以上のレコードが整合する共通キーを有する場合、整合するレコードのペアの各々について別個の出力レコードが形成される。

【００８９】

ロールアップというデータ操作においては、共通キーを共有するレコードのグループの内容を組み合わせて、同じキーを有する単一の出力レコードを生成する。例えば、取引金額を合計しながら取引レコードを顧客レベルにロールアップすることにより、ある顧客の全取引金額が得られる。スキャンというデータ操作においては、共通キーを共有するレコードのグループにおける各レコードについて、同じキーを有する、これまでに見られた全てのレコードの内容を使用して、出力レコードを算出する。例えば、顧客による取引のス

キャンを用いて、顧客支出の現在高を算出することができる。

【００９０】

ソートというデータ操作においては、レコードをそれらのキー値によって順序付ける。マージというデータ操作においては、１つ以上のデータ・ストリームからのソートされたデータを組み合わせて、出力レコードもまたソートされているように、単一のストリームとする。並列処理におけるキーによる分割というデータ操作においては、キーの値に基づいて、データを処理パーティションに割り当てる。

【００９１】

上記において考察した種類のデータ品質問題を各々が有するかもしれない複数の独立したシステムが共存する場合、レコードを共通データと関係付けるキーは一般的には存在せず、存在するキーは信頼性が無い場合がある。結局のところ、各レコード内のデータは重要な項目である。キーは、データを識別し、データにアクセスするためにデータベースに導入される便利なフィクションであると考えることができる。信頼性のあるキーが無い場合、識別の目的にデータそのものを使用してもよい。

【００９２】

レコードの内容によるレコードへのアクセスは検索に基づく。例えば、１つのデータベースにおける顧客を、第２のデータベースにおいて名前によって探してもよい。名前は曖昧な識別子であるので、それをキーとすることは希である。名前を使用して識別を開始することができるけれども、整合を確証するのに誕生日や住所等の支援情報が必要とされるのが一般的である。

【００９３】

その上、データ品質問題の故に、レコードが正しく整合するために、名前も確証情報も厳密に一致することが必要とされない。厳密な一致は限定的すぎる場合があり、正確な整合を要求することが、多くの正しい識別の欠落に繋がる場合がある。（ファジーな）検索のデータ操作により、しっかりと整合するけれども、必ずしも厳密には整合しないデータ入力が取り出される。例えば、「Ｌｅｓｌｉｅ」についてのファジーな検索により、「Ｌｅｓｌｅｙ」と名付けられた人物についてのレコードが返される場合がある。ファジーな検索において、種々の度合いの類似性又は確証を有する整合レコードが２つ以上存在する場合がある（Ｌａｓｌｉｅについての検索により、Ｌｅｓｌｅｙと名付けられた第２の人物のレコードが取り出される場合もある）。候補整合が十分に確証されず、限定的整合又は許容整合と称される場合がある。例えば、取り出されたＬｅｓｌｅｙのレコードの誕生日がＬｅｓｌｉｅのレコードの誕生日と一致しない場合があり、この場合、当該候補整合は確証されない。

【００９４】

検索時、ルックアップのための厳密なキーを使用する単一段階プロセスを、２段階プロセスによって置き換える。レコードは、検索語を使用する取り出しのために識別され、整合を特定するために評定される。検索語がレコードを一意的に識別することは希であるので、検索語はキーではない。しかしながら、検索語をキーのように使用して、レコードをリンクする。

【００９５】

明確にするために、そこから検索語を選択するためのフィールドを、レコードを比較して整合の品質を評定するのに使用されるフィールドと区別することが有用である。これらを、それぞれ検索フィールド及び比較フィールドと呼んでもよい。

【００９６】

検索語又は比較フィールドが同一ではない場合に整合を見出して判定するために、採点関数を使用して、変形値（ｖａｒｉａｎｔ　ｖａｌｕｅ）を認識してもよい。候補整合を、変形検索語を使用して取り出し、採点関数を使用して評価して、確証フィールド間の整合の品質を定量化してもよい。これらの採点関数は、様々なデータ品質問題を明らかにするように設計される。それらは、これらの問題にもかかわらず、減点されたスコアで整合を認識する。例えば、個人名についての採点関数は、姓（ラスト・ネーム）と名（ファー

スト・ネーム）との入れ替え又はミドル・イニシャルの使用には寛容であるようにしてもよく、企業名用に調整された採点関数は、欠落している単語よりも単語の順序に重きを置くようにしてもよい。

【００９７】

厳密なキーのもう１得ｔの基本的な用途は、共通するキー値を有するレコードの集合（キーグループと呼ばれることが多い）を識別することである。これらのキーグループは、多くのキーに基づくデータ操作において重要な役割を演ずる。厳密に整合するキーの要件が緩和される場合、如何にしてキーをグループ化するのかという疑問が生ずる。緩和された整合基準に基づいて一緒にグループ化されたキーの集合は、クラスタと呼ばれる。

【００９８】

一般に、クラスタは、それらの比較フィールドが比較テストに合格するレコードの集合であってもよい。例えば、１つのアレンジにおいて、クラスタについてのあるレコードのスコアが閾値を超える場合、そのレコードはそのクラスタのメンバである。クラスタについてのレコードのスコアを定義する方法としては種々の方法が多数存在し、概して、クラスタの各メンバについてレコードを個々に採点し、次いでそれらのスコアを足し合わせることを含むけれども、この方法に限定されるものではない。例えば、スコアは、クラスタの各メンバについてのレコードのスコアの最大値であってもよく、又はクラスタの各メンバについてのスコアの平均値であってもよい。アレンジによっては、レコードのペアを採点することは、フィールド値の一方の集合を他方の集合と比較した結果に数を割り付けることを含む。フィールド値の比較が、定性的評価及び定量的評価の両方を含んでもよい。

【００９９】

クラスタの定義における問題は、フィールド値の比較が、採点された関係であるため、曖昧なメンバシップの割り当てが可能であるということである。特に、採点は、２つ以上のクラスタに属する１つのデータを示す場合がある。１つのアレンジにおいて、上記１つのデータを上記クラスタの中の１つに強制的に属させて、それらが厳密なキーの場合にあるとして、それらのクラスタをハッキリと定義することによって、この曖昧さに対処してもよい。この場合、上記キーに基づくデータ操作は、本質的に、それらが厳密なキーの場合にあるときのままである。

【０１００】

厳密なキーに基づくデータ操作は、様々な理由により、必ずしも望まれるほどには正確又は精密ではない場合がある。１つの理由は、１つのデータ又はデータ操作に関連する固有の曖昧さである場合がある。例えば、１つのデータが、２つ以上のグループに正当に属する場合がある。クラスタ化の方法によっては、固有の曖昧さにより、精密な分類が困難又は達成不能となる場合がある。例えば、従業員が、彼又は彼女が属する部門に従って分水される、上述の人材データベースにおいて、従業員は、マーケティングとＲ＆Ｄ等、同時に２つの部門に属する場合がある。当該従業員をいずれかの部門（マーケティング又はＲ＆Ｄ）に強制的に関連付けることは、誤解を招くことになる場合がある。単純に、当該従業員を両方の部民に関連付けることは、二重カウント問題を引き起こす場合がある。例えば、医療費等の費用が、同じ従業員に２回記録される場合がある。

【０１０１】

正確な分類が可能ではない場合があるもう１つの理由は、保留中の事象の結果が現行の分類に強い影響を与える場合があることである。例えば、慈善団体か又は非慈善団体かという組織の法的地位により、その納税義務に変化が生ずる場合がある。ＩＲＳ（国税庁）と当該組織との間に、当該組織が慈善団体として適格であるが故に税控除を受ける資格があるか否かに関する進行中の訴訟が存在すると更に仮定する。当該組織の年間予算において、当該組織の課税上の地位が慈善団体の地位にあると推定され、従って納税を保留して、より小さい予算が設定されている場合、且つ、その後、当該組織は、非慈善団体であるが故に慈善団体にのみ与えられる税控除を受けることができないと裁判所が決定する場合、当該年間予算は改訂されなければならない。このような状況に対処する旧来の方法は、当該予算に影響を及ぼす可能性のある不都合な判決を説明する注記を予算に付加すること

であることが多い。不都合な判決が下された場合、当該予算は改定されなければならない。しかしながら、予算を修正しなければならないことよりも更に悪いことに、当該予算が、他の事業分野、又は他の国において、他のアプリケーションによって使用されていた場合は、その波及効果を追跡するのが不可能である場合があるため、当該予算そのものを超える修正が不可能である場合がある。

【０１０２】

上記２つの例は、従来のデータ・アプローチが、クラスタへの曖昧な割り付け（「部分的メンバシップ」）を取り扱うのに如何に妥当ではない場合があるかを説明している。曖昧な割り付けは、クラスタへの１対１整合を、確実にすることができないか又は確実にしない方が良い場合に起こる。複数のクラスタへの割り付けを許すことによって呈される１つの難題は、会計の完全性を如何に保つかということである。部分的メンバシップの方法は、この目的のために使用することができ、この開示において後に詳細に考察する。曖昧なメンバシップを有する要素を含むクラスタを取り扱うのに、ファジーなデータ操作を使用してもよい。

【０１０３】

重複するメンバシップを伴う要素をクラスタが有し且つ２つ以上のクラスタに何等かのデータが関連付けられている場合、ファジーなロールアップというデータ操作を使用して、会計の完全性を保ち且つ可能な代替の割り付けと関連するエラーの範囲を報告しつつ、計算を実行してもよい。従業員が２つ以上の部門のために働いている場合、その従業員のための費用を、当該従業員の部分的メンバシップに従って、それらの部門の間で配分することができる。

【０１０４】

クラスタ・メンバシップが、上記法的な例におけるように、将来の事象次第であるか、又は上記銀行業務における取引先企業の特定におけるように、曖昧又は不十分な情報のために不確実であるか、の何れかである場合、例えば、グループによる金銭的な合計を演算するファジーなロールアップ操作は、会計の完全性を保ちつつ、この不確実さを反映すべきである。確かに、会社の税分類等の、不確実な将来の事象の場合、ある偶発性が生ずる。特定の選択肢への早まった割り付けは、企画立案及びリスク評価の目的に誤解を生ずる概念を生ずる場合がある。

【０１０５】

例えば、図２Ａにおいて、データ要素１２０がクラスタ１２２、クラスタ１２４、又はクラスタ１２６の何れに属するかは不確実である。データ要素１２０がこれら３つのクラスタ１２２、１２４、及び１２６に同時に属することもあるかもしれない。また、データ要素１２０が、ある特定の時刻には１つのクラスタに属するけれども、上記３つのクラスタの間で交代することもあるかもしれない。クラスタ１２２、クラスタ１２４、及びクラスタ１２６に属するデータ要素１２０のメンバシップはn_1、n_2、及びn_3によって表現され、n_1、n_2、及びn_3は分数である。データ要素１２０が等しい可能性で同時に３つのクラスタに属する場合、n_1、n_2、及びn_3の各々には分数１／３を割り付けることができる。この場合、クラスタ１２２、１２４、及び１２６に属するデータ要素１２０の部分的メンバシップの合計は１（１／３＋１／３＋１／３）である。データ要素１２０が、ある特定の時刻には１つのクラスタに属するけれども、上記３つのクラスタの間で交代する例においては、時刻t_1において、n_1、n_2、及びn_3は、１、０、及び０の値であってもよい。時刻t_2においては、n_1、n_2、及びn_3は、０、１、及び０の値であってもよい。n_1、n_2、及びn_3の値は変化する場合があるけれども、それらの値の合計は常に１であるべきである。

【０１０６】

上記銀行業務の例においては、曖昧に識別された取引先企業へのリスクの代替の割り付けに基づいて、各取引先企業の最大リスク及び最小リスクを知ることにより、何れの任意の取引先企業についても怒りうるリスクのより完全な概念が得られ、不確実な既知知識が明らかとなる。将来又は曖昧さの有望な解消についての現行の確信は、メンバシップの確

からしい確度を使用して暮らす他のメンバに重み付けを行うことによって組み込むことができ、これらの重み付けを時間と共に見直して、変化する既知知識を反映させてもよい。

【０１０７】

ファジーな結合の操作により、２つ以上のデータセットが共通の厳密なキーを共有していない場合でも、それらを組み合わせることができる。例えば、住所が逐語的には同一では無い場合でも、異なるデータベースからの顧客所帯記録を住所について結合することができる。１つのデータセットからの住所が不十分又は不正確である場合、候補整合となる第２のデータセットが複数存在する場合がある。この可能性に、ファジーな結合は適合する。

【０１０８】

ソート操作は、キーによってレコードを順序付け、レコードのグループに作用する（ロールアップ又は結合等の）キーに基づく操作に先立って使用されることが多い。個々のレコードが複数のクラスタのメンバである（と見込まれる又は実際にそうである）場合に、ファジーなソートを使用して、ファジーなロールアップ等の操作に先立ってレコードを順序付けてもよい。ソート順序の観念及び並べ替えの作用は、複数のクラスタの曖昧なメンバである個々のレコードを複製し、それらを最終的な順序付けにおいてそれらの関連するクラスタの各々に位置付けることによって拡張される。

【０１０９】

ファジーなデータ操作は、キーを厳密に整合させることに基づくキーグループの代わりにクラスタを使用するという点において、従来のデータ操作とは異なる。クラスタは、キーがＬｅｓｌｉｅである場合にＬｅｓｌｅｙを取り出す上記例を含む。また、クラスタは、Ｊｏｈｎ　Ｓｍｉｔｈはマーケティング部門において半分の時間しか働いていないので彼は厳密には当該部門には属していないにもかかわらず、彼を当該部門に分類する例をも含む。

【０１１０】

図２Ｂは、例示的なファジーなデータ操作を図解する。この例において、ファジーなデータ操作１８０は、キー１６０に作用し、データセット１５０を取り出す。キー１６０は従来のキーである。取り出されたデータセット１５０は、データ要素１５１、データ要素１５２、データ要素１５３、データ要素１５４、及びデータ要素１５５、の５つのデータ要素を含む。これら５つのデータ要素は、キー１６０に整合しない。しかしながら、それにもかかわらず、それらは上記データ操作によって取り出される。これが、ファジーなデータ操作が従来のデータ操作と異なるところである。キーが与えられると、従来のデータ操作は、そのキーに厳密に整合するデータを取り出す。しかし、ファジーなデータ操作は、そのキーに厳密には整合しないデータを取り出すことができる。

【０１１１】

クラスタの定義において基本的なこととして、重要なファジーなデータ操作は、異なるレコードにおけるデータの比較である。比較テストを使用して、各クラスタにどのレコードが属するかを特定する。アレンジによっては、比較テストは、各レコードから取得された、選ばれたフィールド値の採点された関数であり、２つのデータ（キーは１つのデータである）の間の定量化された差異を距離としてもよい。

【０１１２】

（ａ）２つのデータ間の距離

２つのデータの間の差異は直感的に単純であることが多い。例えば、Ｌｅｓｌｉｅという名前とＬｅｓｌｅｙという名前との間の差異は明らかであり、フル・タイムの従業員とパート・タイムの従業員との間の違いは明白である。しかしながら、２つのデータの間の差異を定量化又は測定するのは必ずしも単純ではない。ここで、２つのデータの間の距離を測定するのに使用することができる２つの方法について簡潔に検討する。データ間の差異を定量化する他の方法を下述する原理に基づいて容易に開発することができることが理解されるべきである。ファジーな整合手法及び距離測定の更なる例は、例えば、米国特許出願公開第２００９／０１８２７２８号に記載されている（引用により、本明細書に組み

