

【公報種別】特許法第 17 条の 2 の規定による補正の掲載

【部門区分】第 6 部門第 3 区分

【発行日】平成20年9月11日(2008.9.11)

【公開番号】特開2006-72985(P2006-72985A)

【公開日】平成18年3月16日(2006.3.16)

【年通号数】公開・登録公報2006-011

【出願番号】特願2005-221802(P2005-221802)

【国際特許分類】

G 0 6 F 17/30 (2006.01)

【F I】

G 0 6 F 17/30 2 2 0 Z

G 0 6 F 17/30 2 1 0 D

【手続補正書】

【提出日】平成20年7月29日(2008.7.29)

【手続補正 1】

【補正対象書類名】特許請求の範囲

【補正対象項目名】全文

【補正方法】変更

【補正の内容】

【特許請求の範囲】

【請求項 1】

プロセッサによって実行されると、

データベースから複数のタプルを含むデータ集合を取得する動作（212）と、

前記データ集合の前記複数のタプルの各々について、前記データ集合内の前記タプルに対して定義された距離関数を使用して前記データ集合内の定義された数の最近隣のタプルを計算する動作（214）と、

各タプルについて最近隣のタプルの定義された数の組を近傍と定義する動作と、

各近傍について近傍の拡張比を計算する動作と、前記近傍の拡張比は前記近傍が定義されたタプルの周囲のスフェア中のタプル数であり、前記スフェアの半径は前記近傍が定義された前記タプルと最近隣の近傍との間の距離よりも大きく、

複数のタプルの複数の分離した分割を各分割に基づいてコンパクトの組及び離散的な近傍に定義する動作（216）と、

ここで、もしタプルの組中の各タプルに対して、前記タプルの組中のタプルと前記タプルの組中のその他のいずれのタプルとの間の距離が前記タプルの組中のタプルと前記タプルの組中の外のデータ集合内のその他のいずれのタプルとの間の距離よりも短ければ、前記データ集合内の前記タプルの組はコンパクトな組であり、

もし前記タプルの組中の全てのタプルの近傍の拡張比の集合された値が閾値よりも小さければ、前記タプルの組は離散的な近傍のグループであり、

ここで、分割内の各あいまいな重複のタプルが共通の実世界の实体または現象を表すよう、各分割内のタプルが相互のあいまいな重複を含み、

前記分割の各々によって1つの重複しないタプルが残るよう、前記複数のあいまいな重複のタプルの複数の分割内の重複を削除する動作（218）と

を備えた動作群を実行するプロセッサ実行可能な命令を有する1つまたは複数のプロセッサ読取可能なプログラムの記録媒体。

【請求項 2】

もしタプレットの組のカーディナリティが1に等しければ前記タプレットの組は離散的な近傍のグループであることを特徴とする請求項1に記載の1つまたは複数のプロセッサ読取可能なプログラムの記録媒体。

【請求項 3】

前記計算する動作は、前記データ集合内の前記タブルの索引によって最近隣の近傍を検索する動作を含むことを特徴とする請求項 1 に記載の 1 つまたは複数のプロセッサ読取可能なプログラムの記録媒体。

【請求項 4】

前記定義動作は、各分割に基づいて、複数のタブルの複数の分割を「コンパクトな組」、「離散的な近傍」および「小さいカーディナリティ」と定義する動作を含み、もし前記分割中のあいまいな重複のグループのカーディナリティが予め定義された定数と等しいか小さい場合に分割のカーディナリティが小さいことを特徴とする請求項 1 に記載の 1 つまたは複数のプロセッサ読取可能なプログラムの記録媒体。

【請求項 5】

各分割内の前記タブルは、互いに完全な重複をさらに含むことを特徴とする請求項 1 に記載の 1 つまたは複数のプロセッサ読取可能なプログラムの記録媒体。

【請求項 6】

あいまいな重複を削除するシステム（100）であって、

データベース（170）からの複数のタブルを含むデータ集合を取得するデータ集合取得手段（110）と、

前記データ集合の前記複数のタブルの各々について、前記データ集合内の前記タブルに対して定義された距離関数を使用して前記データ集合内の定義された数の最近隣のタブルを計算し、

各タブルについて定義された数の最近接のタブルの組を近傍として定義し、

各近傍について近傍の拡張比を計算する計算手段（120）と、前記近傍の拡張比は前記近傍が定義されたタブルの周囲のスフェア中のタブル数であり、前記スフェアの半径は前記近傍が定義された前記タブルと最近隣の近傍との間の距離よりも大きく、