込まれる）。

【０１１３】

（１）２つの単語間の距離

（例えば、所定の文字集合から形成される）２つの単語の間の距離を測定する方法（「編集距離」と称されることが多い）には、一方の単語から他方の単語になるのに何回の文字操作がかかるかをカウントすることが含まれる。この例においては、１回の文字操作には単一の文字が含まれる。文字は、様々な方法の何れでコード化されていてもよい。例えば、文字は、シングルーバイト若しくはマルチーバイトのコード化又は文字集合における文字を表現するのに使用されるコードーポイントの何れを使用しても、コード化することができる。レーベンシュタイン（Ｌｅｖｅｎｓｈｔｅｉｎ）編集距離は、一方の文字を他方の文字に変えるのに必要とされる文字の挿入、削除及び置換の数をカウントする。

【０１１４】

レーベンシュタイン編集距離及びその変形の限界は、それらはオンライン・ファジー整合文脈において（即ち、以前に見たことが無い問合せ単語を有しており、既存の参照集合において整合する変形を見出したい場合に）使用することができないという点である。変形を演算するための削除アルゴリズム（例えば、米国特許出願公開第２００９／０１８２７２８号。引用により、本明細書に組み込まれる）を代わりに適用することもできる。この方法において、単語間の距離は、各単語から整合する単語に至るまでに必要とされる削除の数をカウントすることによって特定される。図２Ｃは、ＬｅｓｌｉｅとＬｅｓｌｅｙとの間の削除距離を如何にして演算するかを示す。操作１０２により、「Ｌｅｓｌｉｅ」から「ｉ」が削除されて「Ｌｅｓｌｅ」が得られる。操作１０４により、「Ｌｅｓｌｅｙ」から「ｙ」が削除されて「Ｌｅｓｌｅ」が得られる。「Ｌｅｓｌｉｅ」と「Ｌｅｓｌｅｙ」との間の距離は１＋１（各単語から１回の削除。あるいは、これらの単語の一方にのみ、１回の削除と１回の挿入が作用する）。

【０１１５】

アレンジによっては、削除される文字の位置及び相対値を比較することによって、より精巧な採点を行うことができる。これにより、異なる種類の変化に異なる重み付けが適用される、重み付け採点が可能となる。例えば、置換は転位よりも重要ではないとすることもでき、又は「ｎ」による「ｍ」の置換は「ｋ」による「ｍ」の置換よりも重要ではないとすることもできる。

【０１１６】

上記削除アルゴリズムを、以下の方法により、参照用データセットのオンライン・ファジー検索に使用することができる。参照用データセットにおける各単語から１つ以上の文字を（必要と考えられる回数まで）削除することによって得られる全ての単語形成することにより、参照用データセットから削除辞書を構築する。（単語の長さに伴って削除の回数が増えて、より大幅な変形が可能となる場合がある。）元の単語及び削除された文字の位置の両方を、削除の結果として生ずる単語と一緒に記録する。検索がなされる場合、問合せ単語を処理して、１つ以上の文字を削除することによって得られる全ての単語を構築する。これらの単語の各々を上記参照削除辞書において調べて、対応する元の単語を見付け出す。（削除の位置の記録は整合を採点するのに使用される。）次いで、整合した元の単語を、データセットにおける通常の厳密な検索／ルックアップにおいて使用することができる。繰り返すと、問合せ単語が参照用データセットには現れない変形である場合でさえ、この方法は役に立つ。

【０１１７】

もう１つの例は、「Ｃｏｒｐ．」と「Ｃｏ．」との間の距離である。「Ｃｏｒｐ．」から「Ｃｏ．」にするのに、一方の単語における２回の削除（文字ｒの削除及び文字ｐの削除）が必要とされる。故に、２つの単語の間の距離が、各単語について整合する単語を得るのに削除操作が（最少で）何回必要であるかで定義される場合、「Ｃｏｒｐ．」及び「Ｃｏ．」は同じ単語「ｃｏｒｐｏｒａｔｉｏｎ」の２つの互換性のある省略形であるけれども、「Ｃｏｒｐ．」と「Ｃｏ．」との間の距離は２＋０であるとすることができる。ユ

ーザによって入力されたデータが「Ｃｏｒｐ．」を使用する一方でデータ操作によって使用されるキーが「Ｃｏ．」を使用する場合、単語を厳密に整合されることに依存する従来の方法は満足の行く結果を与えないであろう。例えば、ファジーなデータ操作において使用されるキーがＡＢＣ　Ｃｏ．である場合、キーに厳密に整合するデータ入力のみを取り出す従来の検索データ操作はＡＢＣ　Ｃｏｒｐ．を与えないであろう。ファジーな検索は、キーの特定の距離（例えば、２＋０又はそれより良好）以内にあるデータ入力を返すように構築することができる。このようなファジー検索においては、ＡＢＣ　Ｃｏ．というキーに対する整合として、ＡＢＣ　Ｃｏｒｐ．を返すことができる。

【０１１８】

あるいは、これら２つの単語は同義語として互換性があるので、「Ｃｏｒｐ．」と「Ｃｏ．」との間の距離をゼロ（０）として定義することができる。ユーザ指定の同義語を含むデータ入力を返すようにファジー検索を構築することができる。この例は、ファジーな操作が扱う必要があるかもしれない複雑さを紹介している。

【０１１９】

上記例においては、挿入又は削除等の操作に基づいて、挿入及び削除を両方とも１回の操作としてカウントして、距離を演算する。他のアレンジにおいては、重み付け操作に基づいて距離を演算することができる。重み付けを使用して、別のタイプの操作（例えば、削除）に対して、あるタイプの操作（例えば、挿入）にバイアスをかけることができる。あるいは、重み付けを使用して、別の個々の操作に対して、ある個々の操作にバイアスをかけることができる。例えば、スペースの削除に対応する操作に文字ｚの削除に対応する操作よりも少なく重み付けして、スペースの脱落は一般的なミススペル間違いであるのに対し、英単語におけるｚの挿入は、ことによると、ミススペル間違いというよりも、２つの英単語の間での真の差異であるという事実を反映させてもよい。

【０１２０】

例えば、「ｓｕｎｓｈｉｎｅ」と「ｓｕｎ　ｓｈｉｎｅ」との間の距離は、スペースの１回の挿入である。「ｚｉｎｃ」と「Ｉｎｃ」との間の距離は、文字ｚの１回の挿入である。距離の計算において個々の操作が重み付けされない場合、これら２つの距離は等しい（１操作）。１操作の距離以内であれば何れの整合をも返すようにファジーな検索操作が構築されている場合、キー「ｓｕｎｓｈｉｎｅ」による検索は「ｓｕｎ　ｓｈｉｎｅ」を返すであろうし、キー「Ｉｎｃ」による検索は「ｚｉｎｃ」を返すであろう。

【０１２１】

しかし、重み付け操作を使用する場合、これら２つの距離（「ｓｕｎ　ｓｈｉｎｅ」と「ｓｕｎｓｈｉｎｅ」との間の距離、及び「ｚｉｎｃ」と「Ｉｎｃ」との間の距離）を異なるようにすることができる。例えば、スペースの挿入を０．５の係数によって重み付けして、スペースの挿入が、タイプミスによって、より起こり易いという事実を反映させてもよい。文字ｚの挿入を１の係数によって重み付けして、余分な文字ｚは手違いによって追加され難いという事実を反映させてもよい。図２Ｄは、「ｓｕｎ　ｓｈｉｎｅ」と「ｓｕｎｓｈｉｎｅ」との間の距離が０．５操作である一方、「Ｚｉｎｃ」と「Ｉｎｃ」との間の距離が１操作であるように操作を重み付けした場合を示す。

【０１２２】

０．５文字操作の距離以内であれば何れのキーの整合をも返すようにファジー検索のデータ操作が構成されるアレンジにおいては、キー「ｓｕｎｓｈｉｎｅ」の検索は「ｓｕｎ　ｓｈｉｎｅ」を返すであろう。しかし、キー「Ｉｎｃ」の検索は「Ｚｉｎｃ」を返さないであろう。

【０１２３】

アレンジによっては、より巧妙な重み付けオプションを定義してもよい。

【０１２４】

（２）英国郵便番号（Ｂｒｉｔｉｓｈ　ｐｏｓｔａｌ　ｃｏｄｅ）間の距離

ファジーな整合が有用なもう１つの用途は、同じ所帯について重複レコードを含む会社の顧客住所データベースで役に立つ。同じ所帯についての複数の入力は、当該所帯に関連

する郵便番号における誤植又は当該所帯に関連する名前のミススペルによって起こる場合がある。

【０１２５】

起こり得る誤植には、スペースの脱落又は挿入、文字の脱落又は挿入、及び文字のミスタイプを含む。同じ郵便番号においてユーザが２つの誤植をすることはあまり起こらないけれども、珍しくはない。同じ郵便番号においてユーザが３つの誤植をすることはあまり起こらないけれども、不可能ではない。

【０１２６】

図３は、郵便番号における誤植によって起こり得る重複レコードを示す。顧客住所データベース３００において、Ｊｏｈｎ　Ｓｍｉｔｈという名前の下に、Ｊｏｈｎ　Ｓｍｉｔｈ　ｏｘ２６ｑｔ、Ｊｏｈｎ　Ｓｍｉｔｈ　ｏｘ２　６ｑｔ、Ｊｏｈｎ　Ｓｍｉｔｈ　ｏｘ２６ｑｙ、Ｊｏｈｎ　Ｓｍｉｔｈ　ｏｘ２　６ｑｙ、Ｊｏｈｎ　Ｓｍｉｔｈ　ｏｘ２　６ｑｘ、及びＪｏｈｎ　Ｓｍｉｔｈ　ｏｘ２　７ｑｙという、６つの入力が存在する。レコードの全てのペアの間の距離が、レコードのペアを繋ぐ線の隣に表示されている。

【０１２７】

真正のレコードにおける郵便番号から１＋１の削除距離以内の郵便番号を含む何れのレコードも、偽のレコード（誤って入力されたレコード）である可能性が最も高く、上記真正のレコードの重複として扱われると、会社が決定したと仮定する。更に、検索キーの１＋１削除距離以内にある全てのレコードを求める検索をファジーな検索として当該会社が定義したと仮定する。

【０１２８】

ある単語が別の単語の指定の距離以内にある場合、前者は後者の変形である。本例においては、指定の削除距離は１＋１（各単語から１回の削除）である。顧客住所データベース３００における各郵便番号間の距離情報は図３に列挙されている。図４に示されているように、図４に基づいて、各レコードについての変形を特定することができる。

【０１２９】

図４は可視化ツールであり、独特のシェードのボックスで各レコードを表現し、各レコードのボックスをその変形のボックスと重ね合わせることによって形成される。例えば、レコードＢ、Ｃ、及びＥはレコードＡの変形であるので、レコードＡのボックスはレコードＢ、Ｃ、及びＥのボックスと重なる。レコードＡ及びＦはレコードＥの変形であるので、レコードＥのボックスはレコードＡ及びＦのボックスと重なる。

【０１３０】

場合によっては、どのレコードが真正のレコードであるのかを会社が知っているかもしれないし、他の場合においては、どれが真正であるのか会社は知らないかもしれない。

【０１３１】

最初の例において、真正のレコードが「Ｊｏｈｎ　Ｓｍｉｔｈ　ｏｘ２　６ｑｙ」であるということを会社は知っている。検索キーとして「ｏｘ２　６ｑｙ」を使用するファジー検索を実行すると、以下の２つのレコードが取り出されるであろう：「Ｊｏｈｎ　Ｓｍｉｔｈ　ｏｘ２　６ｑｔ」及び「Ｊｏｈｎ　Ｓｍｉｔｈ　ｏｘ２６ｑｙ」。この会社は、真正のレコード「Ｊｏｈｎ　Ｓｍｉｔｈ　ｏｘ２　６ｑｙ」の重複として、同じクラスタにおけるこれら２つのレコードを扱うであろう。この会社は、これら２つの重複を除去したり、又は当該３つのレコードに共通キーを割り付けることによって、それらをグループ化したりすることを決定することができる。このグループは、ファジーなクラスタの例である。

【０１３２】

ファジーなクラスタ化は、厳密には整合しないけれども、互いに特定の距離以内にあるキーを有するデータを一緒にグループ化するデータ操作である。ファジーなクラスタ化は、上記例に示されているように、ファジー検索と関係している場合がある。上記の場合におけるように真正のレコードが知られている場合、ファジー検索は、真正のレコードの指定の距離以内にあるデータを取り出す。次いで、取り出されたデータはファジーなクラス

タを形成する。
【０１３３】
どのレコードが真正のレコードであるのかを会社が知らない場合においては、例えば、「Ｊｏｈｎ Ｓｍｉｔｈ ｏｘ２ ６ｑｔ」及び「Ｊｏｈｎ Ｓｍｉｔｈ ｏｘ２ ６ｑｙ」の両方が真のレコードであるかもしれず、レコードを如何にして一緒にグループ化するかについての先験的助言は存在しないので、単純なファジー検索によって、どのレコードが互いの重複であるのかを算定して、ファジーなクラスタを作り出すことはできない。このような場合においてファジーなクラスタを生成するのに採用することができる２～３のアプローチにつき、セクション（ｂ）において詳細に説明する。
【０１３４】
（３）定量化された差異の他の例
２つのデータの間の距離は、２つのデータの間の定量化された差異の一例である。２つのデータの間の差異を違う方法で定量化することもできる。
【０１３５】
アレンジによっては、整合するペアの間の類似性に基づいて当該ペアを採点するように採点システムを発展させてもよい。次いで、上記ペアの間の定量化された差異を、正規化された整合スコアの補数（ｃｏｍｐｌｅｍｅｎｔ）として定義することができる。
【０１３６】
ある事象（例えば、訴訟）の結果が保留されているために不確かさが存在する場合、１つのカテゴリ又はもう１つに属する１つのデータの可能性を使用して、当該データと当該カテゴリを表現するキーとの間の距離を定量化することができる。当該データとカテゴリを表現するキーとの間の距離は、カテゴリが２つしか無い場合に当該データ当該カテゴリに分類されるであろう可能性の補数として、又は、より多くのカテゴリが存在する場合に当該データ当該カテゴリに分類されるであろう可能性の共役（ｃｏｎｊｕｇａｔｅ）として、定義することができる。
【０１３７】
（ｂ）変形関係及び変形の（ファジーな）結合
結合操作においてペアとなるデータ要素と比較されるべき、それぞれのデータセットにおけるそれぞれのデータ要素からのオブジェクトは、１つのデータ又は複数のデータの組み合わせとして定義することができる。リレーショナル・データベースにおけるテーブルの行において、オブジェクトは、列における値、値の一部（例えば、サブストリング）、又は２つ以上の列からの値の組み合わせであってもよい。フィールドを含んでなる一連のレコードからなる、単層ファイル・データセットにおいては、オブジェクトは、１つのフィールドにおける値、１つのフィールドの一部又は２つ以上のフィールドの組み合わせであってもよい。文書において、これは、テキストの断片又はテキストのバラバラの断片の組み合わせであってもよい。
【０１３８】
オブジェクト｛ｋ｝の集合Ｓを考察する。Ｓにおける各オブジェクトｋは、変形｛ｖ｝と呼ばれる、変形オブジェクトの（ことによると、空の）関連する集合を有する。当該関係である下式
【０１３９】
【数１】

$$k \sim v$$

【０１４０】
は、「ｖはｋの変形である」と読まれる。アレンジによっては、関数ｓ（ｋ，ｖ）の下での２つのオブジェクトのスコアが閾値Ｔ未満である場合、それらは変形であると特定される。
【０１４１】

【数２】

$$s(k,v) < T .$$

【０１４２】
（採点関数によっては、代わりに、閾値を超えるとした方が便利である場合がある。）オブジェクト間の距離は、（上記において考察した文字列についての編集距離又は削除距離の場合、）単語又は句を比較するための採点関数を構築するための基礎として使用することができる。
【０１４３】
上記変形関係は同値関係（即ち、対称的であり、推移的性質（ｋ～ｋ’，ｋ’～ｋ”＝＞ｋ～ｋ”）を有する）である必要は無いけれども、時として同値関係である。変形関係は、たとえ同値関係でなくとも、対称であると仮定される。
【０１４４】
【数３】

$$k \sim v \Rightarrow v \sim k,$$

【０１４５】
即ち、ｖがｋの変形である場合、ｋはｖの変形である。
【０１４６】
２つ（又はそれ以上）のデータセットＡ及びＢの厳密な（内部）結合は、同一のオブジェクト（Ａにおけるk_A、Ｂにおけるk_B）を含む、レコード（行、文書等）のペアリングとして定義される。
【０１４７】
【数４】

$$k_A = k_B.$$

【０１４８】
上記オブジェクトk_A及びk_Bはキーと呼ばれる。
【０１４９】
変形（「ファジーな」）内部結合は、２段階で定義される。先ず、レコード（行、文書等）等のデータ要素の暫定的なペアリングが作られる。１つのアレンジにおいて、Ａにおけるk_Aは、Ｂにおけるv_{Bn}とペアとなる（Ｋａ～v_{Bn}）。次いで、k_A及びv_{Bn}と関連付けられたレコードのペアが評価され（Ｅ（k_A，v_{Bn}））、どのレコードのペアを保持するかが特定される。（厳密な場合においては、全てのペアが保持され、整合ステップ及び採点ステップが単一の比較に融合する（$k_A = k_B$）。）評価操作は、一般に、ペアリングに使用されるオブジェクト以外の、ペアとなったレコードにおける更なるオブジェクトの比較を含む。アレンジによっては、評価操作はスコアを生成し、このスコアは、整合を識別するには整合閾値を超えなければならない。半結合及び外部結合は、整合するレコードが全く見付からない（又は保持されない）場合に、対抗するレコードに対してヌル値が指定される場合に例えて定義される。
【０１５０】
最も単純な暫定的ペアリングは下式によって与えられる。
【０１５１】
【数５】

$$k_A \sim v_{Bn},$$

【０１５２】
即ち、Ｂにおけるk_Aの変形の集合である。このペアリング（「整合」）ステップは、提案されたペアリングが保持されるべきか否かを特定する評価（「採点」）ステップによって補足される。

【０１５３】
上記変形ペアリング（$k_A \sim v_{Bn}$）への拡張の階層が存在する。第１の一般化は、下式によって与えられる更なるペア（k_A，v_{Bnm}）を付加するによって、上記ペア（k_A，v_{Bn}）を拡張することである。
【０１５４】
【数６】

$$k_A \sim v_{An},\ v_{An} \sim v_{Bnm}.$$

【０１５５】
即ち、k_Aは、Ａにおけるk_Aの変形のＢにおける変形とペアとなる。変形関係が同値関係ではない場合、これは、Ｂにおける要素のより大きい集合に到達する。ここで留意すべきは、この操作は対称的ではない（k_Aに到達することができないオブジェクトv_{Bnm}がＢに存在する場合がある）ということである。即ち、下式が成立しても、
【０１５６】
【数７】

$$v_{Bnm} = k_B, k_B \sim v_{Bi}, v_{Bi} \sim v_{Aij},$$

【０１５７】
$V_{Aij} = k_A$は成立しない。これは、Ｂにおけるk_Bの何れの変形も、変形としてk_Aを有することを必要としないためである。せいぜい、k_Bは、その変形の１つとしてk_Aの変形を有することを要求される程度である。
【０１５８】
変形の変形への更なる拡張等が可能である。特に、下式が成立するペア（k_A，v_{Bnmj}）による（k_A，v_{Bn}）の拡張は、
【０１５９】
【数８】

$$k_A \sim v_{An},\ v_{An} \sim v_{Bnm},\ v_{Bnm} \sim v_{Bnmp},$$

【０１６０】
以下の意味において対称的である。（上記操作によってk_Aとペアとなる）Ｂにおける要素k_Bを仮定すると（即ち、ｎ、ｍ、ｐによっては、$k_B = v_{Bnmp}$）、下式が成立する要素$v_{Aijl} = k_A$が存在する。
【０１６１】
【数９】

$$k_B \sim v_{Bi},\ v_{Bi} \sim v_{Aij},\ v_{Aij} \sim v_{Aijl}.$$

【０１６２】
言い換えれば、同じ変形性号手順を逆に適用すると、逆転したペアが含まれる。Ａにおけるオブジェクトから到達されるＢにおける全てのオブジェクトは、同じ手順により、今度はＡにおける元のオブジェクトに到達することができる。
【０１６３】
３つ以上のデータセットへの拡張は、データセットをペア毎に結合し、結果として得られるペアのカルテシアン積をとることによって定義してもよい。Ａ、Ｂ、及びＣを下式のように結合する。
【０１６４】
【数１０】

$$k_A \sim v_{Bn},$$

$$k_A \sim v_{Cm},$$

$$\Rightarrow \quad (k_A, v_{Bn}, v_{Cm}).$$

【０１６５】
より高次の拡張は、ペア毎のやり方で上記のように（例えば、変形の変形）定義されるより高次の拡張を使用することによって得られる。状況によっては、随意に、ＢとＣとの間の変形関係が要求される場合がある。例えば、幾つかのｎ、ｍについて、下式が成り立つことが要求される場合がある。
【０１６６】
【数１１】

$$v_{Bn} \sim v_{Cm}$$

【０１６７】
より高次の変形の使用は、ＢとＣとの間の直接の繋がり（当該関係は、勿論、Ａによって既に仲介されている）を確立するのに必要とされる場合がある。
【０１６８】
上記において考察したように、変形関係の１つの有用なソースは、編集距離によって関係付けられる単語をペアにすることである。単語間で考察される編集距離が１に限定される場合、これにより、データセット内のペアリングの特定の集合が変形として認められる。例えば、「Ｓｍｉｔｈ」は、「Ｓｍｔｈ」、「Ｓｍｉｔｈ２」及び「Ｓｍｙｔｈ」を変形として有する。「Ｓｍｉｔｈ２０」は「Ｓｍｉｔｈ」の変形ではないけれども、編集距離１の変形関係は推移的ではないので、それは「Ｓｍｉｔｈ２」の変形である。
【０１６９】
上記変形結合は、単一の単語又はフィールド全体を変形キーとして使用することができる場合に使用することができる。例えば、データセットを検索することは、変形キーワードを使用する変形結合として明確化することができる。問合せ句は、キーワードの集合に分解され、その各々は、対象となるデータセットからの単語のインデックスにおける、その変形と整合される。当該インデックスは、対象となるデータセットにおける所定のフィールドにある単語を含むレコードについて、当該単語をレコード識別子（キー）とペアにする。対応するレコード識別子のリストを各キーワードの当該インデックスとの変形整合から得て、これらのリストの積集合を得る（ｉｎｔｅｒｓｅｃｔ）ことにより、１つ以上のキーワードを共有するレコードを見付け出すことができる。返されたレコードのリストは、整合するキーワードの組み合わせにスコアを割り付けることによって順序付けられる。このスコアは、データセットにおける各キーワードの相対頻度（「逆文献頻度」）（対象となるデータセット・レコードにおけるキーワードの位置と比較した、問合せ句におけるそれらの相対的位置（例えば、順序、隣接姓（ａｄｊａｃｅｎｃｙ））、又は問合せ句からの単語の不在）を考慮に入れてもよい。また、キーワードを関連度の他の尺度と関連付けて、採点をより識別力のあるものとすることもできる。
【０１７０】
また、変形結合を、単一の単語についてのルックアップとして使用することもできる。例えば、１つのデータセットにおけるある顧客を、ファースト・ネーム、ラスト・ネーム及び住所によって識別してもよい。この顧客を、第２のデータセットにおいて、ラスト・ネームによって探して、当該整合をファースト・ネーム及び住所によって確証してもよい。当該整合手順は、ソースとなるデータセットにおけるラスト・ネームから、対象となるデータセットにおける様々なラスト・ネームの集合を使用して、整合候補の集合を識別し、取り出すというものである。これらの候補は、ファースト・ネーム及び住所について更に比較され、一致の度合いが整合を識別するのに十分であるかどうかが特定される。
【０１７１】
例えば、ソースとなるデータセットにおけるレコードが、「Ｐａｕｌ，Ｓｍｉｔｈ，２０　Ｗａｌｋｅｒ　Ｓｔｒｅｅｔ」であり、対象となるデータセットにおける整合する変形の集合が（Ｓｍｉｔｈ，Ｓｍｙｔｈ，Ｓｍｉｔｈｈ）である。対象となるデータセットにおける関連付けられるレコードは以下の通りである。
【０１７２】

【表１】

1,Paul,Smith,20 Walken St

2,Robert,Smith,1532 East Grove Ave

3,P,Smyth,19 Western Ave

4,Pal,Smithh,20 Walker Street

【０１７３】

確証アルゴリズムは、レコード１及び４が整合となるのに十分に近いことを見出すことができる。これらのレコードを、ルックアップ（検索）によって、又は変形結合において、返すことができる（ここで、２つのデータベースは互いに対して編成されている）。

【０１７４】

あるいは、ことによると、ソースとなるデータセットにおいて、元の「Ｓｍｉｔｈ」が、対象において整合するレコード「５，ｐ，Ｓｍｉｔｈ２０，Ｗａｌｋｅｒ　Ｓｔｒｅｅｔ」を有する「Ｓｍｉｔｈ２」という変形を有する。

【０１７５】

「Ｓｍｉｔｈ２０」は「Ｓｍｉｔｈ」の直接の変形ではないけれども、ソースとなるデータセットにおける変形「Ｓｍｉｔｈ２」からは到達することができる。

【０１７６】

変形結合のもう１つの使用は、クラスタ化に先立つスーパークラスタの定義である。これについては、クラスタ化が定義されてから、後に考察する。

【０１７７】

（ｃ）クラスタ化及び部分的メンバシップ

多くの厳密なキーに基づく操作は、共通キー値を共有する集合にレコードをグループ化することを必要とする。これらの集合は、時として、「キーグループ」と呼ばれる。例えば、ロールアップ操作は、あるキーグループにおけるレコードに亘ってデータを組み合わせたり、統合したりする。カウント、総計、最大値又は最小値、値のベクトル、一意的な値への重複等は全て、ロールアップ操作によって演算することができる。レコードのグループを単一のレコードに集約する操作は何れも、ロールアップ操作として解釈することができる。

【０１７８】

（独立した処理のためのデータ・パーティションにデータが分離される）データ並列処理は、同じキーグループに属する全てのレコードが同じデータ・パーティションに存在することを保証するために、キーに基づく分割に依存することが多い。ロールアップ及び結合等の操作は、順次（非並列）処理における結果と同じ結果を生成するために、このことに依存する。

【０１７９】

キーグループの集合は、全てのレコードの集合を互いに素な集合へと分割することを構成する。全てのレコード（オブジェクト）は唯一のキーグループのメンバである。クラスタは、キーグループの観念を重なり合う集合を含むパーティションへと一般化する（メンバシップはキーの厳密な一致によって特定されない）。

【０１８０】

各々が重み付けｗ（ｋ，Ｃ）を有する、オブジェクトｋの重なり合うかもしれない集合｛Ｃ｝（クラスタと呼ばれる）の集まりに集合Ｓを分割（分解）することを考察する。あるオブジェクトｋは２つ以上のクラスタＣのメンバであるかもしれず、もしそうである場合、そのクラスタ・メンバシップは、曖昧又は部分的であると言われる。Ｃにおけるｋに割り付けられる重み付けｗ（ｋ，Ｃ）は、Ｃにおけるｋの「部分的メンバシップ」を定量化し、時として曖昧さの尺度と呼ばれる。クラスタＣは、ペアＣ＝｛（ｋ，ｗ（ｋ，Ｃ））｝の集合として表示することができる。ｗ（ｋ，Ｃ）＝０である場合、ｋは「Ｃのメンバではない」と言われる。ｗ（ｋ，Ｃ）＝１である場合、ｋは「確かにＣのメンバである

」と言われる。固定されたｋについて、Ｃを亘っての重み付けの合計は１に等しい（Ｓにおける確かなメンバシップに対応する）。

【０１８１】

【数１２】

$$\sum_{C} w(k,C) = 1.$$

【０１８２】

重み付けの割り付けは、ルールＲと関連付けられ、Ｒによって表示される。所定の集合Ｓは、概して、クラスタの集まりへの２回以上の分割、及び異なるルールの下での各オブジェクトへの２回以上の重み付けを認める。一般に、異なるルールに関連付けられる重み付けを組み合わせることはできない。

【０１８３】

クラスタＣ＝｛（ｋ，ｗ（ｋ，Ｃ））｝の補数は、集合｛（ｋ，１－ｗ（ｋ，Ｃ））｝であると定義される。特に、当該補数は、重み付け１を用いて、Ｃにはないオブジェクトを表す。クラスタ｛Ｃ｝の集まりがＳにかからないか、又はＣを亘ってのｋについての重み付けの合計が１に等しくない場合、Ｓにおける｛Ｃ｝の和集合の補数は、当該集まりに付加されたものと仮定される。

【０１８４】

２つのクラスタを組み合わせて単一のクラスタとし、それにより、それらの重み付けを合計することによって、分割を粗くすることができる。

【０１８５】

【数１３】

$$C_1 + C_2 = \{(k, w(k,C_1) + w(k,C_2))\}.$$

【０１８６】

このプロセスを逆転させて、新しいクラスタの中の各オブジェクトに重み付けを割り当てることによって、クラスタを更なるクラスタに分解することができ、それ故に、新しい重み付けの合計は、元の重み付けに等しい。例えば選択基準を適用した後に、オブジェクトの重み付けを減じることによって、クラスタからオブジェクトを取り去ることができる。

【０１８７】

状況によっては、重み付けが０＜＝ｗ（ｋ，Ｃ）＜＝１を満足する場合、これらの重み付けは「ｋがＳにおけるクラスタＣのメンバである可能性」との解釈を認めることができるけれども、一般には、クラスタの当該定義は非統計的である。負の重み付け及び１よりも大きい重み付けの可能性は排除されないけれども、固定されたｋについての重み付けの合計は１でなければならない。

【０１８８】

Ｃ$_1$＋Ｃ$_2$等の組み合わせは、原則として、形成することができるけれども、それらは、あるオブジェクトについてあり得る最大の重み付けの値によって反映されるように（例えば、この場合は２）、オブジェクトの複数のコピーを含むクラスタを構築することに該当する。各オブジェクトについての重み付けの合計は１であるという条件は、集合Ｓにおいて各オブジェクトのコピーが１つだけ存在することを想定している。それが当てはまらない場合、重み付けの合計の値は、それ相応に変化する場合がある。一般に、オブジェクトによって重み付けの総計が変動することを排除するものは無い。

【０１８９】

部分的メンバシップを有するクラスタは、集合における各オブジェクトに重み付けを割り付けるメンバシップ関数を有するオブジェクトの集合として記述することができるので、ファジーな集合の観念に類似している。しかしながら、ここでは、孤立しているファジーな集合には重きが置かれないけれども、分割の要素としてのクラスタに重きが置かれる

。特に、重み付けは、分割の性質であって、孤立しているクラスタ内のオブジェクトのものではない。クラスタ内のオブジェクトに割り付けられる重み付けは、他のクラスタに対して起こり得る割り付けによって影響を受ける。焦点は、クラスタ内のメンバシップ関数からクラスタを亘ってのオブジェクトへのメンバシップの割り当てにシフトする。

【０１９０】

部分的メンバシップを有するクラスタは多くの状況において自然に生ずる。最も単純な言葉によれば、部分的メンバシップは、オブジェクトのクラスタへの割り付けにおける曖昧さの結果である。厳密なキーが存在する場合、オブジェクトが何れのキーグループに属するのかという疑問は存在しない。厳密に一致することを必要としない１つのデータ又はデータの組み合わせにメンバシップが基づく場合、メンバシップ決定はあまり明確ではない。