各分割に基づいて複数のタブルの複数の分離した分割をコンパクトの組及び離散的な近傍に定義するための分割手段（130）と、

ここで、もしタブルの組中の各タブルに対して、前記タブルの組中のタブルと前記タブルの組中のその他のいずれのタブルとの間の距離が前記タブルの組中のタブルと前記タブルの組中の外のデータ集合内のその他のいずれのタブルとの間の距離よりも短ければ、前記データ集合内の前記タブルの組はコンパクトな組であり、

もし前記タブルの組中の全てのタブルの近傍の拡張比の集合された値が閾値よりも小さければ、前記タブルの組は離散的な近傍のグループであり、

ここで、分割内の各あいまいな重複のタブルが共通の実世界の实体または現象を表すよう、各分割内の外観上別個のタブルが互いにあいまいな重複を含み、

前記分割の各々によって 1 つの重複しないタブルが残るよう、前記複数の分割内の複数のあいまいな重複のタブルの重複を削除する重複削除手段（140）と

を備えたことを特徴とするシステム。

【請求項 7】

もしタプレットの組のカーディナリティが 1 に等しければ前記タプレットの組は離散的な近傍のグループであることを特徴とする請求項 6 に記載のシステム。

【請求項 8】

前記各分割内のタブルは、相互の完全な重複をさらに含むことを特徴とする請求項 6 に記載のシステム。

【請求項 9】

前記計算手段は、前記データ集合内の前記タブルの索引によって最近隣を検索することを特徴とする請求項 6 に記載のシステム。

【請求項 10】

前記分割手段は各分割に基づいて複数のタブルの複数の分割を、「コンパクトな組」、「離散的な近傍」および「小さいカーディナリティ」と定義し、もし前記分割中のあいまいな重複のグループのカーディナリティが予め定義された定数と等しいか小さい場合は、

分割のカーディナリティが小さいことを特徴とする請求項 6 に記載のシステム。

【請求項 1 1】

データ集合内のあいまいな重複タブルを除去するためのコンピュータによって実行される方法であって、

前記データ集合の前記複数のタブルの各々について、前記データ集合内の前記タブルに対して定義された距離関数を使用して前記データ集合内の定義された数の最近隣のタブルを計算するステップ (2 1 4) と、

各タブルについて最近隣のタブルの定義された数の組を近傍と定義するステップと、

各近傍について近傍の拡張比を計算するステップと、前記近傍の拡張比は前記近傍が定義されたタブルの周囲のスフェア中のタブル数であり、前記スフェアの半径は前記近傍が定義された前記タブルと最近隣の近傍との間の距離よりも大きく、

複数のタブルの複数の分離した分割を各分割に基づいてコンパクトの組及び離散的な近傍に定義するステップ (2 1 6) と、

ここで、もしタブルの組中の各タブルに対して、前記タブルの組中のタブルと前記タブルの組中のその他のいずれのタブルとの間の距離が前記タブルの組中のタブルと前記タブルの組中の外のデータ集合内のその他のいずれのタブルとの間の距離よりも短ければ、前記データ集合内の前記タブルの組はコンパクトな組であり、

もし前記タブルの組中の全てのタブルの近傍の拡張比の集合された値が閾値よりも小さければ、前記タブルの組は離散的な近傍のグループであり、

ここで、分割内の各あいまいな重複のタブルが共通の実世界の实体または現象を表すよう、各分割内のタブルが相互のあいまいな重複を含み、

前記分割の各々によって 1 つの重複しないタブルが残るよう、前記複数のあいまいな重複のタブルの複数の分割内の重複を削除するステップ (2 1 8) と

を備えた方法。

【請求項 1 2】

もしタプレットの組のカーディナリティが 1 に等しければ前記タプレットの組は離散的な近傍のグループであることを特徴とする請求項 1 1 に記載の方法。

【請求項 1 3】

前記計算ステップは、前記データ集合内の前記タブルの索引によって最近隣を検索することを特徴とする請求項 1 1 に記載の方法。

【請求項 1 4】

前記分割ステップは各分割に基づいて複数のタブルの複数の分割を、「コンパクトな組」、「離散的な近傍」および「小さいカーディナリティ」と定義し、もし前記分割中のあいまいな重複のグループのカーディナリティが予め定義された定数と等しいか小さい場合は、分割のカーディナリティが小さいことを特徴とする請求項 1 1 に記載の方法。