【０１９１】

以下は、部分的メンバシップに繋がり得る幅広いデータ品質問題の例である。

【０１９２】

データは、クラスタ化のルールと比べて、本質的に曖昧である。時として、クラスタは、それらの定義が排他的メンバシップを推定していないという単純な理由により、重なり合う。ある会社において２つの部門で働く従業員について考察する。従業員のリストが部門によってクラスタ化されている場合、当該従業員は２つのクラスタに現れるのが正しい（これが事柄の真の状態を反映している）。この場合、部分的メンバシップを設定して、当該従業員が各部門のために働く時間の分配を反映させることができる。このことは次に、別個の部門に同じ識別情報を有する２人の別個の従業員が存在するという誤った結論を導き出す機会を減らす。

【０１９３】

データは不完全である場合がある。フィールドにおける変形の単語が識別を曖昧にする場合がある。住所を所帯に割り付ける際に、１つの住所レコード上の番地が１２であり、その通りには１２という番地は無い場合がある。代わりに、１、２及び２１という番地が存在する場合がある。最初の２つの番地への０．３の重み付け及び最後の番地への０．４の重み付けにより、挿入エラーよりも僅かに大きい転移エラーの可能性を反映させることができる。

【０１９４】

データは不十分である場合がある。クラスタへのぇっていてきな割り付けを行うのに必要な情報が欠落している場合がある。例えば、住所レコードを所帯に割り付ける問題を考察する。各々の一意的な番地、通り、市、郵便番号の組み合わせには、一意的な所帯番号が割り付けられる。クラスタ化アルゴリズムは、通りの名前及び市の名前の変形のスペルに寛容である場合があり、それ故に、正しい所帯に割り付けられる全ての住所が同一である必要は無い。しかしながら、住所から番地が欠落している場合、信頼の置ける割り付けを行うのに十分な情報が無い。できるだけ多くの情報を保持するために、不十分なレコードを、入手可能な情報と一致する各所帯に部分的に割り当ててもよい。可能性がある番地が５つ存在する場合、各番地クラスタにおける重み付けは、各番地の等しい確度を反映させて、０．２とすることができる。異なる見地で、会社名によって表示された会計における負債について考察する。銀行は、このデータを統合して、国毎に各会社と関連付けられる未払いの負債を特定することが必要である。これらの会社名の中には、「ＡＣＭＥ　ＳＥＲＶＩＣＥＳ　ＬＩＭＩＴＥＤ　（ＡＵＳＴＲＡＬＩＡ）」、「ＡＣＭＥ　ＳＥＲＶＩＣＥＳ　ＬＩＭＩＴＥＤ　（ＣＡＮＡＤＡ）」及び「ＡＣＭＥ　ＳＥＲＶＩＣＥＳ　ＬＩＭＩＴＥＤ」がある。最初の２つの各々は別個のクラスタに行くけれども、３つ目は、最初の２つの各々に対して同等の整合であり、国の識別子が足りない。最初の２つのクラスタの各々における３つめの会社に０．５の重み付けを行うことにより、会社情報の不十分さを反映させる。

【０１９５】

データ又は分類は、本質的に不確かである場合がある。クラスタのメンバシップは、将

来の事象の結果に基づいている場合がある。資産及びそれらの価値のリストを含むデータセットについて考察する。資産は、所有者によってクラスタ化されるべきである。しかしながら、特定の資産の所有権についての訴訟が係争中である。その資産を可能性がある所有者の何れかに結び付けることは負け博打（ｌｏｓｔ　ｂｅｔ）である場合があるけれども、その資産を単純に無視することもできない。
訴訟の予想結果の知見の現状を反映する部分的メンバシップを用いて当該資産を各所有者に割り付けることは、現行の知見と一致した最も公平で有益な当該資産の処分を生ずる。
【０１９６】
（ｄ）クラスタ化
クラスタ化は、クラスタ・メンバシップ基準に基づいて、レコードをクラスタにグループ化する行為である。厳密な場合、各レコードからのオブジェクト（キー）が、他のレコードにおける対応するオブジェクトと厳密に整合され、クラスタ（又は「キーグループ」）は共通キーを共有するレコードの集合である。ファジーな場合、クラスタ・メンバシップを、各レコードにおけるオブジェクト間の変形関係によって特定してもよい。（より一般的には、クラスタ・メンバシップ基準があり得る。）データセットにおける全てのレコードを互いに比較しなければならないことを回避するために、スーパークラスタ・キーを使用して、全体集合を部分集合に分割して、１つのスーパークラスタ内のレコードに対して相互比較を制限する。
【０１９７】
多くの場合、スーパークラスタは、厳密なキー（例えば、郵便番号）によって定義される。変形結合は、変形オブジェクトを使用してスーパークラスタを定義することを可能とする。例えば、スーパークラスタは、所定の郵便番号の変形である全ての郵便番号を含むレコードの集合として定義することができる。例えば、英国の郵便番号ＯＸ２　６ＱＹを考えると、変形郵便番号ＯＸ２６ＱＹ及びＯＸ２　６ＱＴは両方とも編集距離が１の変形であり、後者はそれ自体が有効な郵便番号である。各変形郵便番号からのレコードを有効な整合として認めることにより、郵便番号におけるエラーに寛容なクラスタを生ずることが可能となる。
【０１９８】
もう１つのアレンジにおいて、各レコードにおける選ばれたフィールドから（例えば、相対頻度に基づいて、最も長い又は最も重要な単語から）単語の断片を取り、この断片の変形を使用してスーパークラスタを識別することによってスーパークラスタを形成してもよい。これは、同じクラスタのメンバであるべき２つのレコードについて、それらが特定の単語を共有する可能性が非常に高いけれども、それらの単語が直接的な変形である必要は無く、ましてや等しい必要は無い場合に適している。変形断片をスーパークラスタ・キーとみなすことによって、その単語の残りの部分が変形として許容されるよりも異なっているレコードが認められる。クラスタ・メンバシップを特定するには、各レコードにおけるフル・ワード及び他のオブジェクトの徹底的な比較が必要とされる。
【０１９９】
例えば、ドイツ語の通りの名前を比較する場合、Ｇｒａｆ　ｖｏｎ　Ｓｔａｕｆｆｅｎｂｅｒｇという通りは、Ｓｔａｕｆｆｅｎｂｅｒｇを何等かの形で吹くことが必要とされる場合がある。サンプル・データにおいて、Ｓｔｒａｓｓｅが省略され、先行する単語に連結されて、Ｇｒａｆ　ｖ．　Ｓｔａｕｆｆｅｎｂｅｒｇｓｔｒ等の入力を生じていることが観察された。各通りの名前における最も長い単語の最初の５つの文字の編集距離が２の変形によって定義されるスーパークラスタは、「ｓｔａｕｆ」及び「ｓｔａｕｆｆ」の両方を含むであろう。Ｓｔａｕｆｆｅｎｂｅｒｇ及びＳｔａｕｆｆｅｎｂｅｒｇｓｔｒの両方を含むレコードが比較のためにスーパークラスタに含まれ、好適な採点関数がそれらに同じクラスタを割り付けるであろう。一方、最も長い単語の編集距離が２の変形に基づくスーパークラスタは、これら２つの通りの名前を別個のスーパークラスタに区分けするであろう。そして、それらが一緒にクラスタ化されることはできない。
【０２００】

スーパークラスタを思慮深く選ぶことは、クラスタ化方法の性能及び精度のために重要である。例えば、スーパークラスタが大き過ぎる場合、性能を劣化させ得る、多くの報いの得られない比較がなされる場合がある。あるいは、スーパークラスタが狭過ぎる場合、許容可能な整合が欠落し、精度が劣化する場合がある。

【０２０１】

（ｅ）部分的メンバシップ

データ入力オペレータが、新しい顧客をデータベースに追加するために、アプリケーションにおけるフォームに記入していると仮定する。名前をフォームに入力する際に、アプリケーションは、名前の参照リストと対照して、入力を確認する。上述のような削除アルゴリズムによるファジー検索を使用して、アプリケーションは名前の変形スペルを検出し、上記参照リストから他の可能性のリストを返すことができる。オペレータがファースト・ネームのフィールドに「Ｊａｍｅ」を入力すると仮定する。アプリケーションは、アルファベット順の他の可能性の以下のリストを（その名前を含むデータベースにおけるレコードのカウントと一緒に）返すことができる。

【０２０２】

【表２】

Jaime 250

James 13359

Jamie 339

Jane 9975

【０２０３】

これらは全て、１回の挿入及び／又は１回の削除だけ、Ｊａｍｅとは異なっており、候補選択肢である。

【０２０４】

上記リストのオペレータに対する実用性を向上させるために、曖昧さの尺度を特定するための様々なモデルの何れかを使用して、選択肢を優先順位付けすることができる。曖昧さを定量化するための３つの例示的な方法は、１）均等分配、２）統計的頻度、及び３）誤差模型である。

【０２０５】

均等分配アプローチにおいては、各選択肢は等しく起こり得るものとして扱われる。ここで、Ｊａｍｅが上記選択肢の何れかである確度は四分の一である。選択肢のアルファベット順のリストは、概して、黙示的な均等分配アプローチを示唆する。

【０２０６】

統計的頻度アプローチにおいては、（データベースのテーブル自体等の）参照集合を各名前についての観測度数として使用する。上記リストを示されているカウントの降順でソートする場合、最も起こり得る修正はＪａｍｅｓであり、続いてＪａｎｅ等である。

【０２０７】

誤差模型の第３の方法は、とりわけ、言語及び入力のモードによって、特定の種類の誤差が他の種類の誤差よりも起こり易いという観測結果に基づく。熟練したオペレータによるキーボード入力については、文字を飛ばしたり又は余分な文字を挿入したりするよりも、置換エラーの方がよく起こる場合があるかもしれない。同様に、電話越しに顧客によって与えられるスペルを記録しているオペレータについては、音声学的に類似の文字の名前を含む転記エラーの方が、他の種類のエラーよりも起こり易い可能性がある。何れの場合においても、ここでは、Ｊａｎｅが最も可能性のある修正であろう。この方法を使用するために、起こり得るエラー及びそれらの相対的重要度を分類するモデルを開発し、適用することができる。このようなモデルは、米国特許出願公開第２００９／０１８２７２８号において紹介されている、ＷＦＳ（単語頻度優位性：ｗｏｒｄ　ｆｒｅｑｕｅｎｃｙ　ｓｉｇｎｉｆｉｃａｎｃｅ）ファイルの統計的分析から作り出すことができる。

【０２０８】
アプリケーションが顧客住所（ここでは問合せ住所と称する）を含むレコードをマスタ顧客住所テーブルに整合させて、整合が見付かった場合には、存在する所帯キーを取り出し、又はさもなければ新しいキーを作り出すと仮定する。問合せ住所は、マスタ顧客住所テーブルにおける住所と厳密には整合しない場合があるので、ファジーな整合を使用することができる。その上、問合せ住所は不十分又は不正確である場合がある。このことは、２つ以上の実在する住所が問合せ住所に対する整合となる場合があることを意味する。整合の品質を定量化するには、整合の曖昧さの尺度を有することが有用である。
【０２０９】
例えば、住所は番地を有しておらず、マスタ顧客住所ファイルは同じ通りの住所を有する複数の入力を有している場合がある（番地は無視する）。問合せ住所は特定の町及び郵便番号を有するＬｏｗｅｒ　Ｓｔｒｅｅｔであると仮定する。郵便番号についてのファジー検索は、同じ郵便番号又は変形郵便番号を有する、見込みのある住所レコードのリストを返す。問合せ住所上の郵便番号、町、通り、及び番地のフィールドは、ファジーな整合プロセスの一部として、見込みのある住所の対応する各フィールドと比較され、採点される。この例において、問合せ住所の、通り、町及び郵便番号に厳密に整合する２つのマスタ・レコード（２　Ｌｏｗｅｒ　Ｓｔｒｅｅｔ及び３　Ｌｏｗｅｒ　Ｓｔｒｅｅｔ）が存在すると仮定する。各有力候補は問合せレコードとの整合の品質が等しく、問合せレコード上の番地が欠落しているので、既存のデータではこれを改良することはできない。均等分配の尺度の下では、何れの所帯に対しても、実際の整合の確度が等しい。
【０２１０】
あるいは、番地は埋まっているけれども無効であり、実在する住所の何れにも符合しないと仮定する。問合せ住所は１２　Ｌｏｗｅｒ　Ｓｔｒｅｅｔであるけれども、参照用のポスタル・アドレス・ファイル（ポスタル・サービス（Ｐｏｓｔａｌ　Ｓｅｒｖｉｃｅ）から入手可能な全ての有効な住所のリスト）に照らした確認により、その住所を有する家が無いことが示されると仮定する。その郵便番号において整合する住所は、上記のように、２　Ｌｏｗｅｒ　Ｓｔｒｅｅｔ及び３　Ｌｏｗｅｒ　Ｓｔｒｅｅｔである。住所入力についての誤差模型は、３に対する１２の整合よりも、２に対する１２の整合の方を選ぶ。これにより、整合の確度の重み付けにバイアスをかけて、２　Ｌｏｗｅｒ　Ｓｔｒｅｅｔの住所との整合に有利に働くであろう。
【０２１１】
最後に、問合せレコード上の番地が埋まっていて且つ有効な郵便の宛先である場合、住所入力についての誤差模型は、当該住所が実在する住所のエラーに対して新しい確度を定量化することができる。
【０２１２】
（ｆ）データ品質の定量化
曖昧さの尺度は、データ品質を測定する、より広い情況においても適用可能である。企業及び組織は、それらのデータ、特にそれらのマスタ・データの品質について懸念するけれども、現在のところ、最も明らかなデータ品質問題以外のものを定量化することは困難である。上記に示したデータ品質問題の短いリストの中で、幾つかのデータ品質測定システム（例えば、米国特許出願公開第２００５／０１１４３６９号を参照。引用により、本明細書に組み込まれる）は、主に、１つ（データ妥当性）に直接的に取り組む。データは網羅的に目録に載せられ、そのデータ型及びユーザが定義する妥当性の様々な尺度（有効値のリストを含む）に照らして、妥当性がチェックされる。
【０２１３】
レコードのフィールド内の不十分な入力の証拠は、埋まっていない（空又はヌル）入力の数から推察することができるけれども、これでは、欠落した情報の重要性は定量化されない。マスタ顧客住所リストの場合を考察する。顧客住所入力から市が欠落しているけれども、有効な郵便番号及び通りの名前がある場合、これは何等かの曖昧さを加えるであろうか？又は、ことによると、ポスタル・アドレス・ファイルのような参照集合を使用して

、手元の情報から住所を効率的に完成することができるか？もし住所から番地が欠落していたらどうなるか？残りの住所を何軒の家が共有しているか？幾つかのデータ品質問題はデータ・クレンジングによって解決することができ（欠落している市）、他は解決することができない（欠落している番地）。データに存在する本質的な曖昧さの尺度が必要とされている。

【０２１４】

各アドレス入力を参照用データセットと比較することにより、当該入力における曖昧さの尺度を演算することができる。曖昧さのレポートには、曖昧さが全く無い入力の割合が含まれる場合がある。曖昧さを有する入力については、レポートは、Ｋ個の選択肢を有する入力の数のヒストグラム・プロット（又は、特定の瓶／範囲の選択肢、「変形」とも呼ばれる）を示す場合がある。また、最も大きい曖昧さを有する最初のＮ個の入力のリストもあるかもしれない（ここで、Ｎはユーザによって指定される入力数である）。データセット全体について不十分なデータと関連付けられる曖昧さを定量化する１つの要約統計量は、入力当たりの選択肢の数の平均偏差及び標準偏差から構築される。

【０２１５】

曖昧さの統計的頻度を適用して住所の完成の可能性を定量化する場合、興味深い尺度は、ａ）Ｋ個の選択肢を有する入力の数のヒストグラム・プロット、ｂ）選択肢の頻度の分布のヒストグラム・プロットを伴う、選択肢の最も大きい範囲内のＮ個の入力のリスト、ｃ）単一の選択肢への最強の関連を伴うＮ個の入力のリスト、ｄ）選択肢の数の平均偏差及び標準偏差である。

【０２１６】

類似の尺度が、曖昧さの誤差模型尺度に当てはまる。

【０２１７】

変形入力のデータ品質問題（フィールドにおける入力が意図したものではない）は、無効値の問題及び不十分な情報の問題の両方に類似している。レベルによっては、ある入力が意図したものではないと言うことは、特定の（必ずしも枚角ではないけれども）検証基準に対して無効であるということを断言することである。通りの名前の妥当性は、当該通りの名前を参照用のポスタル・アドレス・ファイルに含まれているものと比較することによって定義することができる。あるいは、参照用データセットが無い場合、関連付けられた変形に整合する入力の相対頻度から妥当性を推察することができる。ある入力が相対的に高い頻度で起こり、高い頻度の選択肢が無い場合、それを妥当として解釈することができる。それが相対的に低い頻度で起こり、単一の高い頻度の選択肢が有る場合、それを無効として解釈することができる。複数の高い頻度の選択肢が有る場合、妥当なデータへの修正は、曖昧である場合があり、当該データが欠落していたらどうなっていたであろうかということと同様にして定量化することができる。

【０２１８】

フィールドが有効値の列挙集合の範囲外の値を有すると仮定する。例えば、性別フィールドが、Ｍ又はＦの代わりにＧである。曖昧さの均等分配尺度によれば、当該入力には２つの選択肢があったと言われる。頻度尺度もやはり２つの選択肢を示すけれども、Ｍ又はＦのいずれかへのバイアスを含む場合がある。選択肢の数に変動が無い、単純な列挙された場合においては、当該データセットについての曖昧さの尺度は、入力当たりの曖昧さの尺度掛ける無効値の割合の積から形成される。

【０２１９】

もし通りの名前にミススペルがあったらそうであろうように、選択肢の数に変動が有る場合、曖昧さの尺度が、当該データに存在する曖昧さを定量化する助けとなるであろう。頻度及び誤差模型尺度は、閉経のスペルについて最も信頼性のある結果を与えるであろう。不十分な情報と同様に、曖昧さの尺度は、究極的には、当該データセットをクレンジングによってどれくらい向上させることができるか、及び不確かさが未だどれくらい残るかを反映する。

【０２２０】

間違ったフィールドに入力されたデータもまた、同様の方法で定量化することができる。ここで、データ配置が実際に間違っているか否かに関する更なる曖昧さが存在する場合がある。一般的なファースト・ネームの値を有する名字の場合、ある順序付けが別の順序付けよりも正しいか否かがハッキリしない場合がある。一般的なファースト・ネームが名字フィールドにどれくらい頻繁に存在するか（又はその逆）を知ることは、間違って順序付けされた名前の問題が如何に深刻であるかを明確化するのに役立つ。正確な参照（又は他の検証手段）が入手可能である場合、エラー率の尺度を得ることができる。１００，０００個の入力のデータセットのうち、一般的なファースト・ネームである名字が５００個あるけれども、これらのうち２５個だけが実際に間違って順序付けされている。そのときのエラー率は、２５／５００＝１／２０である。測定されたエラー率が無い場合でも、上記問題に対して脆弱な入力が５００／１００，０００＝１／２００しかないことを知ることは、そのデータにおける信頼度を向上させる。

【０２２１】

名前フィールドについての他の興味深いチェックは、ａ）名及び姓（名字）が一般的なファースト・ネームである場合、ｂ）名が一般的なラスト・ネームであり、姓（名字）が一般的なファースト・ネームである場合、ｃ）名が一般的なラスト・ネームであり、ラスト・ネームも同様である場合、において知られるべきであろう。名前の参照用データセットから導かれる曖昧さの頻度尺度（ことによると、データセットそのもの）を使用して、順序付けが正しい確度を演算することができる。例えば、Ｄａｖｉｄ　Ｐａｕｌという名前について考察する。Ｄａｖｉｄがファースト・ネーム又はラスト・ネームである可能性（及びＰａｕｌについても同様に）から、Ｄａｖｉｄがファースト・ネームである確度を演算することができる。

【０２２２】

例えば、住所におけるフィールド間での並べ替えは、選ばれた参照用データセットと比べて、間違いである場合があるけれども、順序付けを指定する標準が弱い又は存在しないの何れかであるため、正確に言えば、間違いではない。ここで、参照用データセットと比較した特定の住所の順序付けの曖昧さの尺度を使用して、異なる順序付けは、データへの関連付けの曖昧さを導入しないので、深刻なデータ品質問題ではないことを示すことができる。このことは、データ品質を向上させるために、どこに労力を注ぎ込むべきかの決定に重要な情報である。

【０２２３】

フリー・テキスト・フィールドに存在する参照用データの量を定量化するために、当該フリー・テキストを、選ばれた参照用データセットに照らしたファジー検索において使用される単語に分解することができる。例えば、製品番号が請求書のコメント・フィールドに格納されていると会社が予想していると仮定する。当該コメント・フィールドにおける各単語を使用して、製品番号テーブルに照らしてファジー検索を行うことにより、当該コメント・フィールドに存在する製品番号の数を見出すことができる。より一般的には、各単語をＷＦＳファイル（他の特許を参照）に照らして検索して、その単語又はその変形が現れるフィールドの全てを、どれほどの配分率であるかと共に、特定することができる。これにより、あるフィールドにおけるデータの他のフィールドにおいて見られるデータに対するファジーな相互相関が与えられる。（このアプローチを使用して、間違ったフィールドに配置されたデータを識別することもできる。）

【０２２４】

多くの参照整合性問題は全て、定量化することができる。先ず、リンクされたデータが実は正しくリンクされていない厳密なキーのペアの割合として、確証尺度を定義することができる。この種の確証は、比較可能なフィールドがリンクの両側に存在する（又は、更なる結合を通してリンケージの点まで繋ぐことができる）場合にのみ行うことができる。概して、これは、データベース間に結合キーが確立されている２つ以上のデータベースに保持された（顧客の名前及び住所等の）類似のフィールド間での比較である。この状況は、データ・ウェアハウスに保持されている比較データが、データ・ウェアハウスを占める

ソース・システム中のデータと比較される場合に起こる。異なるソース・システムが相反する情報を保持している場合、又はそれらが一貫性無く更新されている場合、データ・ウェアハウスも同様に１つ以上のソース・システムと相容れない場合がある。データ・ウェアハウスをそのソースとの一貫性について妥当性を確認することにより、企業データ品質についての新しく重要なチェックが提供される。

【０２２５】

第２のチェックは、リンクの向こう側にレコードが無いリンクである、欠落リンク及び孤児リンクを探すことである。ファジー検索により、リンクされるべき別のデータが有る（欠落リンク）のか又は無い（孤児リンク）のかを特定することができる。これらの条件の各々の割合の尺度が重要である。リンクの向こう側のレコードへの整合が曖昧である場合、曖昧さの尺度は、それを定量化することができる。このことは、このことを一意的に行うことができるリンクを再投入し、どのリンクが曖昧であるが故に更なる調査を必要としているのかを識別する、データ・クレンジング操作の基礎を形成する。（以下、部分的メンバシップの可能性について考察する。）

【０２２６】

２つのデータベースが関係する情報を含むけれども、それらの間にキー関係が存在しない場合、データセット間でのファジーな検索又は結合により、これらのデータセット間に予想されるリンクが見付け出されるであろう。予想される各リンクの曖昧さの尺度は、これらのデータセットの間でのマッピングが如何にクリーンであるかを示すであろう。このことは、例えば、２つの会社が合併する際に、それらのマスタ参照用データ（例えば、顧客の名前及び住所）を組み合わせる場合に、非常に有用である。同様に、それを使用して、企業の異なる部分の参照用データをマージすることができる。このことは、マスタ・データ管理ソリューションの準備の初期段階で重要であろう。マスタ・データ管理ソリューションの創作におけるギャップ分析の部分は、異なるシステムによって使用される現存する参照用データセット間での調整の品質を決定する。システム間での初期調整は、この分析の副産物である。次いで、曖昧さの尺度を有することが、これらのシステムを調和させるのにかかる更なる労力を定量化するのに役立つ。

【０２２７】

（ｇ）クラスタ及び部分的メンバシップ

（１）ファジーなクラスタの創作

前に述べたように、如何にして要素を纏めてグループ化するかについての先験的助言が存在しない場合、原理又はアルゴリズムを使用してグループを識別する。実際の場合は、どの要素が他の要素を引きつけてグループを形成する中核的要素として役に立つかは明らかではないことが多いので、このことは実用的見地から重要である。例えば、顧客住所データベースにおける重複レコードの上記例において、どのレコードが真正のレコードであるかを会社が見分けるのは時として不可能である。以下の考察により、それらの中で幾つかのデータが互いの仲間であるとみなされる、レコードのファジーなクラスタを形成するのに使用することができる幾つかのアルゴリズムが提案される。

【０２２８】

上述の顧客住所の例において、「Ｊｏｈｎ　Ｓｍｉｔｈ」と関連付けられる６つのレコード：（Ａ）Ｊｏｈｎ　Ｓｍｉｔｈ　ｏｘ２　６ｑｔ；（Ｂ）Ｊｏｈｎ　Ｓｍｉｔｈ　ｏｘ２　６ｑｘ；（Ｃ）Ｊｏｈｎ　Ｓｍｉｔｈ　ｏｘ２　６ｑｙ；（Ｄ）Ｊｏｈｎ　Ｓｍｉｔｈ　ｏｘ２　７ｑｙ；（Ｅ）Ｊｏｈｎ　Ｓｍｉｔｈ　ｏｘ２６ｑｔ；（Ｆ）Ｊｏｈｎ　Ｓｍｉｔｈ　ｏｘ２６ｑｙ、が存在する。どのレコードが実在の所帯に該当するのかを知ること無く、上記レコードを、各クラスタが実在の所帯を表現する２つ又は３つのクラスタにグループ化することに会社が感心を抱く場合がある。この方法において、会社は、偽の郵送先住所に送られる郵便物を減らすことにより、郵便物の量を減らすことができる場合がある。

【０２２９】

クラスタを作り出すのに使用することができる１つのアルゴリズムは、指定された距離

内にある最大数の要素を含む最大の互いに素な部分集合を見付けることである。このアプローチについて、図５を参照しながら説明する。含まれるステップは、フロー・チャートに図解されている。図５もまた、上記顧客住所データベースの例を使用して、当該アルゴリズムを確証する。各ステップからの結果は、対応するステップの右側に示されている。この例において、指定された距離は２操作である。

【０２３０】

図５におけるフロー・チャートに言及して、最大の互いに素な部分集合からクラスタを作り出す最初のステップは、各要素について、その要素の変形をカウントすることである（５０２）。上記において定義したように、要素の変形は、その特定の要素から指定された距離内にある要素である。顧客住所の例においては、レコードＡについて、２操作の距離内にあるレコードが３つ（Ｂ、Ｃ、及びＥ）存在する。レコードＢについては、２操作の距離内にあるレコードが２つ（Ａ、Ｃ）存在する。Ｃについては４つ、Ｄについては１つ、Ｅについては２つ、そしてＦについては２つ存在する。

【０２３１】

次いで、最大数の変形を有する要素を選択し（５０４）、当該要素及びその変形を、当該要素によって表示されたグループにする。顧客住所データベースの例においては、レコードＣが最大数（４）の変形を有する。レコードＣ及びその変形（Ａ、Ｂ、Ｄ、Ｆ）が第１のクラスタを形成する。

【０２３２】

次に、全ての要素の集合から、最大のグループの要素を除去する（５０６）。顧客住所の例においては、これにより、レコードＥだけが残る。

【０２３３】

次いで、残っている要素において、最大数の変形を有する要素を見付ける（５０８）。このステップにより、第２のクラスタが生ずる。顧客住所の例においては、第２のクラスタは、その中に１つの要素Ｅしか有していない。

【０２３４】

全ての要素がクラスタにグループ化されるまで継続する（５１０）。顧客住所の例においては、全ての要素がそのグループを見付けてしまったので、更に続ける必要は無い。

【０２３５】

このアルゴリズムにより、顧客住所の例において、２つのクラスタ（Ａ、Ｂ、Ｃ、Ｄ、Ｆからなるグループ、及びＥのみからなるグループ）が生成する。当該会社は、各グループに含まれるレコードを互いの重複として扱うことができ、これらのレコードを統合して、郵便物の量を減らすことができる。

【０２３６】

上記アルゴリズムに若干の調節を加えてもよい。例えば、レコードＡ及びＦは両方とも、Ｃ及びＥから同じ距離にある。上記アルゴリズムにおいては、レコードＡ及びＦをＣの重複として割り付けることは人為的であり、レコードＡ及びＦがＥよりもＣに近いことを必ずしも示さない。

【０２３７】

１つの調節は、クラスタの表現についての不確かさを注記することである。例えば、Ｃ５－２という表現を使用して、レコードＣ及びその変形を含むＣのクラスタを、レコードの総数を示す５及び不確かさを示す－２で表現することができる。Ｅ　１＋２という表現を使用して、レコードＥ及びその変形を含むＥのクラスタを、そのグループにおける１及び不確かさを示す２で表現することができる。クラスタの正の不確かさは、他のところにグループ化されている要素が、このクラスタに属しているかもしれないことを反映している。クラスタの負の不確かさは、このクラスタにおける要素が別のグループに属しているかもしれないことを反映している。

【０２３８】

もう１つの調節は、Ａ及びＦをＥのクラスタに加えることであってもよい。従って、グループＣはレコードＡ、Ｂ、Ｄ、Ｆを有し、グループＥはレコードＡ及びＦを有する。し

かしながら、レコードＡ及びＦは２つのグループに属するため、全てのグループにおけるレコードの総数は８であり、レコードの真の総カウントよりも２つ多い。真のカウントを維持するために、部分的メンバシップを使用してもよい。

【０２３９】

ファジーなクラスタを構築する第２の方法は、会社名等の、複数の単語のフィールド上のデータをクラスタ化する際にふさわしく、この場合、単一の単語ではなく句（又は、英国の有文番号のようにフィールド全体）を採点することによってレコード間の変動が評価される。句の採点は、単語の変形スペルのみならず、単語の順序、欠落した単語、及び詳細な単語の連なりを変化させる単語間への挿入をも考慮に入れる。例えば、Ｂａｎｋ　ｏｆ　Ａｍｅｒｉｃａという会社名について、識別され、区別される必要がある名前の変動の４つのタイプを如何に説明する。

【０２４０】

【表３】

1) Bank of Amrica (fuzzy match of a word)

2) America Bank (word order, missing word)

3) Bank America (missing word)

4) Bank of South America (inserted word)

【０２４１】

句に基づくクラスタ化の例として、マスタ顧客リストにおける同じ法人に属する会計の全てを識別することを銀行が試みると仮定する。法人は、会社名、住所、及び（存在する場合は）会社登録番号によって識別されるであろう。会社名は法人と非常に相関性があり、常に入力されているので、クラスタ化に使用される主フィールドは会社名である。住所は、類似の名前を偶然有する会社を区別するのに使用される第２のフィールドである。会社登録番号は法人の識別において最も確実であることが期待されるけれども、単独で使用するのに十分なほど入力されてはいない。

【０２４２】

ファジーなクラスタ化操作は、元のデータセットを、クラスタ・メンバシップについての全ての要素の比較を可能とするのに適切な大きさの、より小さい部分集合に分割するスーパークラスタ・キーを識別することによって始まる。異なるスーパークラスタ・キーを有するレコードは、構築によって、異なるクラスタに入るであろう。住所等の地理的なデータについては、郵便番号が適切なスーパークラスタ・キーであることが多い。整合する変形郵便番号を有するレコードは、当該スーパークラスタに含まれる場合がある。整合しない郵便番号を有するレコードは、高い確率で別のクラスタに属すると予想され、それ故に、性能を向上させるために、スーパークラスタ・キーを導入することによってクラスタを演算する際に、それらは排除される。

【０２４３】

各スーパークラスタ内で、データは、会社名のフィールドの長さの降順且つ会社名の降順でソートされ、再現性のある順序で、最も長い名前を最初にクラスタ化アルゴリズムに提供する。スーパークラスタ・グループにおける最初のレコードは、最初のクラスタの主レコードとされる。その後の各レコード（ここでは、現レコードと呼ばれる）は、当該クラスタの主レコードの会社名に照らして現レコードの会社名を採点することによって、存在する各クラスタの主レコードと比較される。スコアが疑わしい整合閾値を超える場合、当該クラスタは、現レコードについての疑わしいクラスタのリストに加えられる。現レコードを存在する全ての主レコードと比較した後、疑わしいクラスタのリストが空である場合は、現レコードが、新しいクラスタの主レコードとされる。疑わしいリストが入力を１つだけ有しており、且つスコアが整合閾値を超えている場合は、現レコードは、当該疑わしいリスト上のクラスタに加えられる。疑わしいリストが２つ以上の入力を有している場合は、現レコード上の会社名が、当該疑わしいリスト上のクラスタの各々における全ての

レコードに照らして採点される。現レコードは、それが整合閾値を超える最高スコアを有するクラスタに加えられる。２つ以上のクラスタにおけるレコードについての最高スコアによる互角の整合が存在する場合、現レコードは、このようなクラスタの中の最初のクラスタに加えられる。整合閾値を超えるスコアが無い場合、現レコードは、新しいクラスタの主レコードとなる。

【０２４４】

このアルゴリズムは、２つの重要な特徴を有する。複数のクラスタへの曖昧な整合は最初に整合したクラスタの方に決定されるので、クラスタの中には、曖昧なメンバで比較的過剰に入力されているものもある。また、レコードがアルゴリズムに提供される順序は、特定のメンバシップ決定に影響を及ぼす。会社名の長さ及び値についての初期ソートは、名前の固定された順序を確立することによって、この問題を改善することを目的とする。下記において考察する部分的メンバシップの観念は、クラスタ・メンバシップの曖昧さをより正確に反映する、より高品位なソリューションを与える。

【０２４５】

曖昧なメンバシップの例を、以下の会社名の集合に示す。

【０２４６】

【表４】

ACME Services Australia Limited

ACME Services Canada Limited

ACME Services Limited

【０２４７】

特定の採点において、ＡＣＭＥ　Ｓｅｒｖｉｃｅｓ　Ａｕｓｔｒａｌｉａ　ＬｉｍｉｔｅｄのＡＣＭＥ　Ｓｅｒｖｉｃｅｓ　Ｃａｎａｄａ　Ｌｉｍｉｔｅｄに対するスコアは、整合閾値０．７５未満である０．６５であり、これらのレコードは別個のクラスタに配置される。ＡＣＭＥ　Ｓｅｒｖｉｃｅｓ　Ｌｉｍｉｔｅｄは、両方のクラスタに対して等しいスコア０．９５を有する。それは、ＡＣＭＥ　Ｓｅｒｖｉｃｅｓ　Ａｕｓｔｒａｌｉａ　Ｌｉｍｉｔｅｄと最初に出くわすので、ＡＣＭＥ　Ｓｅｒｖｉｃｅｓ　Ａｕｓｔｒａｌｉａ　Ｌｉｍｉｔｅｄのクラスタのメンバとなる。

【０２４８】

（２）部分的メンバシップ

前のセクションにおける最初の例において、レコードＡ及びＦは、クラスタＣ及びＥの両方に属する。クラスタにおけるレコードの全ての出現が１とカウントされる場合、６つのレコードしか存在しないにもかかわらず、クラスタＣ及びＥにおけるレコードの総カウントは８である（グループＣ（Ｃ、Ａ、Ｂ、Ｄ、Ｆ）に５、及びグループＥ（Ｅ、Ａ、Ｆ）に３）。この場合、部分的メンバシップを使用して、総カウントを維持することができる。１つのデータが２つ以上のグループに属する場合、そのデータの出現は１未満、即ち分数としてカウントされる。しかしながら、そのデータの全ての出現の合計は、やはり１となって、総カウントを保つべきである。

【０２４９】

アレンジによっては、例えば、上述の曖昧さの尺度を使用して、ある要素が特定のグループに属する確度を反映するように、そのグループにおける当該要素の部分的メンバシップを定義してもよい。

【０２５０】

例えば、レコードＡが４０％の確率でグループＣに属し、６０％の確率でグループＥに属すると仮定する。グループＣにおけるレコードＡには０．４の部分的メンバシップを割り付けることができ、グループＥにおけるレコードＡには０．６の部分的メンバシップを割り付けることができる。

【０２５１】

同様に、レコードＦが１０％の確率でグループＣに属し、９０％の確率でグループＥに

属すると仮定する。グループＣにおけるレコードＦには０．１の部分的メンバシップを割り付けることができ、グループＥにおけるレコードＦには０．９の部分的メンバシップを割り付けることができる。

【０２５２】

レコードＡ及びＦに割り付けられた部分的メンバシップにより、総カウントは、グループＣのカウント（１＋１＋１＋０．１＋０．４＝３．５）とグループＥのカウント（１＋０．９＋０．６＝２．５）との合計（６となる）である。故に、総カウントは維持される。

【０２５３】

部分的メンバシップの発端は、特定の要素のメンバシップに関する不確かさであり、各グループの総メンバシップは不確かさの度合い（即ち、誤差の範囲）によってのみ知られる。各グループの総カウントは、誤差の範囲によって調整される、全体の及び部分的な、メンバシップの合計として表現することができる。この範囲は、メンバシップについての全ての不確かな決定が包括又は排他の何れかに分類されるとみなすことによって得られる最大及び最小の境界値によって示すことができる。これらは、クラスタ間でのメンバの分布についての最悪の事態のシナリオに相当する。ここで、境界値について、Ｃにおける総メンバシップは３．５（３，５）であろう。これは、Ｃのメンバの予想数は３．５であるけれども、Ｃは最少で３つのメンバ、最大で５つのメンバを有することを言うものと解釈される。同様に、Ｅにおける総メンバシップは２．５（１，３）であろう。Ｅの予想される総メンバシップは２．５であるけれども、Ｅは最少で１つのメンバ、最大で３つのメンバを有する。

【０２５４】

異なるクラスタに属する境界値は相関性があるけれども、ここで使用される観念はそれを示唆しない。データセットにおける異なるレコード間の相関関係は可能であり、境界値を演算する際に考慮に入れられるべきである。例えば、時として、Ａ及びＦが、どのクラスタに属するかを知ること無く、それらが同じクラスタに無いことを知ることは可能である。

【０２５５】

上記第２の例において、特に、２つ以上のクラスタに対する曖昧な整合又は疑わしい整合が存在する場合に、上記疑わしい閾値を超える主レコードと整合する全てのレコードを各クラスタに関連付けることによって、クラスタ化の品質を高めることができる。当該整合の品質は、疑わしい各レコードに照らして記録され、曖昧さの尺度によって定量化されるべきである。部分的メンバシップは、あるアレンジにおいて、完全な（ｆｕｌｌ）メンバとは別個に保持され、それらの部分的メンバシップの尺度によって表示されるであろう。例えば、あるクラスタのメンバは、（部分的メンバシップを有する完全なメンバと共に）部分的メンバシップの降順で列挙することができる。

【０２５６】

部分的メンバシップが一緒に特定されたレコードをリンクし且つ部分的メンバシップの割り当てを特定したルール、事象又は決定を識別するために使用することができるルール表示を各レコードにも添付すべきである。このルール・ラベルは、レコードを異なる部分的メンバシップと組み合わせる際に部分的メンバシップを調整する場合に有用であろう。

【０２５７】

ある見地からは、部分的メンバシップは、メンバシップの不確かさから生ずる曖昧さを反映するけれども、別の見地からは、部分的メンバシップは、２つの部門のために働いている従業員の例におけるように、単純に、複数のクラスタ間でのメンバシップの割り当てである。不確かな場合、既知知識における変化が、メンバシップの割り当てを変化させることが予想される。あるいは、部分的メンバシップを、単純に、信頼性のあるものとして許容してもよい。上記比率を現実のものとして許容するのに全くコストは掛からない。

【０２５８】

部分的メンバシップが、異なるクラスタに属する要素の確度を表現する場合、部分的メ

ンバシップは常に負でない値であり、異なるクラスタに属する要素の部分的メンバシップの合計は１であろう。

【０２５９】

しかしながら、部分的メンバシップは、アレンジによっては、負の値であってもよい。しかしながら、異なるクラスタに属するオブジェクトの部分的メンバシップの合計は、やはり１であるように制約されなければならない。

【０２６０】

アレンジによっては、要素の部分的メンバシップは、当該要素と上記主レコードとの間の距離の関数として、又は当該要素と上記主レコードとの間の整合スコアの関数として、定義することができる。ファジーなスコアから部分的メンバシップを構築する１つの方法は、上述のように、曖昧さの尺度によるものである。異なるファジーなスコアは、要素と上記主レコードとの間の異なる距離を反映し、故に、曖昧さの異なる尺度を反映する。ファジーなスコアは変形と上記種レコードとの間の類似性を反映し、確率と同じではないことが多いということに留意されたい。

【０２６１】

（ｈ）ファジーなデータ操作

（１）部分的メンバシップの存在下でのフィルタリング

選択基準を適用して、共通の性質を共有するレコードの部分集合を分離することが有用であることが多い。例えば、国際的なレコードのデータセットにおいて、特定の国からのレコードを選択することができる。当該選択操作（時として、「フィルタリング」と称される）は、選択を特定する表現において使用されるフィールドがキーである必要は無いので、キーに基づくものとはみなされない。レコードが複数のクラスタにおいて部分的メンバシップを有することが許される場合、フィルタリングは、部分的メンバの幾つかを失わせる場合がある。その結果は、１未満であるかもしれない選択された部分集合に亘るレコードに関連付けられた総メンバシップ割り当てである。これについての説明は、総割り当ては、選ばれた部分集合におけるメンバシップを、当該選ばれた部分集合の外にある選択肢に照らして測定するということである。

【０２６２】

ＡＣＭＥ　Ｓｅｒｖｉｃｅｓ　Ｌｉｍｉｔｅｄは、ＡＣＭＥ　Ｓｅｒｖｉｃｅｓ　Ａｕｓｔｒａｌｉａ　Ｌｉｍｉｔｅｄを含むグループに対して０．５の割り当てと、ＡＣＭＥ　Ｓｅｒｖｉｃｅｓ　Ｃａｎａｄａ　Ｌｉｍｉｔｅｄを含むグループに対して０．５の割り当てとを有する。データセット全体に亘るＡＣＭＥ　Ｓｅｒｖｉｃｅｓ　Ｌｉｍｉｔｅｄについての総割り当ては１．０である。フィルタを適用してＣａｎａｄａと関連するレコードのみを保持する場合、ＡＣＭＥ　Ｓｅｒｖｉｃｅｓ　Ｌｉｍｉｔｅｄは、結果として得られるデータセットにおいて、０．５の総割り当てを有するであろう。このことは、ＡＣＭＥ　Ｓｅｒｖｉｃｅｓ　Ｌｉｍｉｔｅｄは、Ｃａｎａｄａ部分集合中に無い選択肢については５０％の可能性を有するのに対して、Ｃａｎａｄａ部分集合中にある選択肢については５０％の可能性を有することを示している。

【０２６３】

（２）キーによる並列処理及び部分的メンバシップ

並列処理において、キーの値に基づいて、異なる処理パーティションにレコードを割り当てることができる（時として、「キーによる分割」と称される）。レコードが曖昧なクラスタ・メンバシップを有することが許される場合、各クラスタに関連付けられたキーに基づいて分割を行ってもよい。この分割スキーム下では、所定のレコードに関連付けられたパーティション内での総割り当てが１未満であってもよい。このことの解釈は、フィルタリングについてのそれと似ている。それは、そのパーティションには無い選択肢に照らして、そのパーティションへのレコードの割り付けを測定する。

【０２６４】

ＡＣＭＥ　Ｓｅｒｖｉｃｅｓ　Ｌｉｍｉｔｅｄは、ＡＣＭＥ　Ｓｅｒｖｉｃｅｓ　Ａｕｓｔｒａｌｉａ　Ｌｉｍｉｔｅｄを含むグループに対して０．５の割り当てと、ＡＣＭＥ

Ｓｅｒｖｉｃｅｓ Ｃａｎａｄａ Ｌｉｍｉｔｅｄを含むグループに対して０．５の割り当てとを有すると仮定する。キーによる分割操作は、ＡＣＭＥ Ｓｅｒｖｉｃｅｓ Ａｕｓｔｒａｌｉａ Ｌｉｍｉｔｅｄを含むグループを１つのパーティションに、ＡＣＭＥ Ｓｅｒｖｉｃｅｓ Ｃａｎａｄａ Ｌｉｍｉｔｅｄを含むグループをもう１つのパーティションに割り当てることができる。ＡＣＭＥ Ｓｅｒｖｉｃｅｓ Ｌｉｍｉｔｅｄのレコードに関連付けられた後者のパーティションにおける総割り当ては、何等かの他のパーティション（複数であってもよい）それがある０．５の選択肢に対して、０．５である（ＡＣＭＥ Ｓｅｒｖｉｃｅｓ Ｃａｎａｄａ Ｌｉｍｉｔｅｄのクラスタとのその関連性を反映している）。

【０２６５】

よく知られているデータ操作の並列バージョンは、パーティション間でのやりとりが全く無い、単一のパーティション内でのそれらの挙動によって定義することができる。あるレコードについてのパーティション内での総割り当てが１未満である場合、このことは、ここで定義された意味に解釈される。

【０２６６】

（３）ロールアップ及び部分的メンバシップ

ロールアップ操作は、個々のレコードのレベルからグループのレベルへとデータを統合又は要約する。厳密なキーの場合、キーグループは、共通キー（値）を共有するレコードの集合として定義される。クラスタの場合、グループは、１つ以上のレコードが１つ以上のグループのメンバである可能性を伴う、メンバが比較によって特定されるレコードの集合として定義される。

【０２６７】

クラスタ・グループにおける付加的な（また、例えば、対数を加えることによる、乗法的な）数学的統合（時として、「演算加法的測度（ｃｏｍｐｕｔｉｎｇ ａｄｄｉｔｉｖｅ ｍｅａｓｕｒｅｓ）」と呼ばれる）は、重み付けについての割り当て尺度を使用する加重統合として行われる。境界値は、（集合への部分的割り付けを有する全てのレコードが当該集合に含まれているか又は当該集合から排除されているかの何れかである）選択肢についての（重み付けをしない）統合を演算することによって演算される。以下のレコードのリストは、会社名に基づくクラスタ・グループである。

【０２６８】

【表５】

ｸﾗｽﾀ･ｷｰ	割り当て尺度	会社名	ｶｳﾝﾄ
c1	1.0	ACME Services Australia Limited	80
c1	1.0	ACME Services (AUS) Limited	60
c1	0.5	ACME Services Limited	100

【０２６９】

クラスタにおける総カウントを特定するロールアップは、重み付き合計が８０＊１．０＋６０＊１．０＋１００＊０．５＝１９０であり、境界値が８０＊１．０＋６０＊１．０＋１００＊０．０＝１４０（排他的）及び８０＊１．０＋６０＊１．０＋１００＊１．０＝２４０（包括的）である。クラスタ・グループにおける総カウントについての結果は、１９０（１４０，２４０）と表現することができる。

【０２７０】

部分的割り付けを有するレコードが集合に含まれるか又は集合から排除されるかの何れかである極端な場合を考察することによって、非付加的な要約を行う。割り付け尺度を使用して、部分的メンバを含めることによって得られる結果に信頼度を割り付けることができることが多い。例として、何等かの二次的なキーに基づいて、クラスタ・グループ内でレコードをソートしてもよく、そのソート順序において、どのレコードが最初であるのかをロールアップによって特定してもよい。以下のリストは、レコードの前記リストをカウ

ントの降順でソートしたものである。
【０２７１】
【表６】

ｸﾗｽﾀ･ｷｰ	割り当て尺度	会社名	ｶｳﾝﾄ
c1	0.5	ACME Services Limited	100
c1	1.0	ACME Services Australia Limited	80
c1	1.0	ACME Services (AUS) Limited	60

【０２７２】
このソート順序でのクラスタ・グループにおける最初のレコード（即ち、最大カウント）を特定するロールアップにより、以下の境界の結果が与えられる。
【０２７３】
【表７】

（包括的）c1 0.5 ACME Services Limited 100

（排他的）c1 1.0 ACME Services Australia Limited 80

【０２７４】
０．５の割り当て尺度を包括的な結果に規定して、包括的結果に関連付けられる信頼度を示すことができる。この例においては、排他的結果は最悪の事態の結果とみなすことができる（最大値が少なくとも８０である）。
【０２７５】
部分的メンバシップの存在下でのロールアップ操作は、並列して実行してもよい。このことを理解するために、先ず付加的な数学的統合について考察する。これは加重和である。このような合計は、別個に演算される部分和に分け、次いで組み合わせることができる。各部分和は、その独自の並列パーティションにおいて演算することができる加重和である。これは、付加的ロールアップの並列化である。
【０２７６】
定義によって全ての部分的メンバが排除されるので、付加的ロールアップ及び非付加的ロールアップの両方の排他的境界の演算は並列化可能である。従って、当該演算により、（殆どの場合において）並列化可能である通常のロールアップが減少する。
【０２７７】
包括的境界の演算は、包括の二重カウントを防ぐ特定の条件下で並列化可能である。クラスタ内でのロールアップについて、各部分的メンバは当該クラスタ内で１回だけ発生する。故に、包括的境界は、何れのメンバも二重にカウントすること無くクラスタ内での部分和の合計として演算することができる。
【０２７８】
ロールアップをクラスタに亘って実行する場合、異なるクラスタにおいて出現する同じレコードからの寄与があるかもしれない。このことは、当該レコードの各インスタンスと関連付けられた重み付けが加わり、新しい全体としての重み付けが与えられるので、問題無い。しかしながら、包括的境界については、各レコードは１回だけ含まれるようにしなければならない。一般に、このことは、レコードが発生する経過を何等かの方法で追うことを必要とし、この操作は並列化可能ではない。
【０２７９】
しかしながら、個々のレコードが、クラスタ化に先立って、キー（例えば、ｒｅｃ＿ｋｅｙ）によって識別され、（クラスタ化後の）データがｒｅｃ＿ｋｅｙに基づいて並列分割される場合、同じｒｅｃ＿ｋｅｙを有する全てのレコードは同じパーティションにおいて発生するであろう。このパーティション内でのロールアップは、たとえクラスタに亘って行われても、全ての関係のあるレコードが存在するので、包括的境界を正しく演算することができる。個々のレコードにおいて、２つ以上のパーティション上にインスタンスを

有するものは無く、従って、二重カウントの可能性は無いので、次いで、パーティションを亘る包括的境界を、パーティションを亘って安全に組み合わせることができる。

【０２８０】

（４）検索

厳密なデータ操作においては、厳密なキーが使用される。例えば、検索操作において、キーが使用され、そのキーに厳密に整合する全てのレコードが取り出される。ファジーなデータ操作においては、ファジーなキーが使用される。

【０２８１】

アレンジによっては、例としての検索操作を使用して以下に説明するように、一連の厳密なデータ操作として、ファジーな操作が実行される。

【０２８２】

上記顧客住所データベースの例において、会社は、Ｊｏｈｎ　Ｓｍｉｔｈと名付けられた人物に送られた全ての保留中の郵便小包を見付け出すことに興味を覚える。その目的に、検索操作を使用することができる。検索操作は、二値キー「Ｊｏｈｎ　Ｓｍｉｔｈ；ｏｘ２　６ｑｔ」を使用して実行することができる。ここで、郵便番号「ｏｘ２　６ｑｔ」は、Ｊｏｈｎ　Ｓｍｉｔｈと関連付けられた正しい郵便番号である。しかしながら、この厳密な検索操作は、入力を行う際に郵便局員によってなされた誤植のために、Ｊｏｈｎ　Ｓｍｉｔｈに送られた保留中の郵便小包を取り出さず、「Ｊｏｈｎ　Ｓｍｉｔｈ；ｏｘ２　６ｑｘ」又は「Ｊｏｈｎ　Ｓｍｉｔｈ；ｏｘ２６ｑｔ」というキーと誤って関連付けられた保留中の郵便小包を取り出すであろう。

【０２８３】

この制約を克服するために、ファジーなキーと関連付けられる検索である、ファジー検索を使用することができる。ファジーなキーは、主キー及びそのキーの指定された距離内に収まる全ての変形を含むキーの集合の１つである。上記顧客住所データベースにおいては、上記保留中の郵便小包の検索において使用されるファジーなキーを、主キーである「Ｊｏｈｎ　Ｓｍｉｔｈ；ｏｘ２　６ｑｔ」に加えて、２操作の距離内に収まる全ての変形を含むように定義することができる。

【０２８４】

アレンジによっては、ファジーなキー（主キーであるＪｏｈｎ　Ｓｍｉｔｈ　ｏｘ２　６ｑｔ及び４つの変形であるｏｘ２　６ｑｘ、ｏｘ２　６ｑｙ、ｏｘ２６ｑｙ及びｏｘ２６ｑｔ）に基づいて実行される上記ファジー検索操作は、以下のやり方で実行することができる。ステップ１において、主キーである「Ｊｏｈｎ　Ｓｍｉｔｈ；ｏｘ２　６ｑｔ」に基づく厳密な検索を実行する。次いで、ステップ２から５において、上記ファジーなキーの一部である変形に基づいて厳密な検索を４回実行する。最後のステップ（ステップ６）において、上記ステップ１から５から取り出された結果を組み合わせる。当該組み合わされた結果は、ファジーなキーを使用するファジー検索の結果である。

【０２８５】

会社の名前又は住所等の多単語フィールドにおいて検索を行う場合、前の手順に先だって、そのフィールドの変形の集合を特定し、直接的に使用することが可能ではない場合がある。２つの代替戦略を用いることができる。２　Ｐｌａｔｅｒ　Ｄｒｉｖｅ，　ＯｘｆｏｒｄのＡＣＭＥ　Ｓｅｒｖｉｃｅ　Ｌｔｄと関連付けられる全ての会計レコードについての検索がなされていると仮定する。第１の戦略においては、単一単語フィールド（例えば、郵便番号ｏｘ２　６ｑｔ）を検索キーとして使用する。当該キー「ｏｘ２　６ｑｔ」又は固定された距離以内のその変形の１つを有する全てのレコードが、「見込みのある整合」又は「有力候補」として、厳密な検索によって取り出される。各有力候補の会社の名前及び住所は、問合せ用の会社の名前及び住所に照らして、別個に採点される。

【０２８６】

概して、住所は、最初に、当該住所を構成する全てのフィールドを単一の文字列に連結することによって比較される。これには幾つかの利点がある。先ず、それは、異なるソースからの住所を単一の共通フォーマットに整える。句採点関数は単語の欠落や単語の順序

の変化に寛容であるので、ソースとなる住所の要素を保持している元のフィールドが不十分に又は一貫性無く埋められていても問題無い。同様に、住所を構文解析して標準的な住所の要素とする必要無しに、共通フォーマットに達する。構文解析は、概して、ポスタル・アドレス・ファイル等の正確な参照用データを必要とするけれども、特に外国においては、ポスタル・アドレス・ファイルは入手可能ではない場合がある。また、構文解析は、計算的には、相対的に高価であるので、それを回避することが、住所の比較の実行をより良好にする。連結された住所フィールドの比較が要領を得ない場合には、採点を改良するためのバックアップ的な試みとして、構文解析を使用してもよい。

【０２８７】

結果として得られた会社の名前及び住所のスコアは、（ユーザ指定の）整合閾値及び疑わしい閾値と比較される。整合閾値を超えるスコアは、比較された２つの句が整合を構成するのに十分に類似していることを示す。疑わしい閾値を超えるスコアは、それらの句が類似しているけれども、信頼度を以て整合を特定するのに十分な程には近くないことを示す。レコード全体が整合であるか否かについてのスコアは、個々のフィールド又はフィールドの組み合わせ（即ち、連結された住所）についてのスコアを組み合わせることによって得られる。どの情報がどれほど近く一致していなければならないかを特定する基準は、ユーザが指定する。例えば、市のフィールドは、郵便番号が一致する場合は、一致しなくても許される場合がある。（英国においては、例えば、何が有効な住所を構成するのかにおいて、驚くべき寛容度がある。町の名前及び番地の有無の両方についての変動が許容される。）

【０２８８】

第２の戦略は、１つ以上のフィールドから単語を選択し、これらをファジー検索用キーワードとして使用することである。このような単語は、フィールドにおける主要な単語として、又はそれらの有意性に基づいて、選ぶことができる。有意性は、フィールドにおける単語又はその変形の１つの出現数のそのフィールド（複数であってもよい）が埋められた（入力された）回数に対する比率のネガティブ・ログから演算される。

【０２８９】

ファジー検索用キーワードが選ばれると、各単語（及びその変形）がソース・レコードにおいて単語毎に調べられる。これは、選ばれた単語を所定のフィールドに含むソース・レコードのインデックスのリストを返す。検索用キーワードが、ポスタル・アドレス・ファイルの何れかの住所フィールドにおいて探される「Ｌｏｗｅｒ」、「Ｉｓｌｉｐ」であると仮定する。「Ｌｏｗｅｒ」という単語は、組織、通りの名前及び町のフィールドにおいて出現するかもしれない。「Ｉｓｌｉｐ」は町としてのみ出現するかもしれない。各単語についての検索により、上記単語が見付かるレコードのインデックスのリストが与えられる。これら２つのリストの積集合を得る（ｉｎｔｅｒｓｅｃｔ）ことにより、両方の単語（又はそれらの変形）を含むレコードの集合が与えられる。これらは有力候補であるレコードの集合を形成する。有力候補であるレコードは特定の数の検索用の単語を含むことが知られているので、それらは効率的に事前承認される。共通して保持されることが要求される検索用の単語が多いほど、フィールド間のスコアがより高くなると見込まれる。

【０２９０】

例えば、２つ以上の検索用単語を含むレコードのみを保つためにフィルタを適用した後、全部のレコードが取り出され、互いに照らして採点される。結果として得られるスコアは、降順にソートされる。曖昧さの尺度が整合の集合について演算され、これにより、参照に対する問合せの整合の品質を特徴付ける情報が増える。

【０２９１】

（５）ファジーな結合

レコードを取り出すのにルックアップを使用する代わりに、参照用データセット全体が読み込まれ、問合せデータセットに照らして処理される結合が行われることを除き、ファジーな結合はファジー検索に類似している。このことは、性能及び制御の両方について有用である場合がある。参照用データセットが、通常のルックアップとしてメモリに納める

には大き過ぎる場合、代わりに、ディスク上に保持されている（ことによると、ブロック圧縮された）ロード可能なルックアップとしてアクセスしてもよい（Ａｂ　Ｉｎｉｔｉｏのロード可能なルックアップを参照）。各検索語が処理される際に、ルックアップ・テーブルの適切なページがディスクからアクセスされる。参照用データセットの十分に大きい断片（例えば、１０％）にアクセスする必要がある場合は、ランダム－アクセス検索を始めるのではなく、単一のパスにおいてソートされた順序で参照用データセット全体を読み込む方が、結局、より効率的となる。これにより、検索プロセスによって繰り返し実行されるディスク・アクセスが減少する。

【０２９２】

制御の見地からは、検索が複数の参照用データセットへのアクセスを必要とする場合、結合が、これらの参照用データセットからのデータを採点のために一緒にするのに、より便利な方法である。問合せが単一の入力における顧客の名前及び住所を問合せが含むけれども、参照用データは顧客の名前及び住所を別個のテーブルに保持していると仮定する。顧客の名前と住所とを繋ぐ第３のリンク・テーブルが存在する場合がある。対象となる参照用データが異なるテーブルに保持されている場合、インデックスは異なるデータセットにおけるレコードを参照するので、検索用インデックスについてインデックス・ファイルを検索することによって返されるレコード・インデックスを直接的に比較することは可能ではない。問合せにキーが割り付けられている場合は、各データセットに照らして別個の検索を行うことができ、これらの検索の結果を、問合せキーに基づく結合によって組み合わせて、有力候補を取ってきて、採点することが可能となる。

【０２９３】

曖昧さの尺度がファジーな結合の出力の一部として演算される場合、各々が部分的メンバシップを有する複数の整合を結合操作の結果として生じさせることが可能である。例えば、番地が無い住所をポスタル・アドレス・ファイルに照らして結合して、郵便の宛先のファイルがその住所に記録されている組織の名前を取得すると仮定する。ポスタル・アドレス・ファイルにおいて適切な通りにある３つの住所には組織が無く、故に当該組織を識別するために、これらの住所を組み合わせることができる。他の整合する住所には、２つの別個の組織であるＡＣＭＥ　Ｌｔｄ及びＳｔａｎｄａｒｄ　Ｃｏｒｐ．が存在する。曖昧さの均等分配尺度により、同等な存在の数がカウントされる。当該結合の出力は、最初は、整合する住所の各々に１つずつ、５つのレコードであろう。続いて、（曖昧さの尺度を特定するために）当該住所における組織に対するロールアップを行うことにより、当該組織が、空欄（３／５）、ＡＣＭＥ　Ｌｔｄ（１／５）、Ｓｔａｎｄａｒｄ　Ｃｏｒｐ．（１／５）のいずれかである確度を示す曖昧さの均等分配尺度が得られる。次いで、この結果を、各々が別個の組織及び関連付けられた部分的メンバシップを伴う、３つのレコードに正規化する。

【０２９４】

【表８】

問合せ:住所に基づく結合により組織を取得する
問合せ住所: Lower Street, ox2 6qt

組織	住所
--	2 Lower Street, ox2 6qt
--	3 Lower Street, ox2 6qt
ACME Ltd	4 Loower St, ox2 6qt
--	5 Lower St, ox2 6qy
Standard Corp.	8 Lower St., ox2 6qt

結果:

組織	住所	割り当てキー	部分的ﾒﾝﾊﾞｼｯﾌﾟ
--	Lower Street, ox2 6qt	a1	0.6
ACME Ltd	Lower Street, ox2 6qt	a1	0.2
Standard Corp.	Lower Street, ox2 6qt	a1	0.2

【０２９５】

（６）ソート及び部分的メンバシップ

メンバシップが部分的である場合にレコードを順序付ける（ファジーな）ソートは定義が容易である。曖昧な関連性を有するレコードについては、その割り当て（曖昧さの尺度）と一緒に、各選択肢についてのレコードが作り出される。前の例における参照用レコードが、問合せレコードに照らして、結合の結果でソートされると仮定する。ルールは、部分的メンバシップは、完全なメンバシップ後に、メンバシップの降順でソートされるというものである。他のフィールドについての下位ソート（ｓｕｂｓｏｒｔ）は、部分的メンバシップが適用された後に適用される。従って、１つのフィールドに対する部分的メンバシップは、後のキーに対して上位にある。これにより、追加のソート・フィールドの適用は、より高いレベルにある順序を変更すること無く、確率された順序でレコードをソートする。

【０２９６】

【表９】

ソート{組織部分住所}:

--	2 Lower Street, ox2 6qt	--	1.0
--	3 Lower Street, ox2 6qt	--	1.0
--	5 Lower St, ox2 6qy	--	1.0
--	Lower Street, ox2 6qt	a1	0.6
ACME Ltd	4 Loower St, ox2 6qt		
ACME Ltd	Lower Street, ox2 6qt	a1	0.2
Standard Corp.	8 Lower St., ox2 6qt		
Standard Corp.	Lower Street, ox2 6qt	a1	0.2

【０２９７】

この種のファジーなソートを適用した場合、組織への（ソートされた）ファジーなロールアップは、全てのデータが見られるまで一時的な結果を格納することを必要とせずに、組織名に基づいて行うことができる。これは、ソートされたデータの主な用途の１つであ

り、ロールアップ操作は、各キー・グループが完成する際に完了する。

【０２９８】

ファジーなマージ操作は、ファジーなソートに似ている。それは、単に、その入力上の各レコードに順序付けルールを適用して、そのソート順助において、どのレコードが次にあるのかを特定する。上記データセットが以下のデータセットとマージされると仮定する。

【０２９９】

【表１０】

ACME Ltd	Lower Street, ox2 6qt	a2	0.9
Standard Corp.	Lower Street , ox2 6qt	a2	0.1

【０３００】

マージされたデータは以下の通りである。

【０３０１】

【表１１】

マージ{組織部分住所}:

--	2 Lower Street, ox2 6qt	--	1.0
--	3 Lower Street, ox2 6qt	--	1.0
--	5 Lower St, ox2 6qy	--	1.0
--	Lower Street, ox2 6qt	a1	0.6
ACME Ltd	4 Lower St, ox2 6qt		
ACME Ltd	Lower Street, ox2 6qt	a2	0.9
ACME Ltd	Lower Street, ox2 6qt	a1	0.2
Standard Corp.	8 Lower St., ox2 6qt		
Standard Corp.	Lower Street, ox2 6qt	a1	0.2
Standard Corp.	Lower Street , ox2 6qt	a2	0.1

【０３０２】

（ｈ）ファジーなデータ操作の有用性：間違い及び不確かさに取り組み、会計の完全性を保つ。

【０３０３】

上記ファジーな検索の例において示したように、ファジーな検索操作により、厳密なキーを使用する従来の検索では見逃すであろうレコード（例えば、誤植を含むレコード）を取り出すことができる。

【０３０４】

また、上述のように、データの分類が保留中の結果に依存する場合、クラスタ化又は部分的メンバシップを使用して、不確かさを正確に捉えることができる。直感的には、クラスタ化又は部分的メンバシップは、別個のファジーではない操作の組み合わせ又は連なりと同等であると見ることができる。しかしながら、クラスタ化又は部分的メンバシップは、より良い取り扱い及び予測を可能とするであろう。組織の年間予算が、当該組織が非慈善団体であるか否かに関する判決に依存している上記の例において、有利又は不利な判決が出される確率に基づいて年間予算を準備することができる。

【０３０５】

より具体的には、年間予算は以下のように予定しておくことができる。

【０３０６】

【数１４】

予定される納税額＝

非慈善団体の地位に基づく納税額×
不利な判決の確率
＋
慈善団体の地位に基づく納税額×
有利な判決の確率

【０３０７】
上記の式を使用して納税予想額を計算することは、組織の財政状態のより良好な全体像を提供し、経営幹部によるリスク評価を容易にする。それはまた、より信頼性の高い数字を下流のアプリケーションに与え、例えば、組織の財務見通しをより良好に予想することを可能とする。

【０３０８】
部分的メンバシップはまた、会計の完全性を保つのにも有用である。例えば、ＡＢＣ社の人材データベースにおいて、マーケティング部門及びＲ＆Ｄ部門におけるＪｏｈｎ　Ｓｍｉｔｈのメンバシップがそれぞれ０．５である場合、彼の医療費が二重にカウントされるようなことはないということが判る。

【０３０９】
上述のアプローチは、コンピュータ上で実行されるソフトウェアを使用して実装することができる。例えば、かかるソフトウェアは、少なくとも１つのプロセッサ、少なくとも１つのデータ記憶システム（揮発性及び不揮発性メモリ、並びに／又は記憶素子）、少なくとも１つの入力装置又は入力ポート、及び少なくとも１つの出力装置又は出力ポートを各々が含む、１つ以上のプログラムされたまたはプログラム可能なコンピュータ・システム（分散型、クライアント／サーバ型、又はグリッド型等の種々のアーキテクチャのものであってよい）上で実行される、１つ以上のコンピュータ・プログラムにおけるプロシージャを形成する。上記ソフトウェアは、例えば、演算グラフの設計及び構成に関係する他のサービスを提供する、より大きいプログラムの１つ以上のモジュールを形成していてもよい。上記グラフのノード及び要素は、コンピュータ可読媒体に格納されたデータ構造、又はデータ・リポジトリに格納されたデータ・モデルに合致する他の系統的データとして、実装することができる。

【０３１０】
上記ソフトウェアは、汎用の又は特殊用途のプログラム可能なコンピュータによって読み取ることができる記憶媒体（例えば、ＣＤ－ＲＯＭ）上に提供したり、又は実行時に、ネットワークの通信媒体越しに（伝搬信号中にコード化して）コンピュータに配信したりすることができる。上記機能の全ては、特殊用途のコンピュータ上で、又はコプロセッサ等の特殊用途のハードウェアを使用して、行うことができる。上記ソフトウェアは、当該ソフトウェアによって規定される計算の異なる部分が異なるコンピュータによって行われる分散方式にて実装することもできる。このような各コンピュータ・プログラムは、汎用の又は特殊用途のプログラム可能なコンピュータによって読み取ることができる記憶媒体又は記憶装置（例えば、固体メモリ若しくは媒体、又は磁気媒体若しくは光媒体）上に格納したり、又はダウンロードしたりして、コンピュータ・システムが上記記憶媒体又は記憶装置を読み取って本明細書に記載されたプロシージャを実行しようとする際に、上記コンピュータが構成され、運転されるようにするのが好ましい。また、本発明のシステムは、コンピュータ・プログラムによって構成される、コンピュータ可読記憶媒体として実装されると考えることもできる。ここで、このように構成された記憶媒体は、コンピュータ

・システムを、特定の予め定められた方式で運転して、本明細書に記載された機能を実行する。

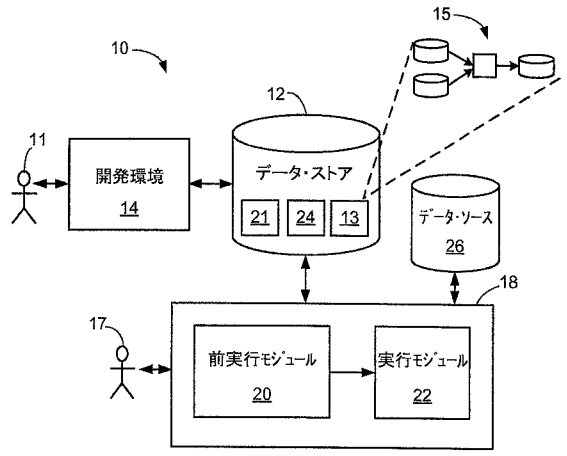
【０３１１】

本発明の幾つかの実施態様について説明してきた。しかしながら、本発明の精神及び範囲から逸脱すること無く、様々な変形を行うことができることが理解されるであろう。例えば、上述のステップの幾つかは、順序に依存しないものであってもよく、従って、上述のものとは異なる順序で実行することができる。

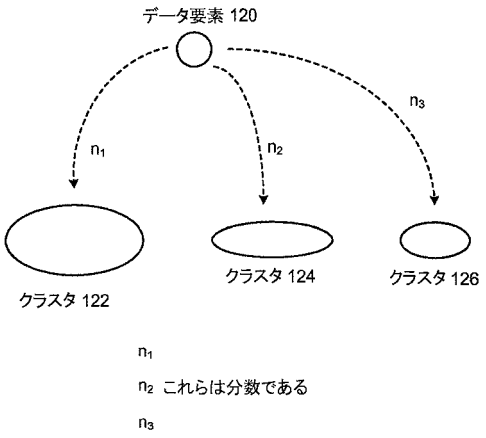
【０３１２】

先の記述は本発明の範囲を説明することを目的とするものであり、添付した特許請求の範囲によって定義される本発明の範囲を限定することを目的とするものではないこともまた理解されるべきである。例えば、上述の機能ステップの多くは、全体としての処理に実質上の影響を及ぼすこと無く、異なる順序で実行することができる。他の態様は、以下の特許請求の範囲の範囲内にある。

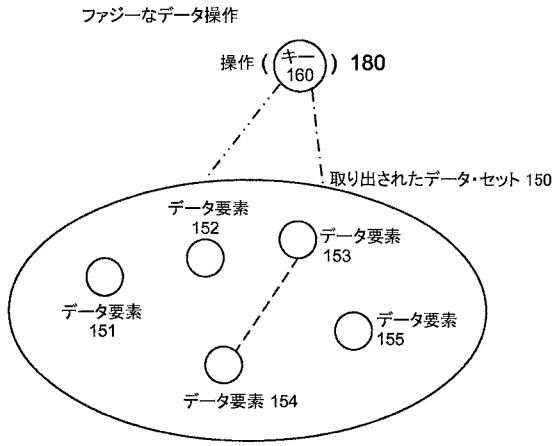
【図１】



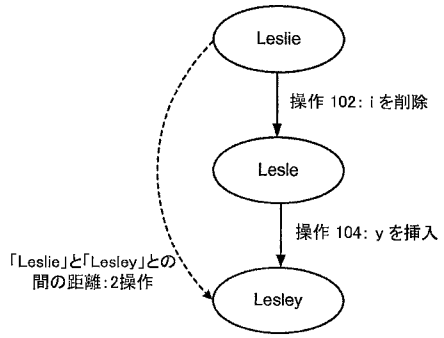
【図２Ａ】



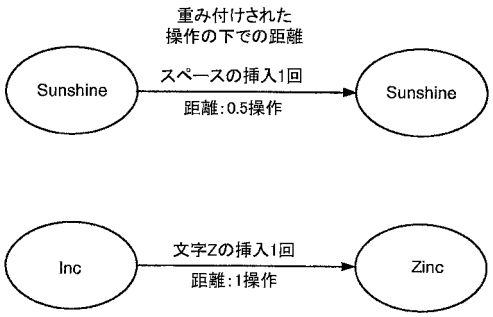
【図２Ｂ】



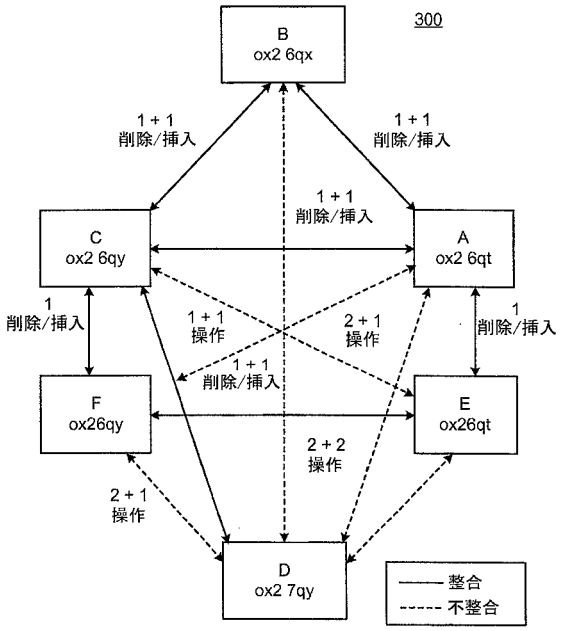
【図２Ｃ】



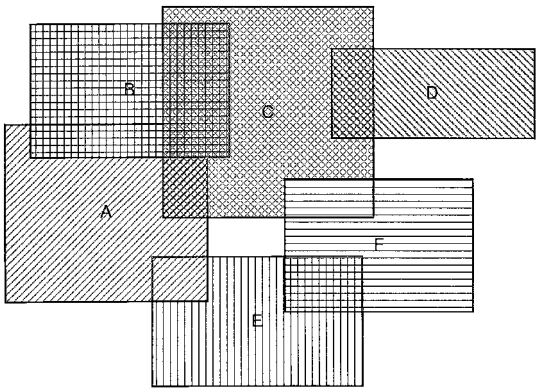
【図２Ｄ】



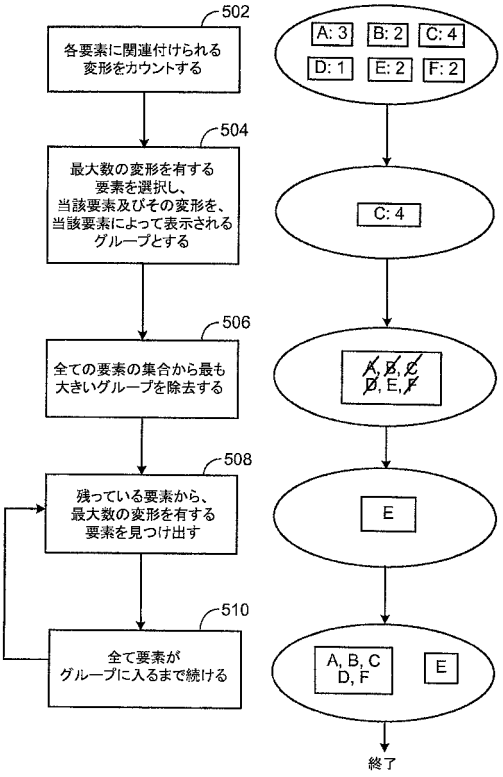
【図３】



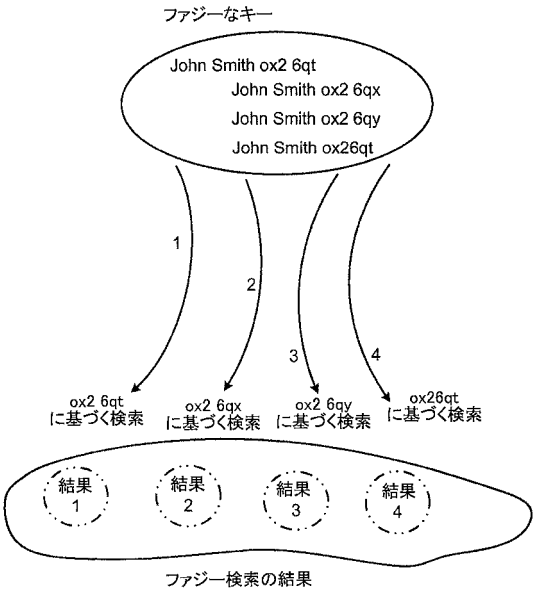
【図４】



【図５】


502
各要素に関連付けられる変形をカウントする
A: 3
B: 2
C: 4
D: 1
E: 2
F: 2
504
最大数の変形を有する要素を選択し、当該要素及びその変形を、当該要素によって表示されるグループとする
C: 4
506
全ての要素の集合から最も大きいグループを除去する
A, B, C D, E, F
508
残っている要素から、最大数の変形を有する要素を見つけ出す
E
510
全て要素がグループに入るまで続ける
A, B, C D, F
E
終了

【図６】


ファジーなキー
John Smith ox2 6qt
John Smith ox2 6qx
John Smith ox2 6qy
John Smith ox26qt
1
2
3
4
ox2 6qt に基づく検索
ox2 6qx に基づく検索
ox2 6qy に基づく検索
ox26qt に基づく検索
結果 1
結果 2
結果 3
結果 4
ファジー検索の結果

フロントページの続き

(56)参考文献 特開２００３－００６２２６（ＪＰ，Ａ）
特開平０９－０４４５１８（ＪＰ，Ａ）
特開平０６－０４４３０９（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｎ　　３／００－９９／００
Ｇ０６Ｆ　　１７／３０