(19)日本国特許庁(JP)

再 公 表 特 許(A1)

(11)国際公開番号
WO2016/117698

発行日　平成29年11月24日(2017.11.24)

(43)国際公開日　平成28年7月28日(2016.7.28)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 17/30	(2006.01)	G06F	17/30	350C

審査請求 未請求　予備審査請求 未請求　（全 29 頁）

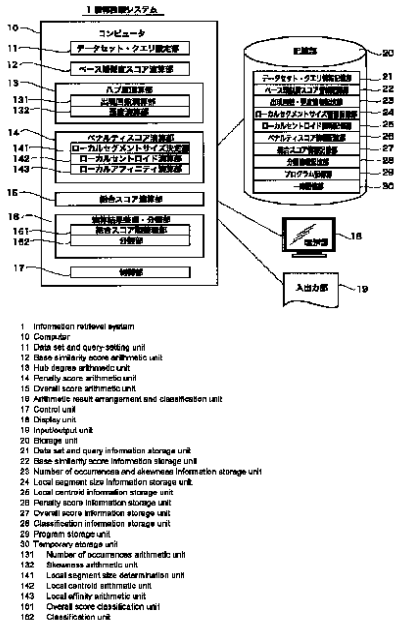
出願番号	特願2016-570732 (P2016-570732)
(21)国際出願番号	PCT/JP2016/051909
(22)国際出願日	平成28年1月22日 (2016.1.22)
(31)優先権主張番号	特願2015-11853 (P2015-11853)
(32)優先日	平成27年1月23日 (2015.1.23)
(33)優先権主張国	日本国(JP)

(71)出願人　504202472
大学共同利用機関法人情報・システム研究機構
東京都立川市緑町１０番３号
(74)代理人　100106909
弁理士　棚井　澄雄
(74)代理人　100188558
弁理士　飯田　雅人
(74)代理人　100161207
弁理士　西澤　和純
(74)代理人　100141139
弁理士　及川　周

最終頁に続く

(54)【発明の名称】情報処理装置、情報処理方法、プログラム、及び非一時記憶媒体

(57)【要約】
情報処理装置は、複数のデータにより構成される第１データセット内の処理対象データと、検索対象データとしてのクエリとの間の類似度を算出する第１算出部と、前記第１データセット内の一部のデータにより構成される第２データセットを、前記処理対象データごとに特定する特定部と、前記特定部が特定する前記第２データセットから、前記処理対象データごとの基準を算出する第２算出部と、前記第１算出部が算出する前記類似度と、前記第２算出部が算出する基準とを用いて、前記処理対象データごとにスコアを算出する第３算出部と、を含む。



【特許請求の範囲】
【請求項１】
複数のデータにより構成される第１データセット内の処理対象データと、検索対象データとしてのクエリとの間の類似度を算出する第１算出部と、
前記第１データセット内の一部のデータにより構成される第２データセットを、前記処理対象データごとに特定する特定部と、
前記特定部が特定する前記第２データセットから、前記処理対象データごとの基準を算出する第２算出部と、
前記第１算出部が算出する前記類似度と、前記第２算出部が算出する基準とを用いて、前記処理対象データごとにスコアを算出する第３算出部と、
を含む情報処理装置。
【請求項２】
前記クエリについて、前記第１データセットから第１の所定数のデータを、前記スコアに基づいて抽出する抽出部
をさらに備える請求項１に記載の情報処理装置。
【請求項３】
前記特定部は、前記第１データセット内のデータと前記処理対象データとの間の類似度に基づいて、前記第２データセットを特定する
請求項１又は請求項２に記載の情報処理装置。
【請求項４】
前記特定部は、前記第１データセット内のデータと前記処理対象データとの間の類似度が高い順に第２の所定数のデータを抽出することにより、前記第２データセットを特定する
請求項１から請求項３のいずれか一項に記載の情報処理装置。
【請求項５】
前記第２の所定数とは、前記第１データセット内の２つのデータの全ての組み合わせにおける類似度に基づいて、前記第１データセット内の各々のデータに対して当該類似度が高い順に第１の所定数のデータを抽出する場合に、前記処理対象データが抽出される回数と、前記処理対象データと当該処理対象データの基準との間の類似度と、の間の相関が最大になる数である
請求項４に記載の情報処理装置。
【請求項６】
前記第２の所定数とは、前記第１データセット内の２つのデータの全ての組み合わせにおける類似度に基づいて、前記第１データセット内の各々のデータに対して当該類似度が高い順に第１の所定数のデータを抽出する場合に、前記処理対象データが抽出される回数に関する分布の歪度が最小になる数である
請求項４に記載の情報処理装置。
【請求項７】
前記第１算出部は、内積と、コサインと、距離と、カーネルとのうちの少なくともいずれか１つに基づいて前記類似度を算出する
請求項１から請求項６のいずれか一項に記載の情報処理装置。
【請求項８】
情報処理装置が、
複数のデータにより構成される第１データセット内の処理対象データと、検索対象データとしてのクエリとの間の類似度を算出する第１ステップと、
前記第１データセット内の一部のデータにより構成される第２データセットを、前記処理対象データごとに特定する第２ステップと、
前記第２ステップにおいて特定した前記第２データセットから、前記処理対象データごとの基準を算出する第３ステップと、
前記第１ステップにおいて算出した前記類似度と、前記第３ステップにおいて算出した

基準とを用いて、前記処理対象データごとにスコアを算出する第４ステップと、
を含む情報処理方法。
【請求項９】
コンピュータに、
複数のデータにより構成される第１データセット内の処理対象データと、検索対象データとしてのクエリとの間の類似度を算出する第１ステップと、
前記第１データセット内の一部のデータにより構成される第２データセットを、前記処理対象データごとに特定する第２ステップと、
前記第１ステップにおいて特定した前記第２データセットから、前記処理対象データごとの基準を算出する第３ステップと、
前記第１ステップにおいて算出した前記類似度と、前記第３ステップにおいて算出した基準とを用いて、前記処理対象データごとにスコアを算出する第４ステップと、
を実行させるためのプログラム。
【請求項１０】
コンピュータに、
複数のデータにより構成される第１データセット内の処理対象データと、検索対象データとしてのクエリとの間の類似度を算出する第１ステップと、
前記第１データセット内の一部のデータにより構成される第２データセットを、前記処理対象データごとに特定する第２ステップと、
前記第１ステップにおいて特定した前記第２データセットから、前記処理対象データごとの基準を算出する第３ステップと、
前記第１ステップにおいて算出した前記類似度と、前記第３ステップにおいて算出した基準とを用いて、前記処理対象データごとにスコアを算出する第４ステップと、
を実行させるためのプログラムを記録したコンピュータ読み取り可能な非一時記憶媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、情報処理装置、情報処理方法、プログラム、及び非一時記憶媒体に関する。詳しくは、高次元又は／及び大規模データセットに対するｋ近傍検索で発生するハブを軽減する類似度演算システム及び類似度演算方法に関する。
２０１５年１月２３日に、日本に出願された特願２０１５－１１８５３号に基づき優先権を主張し、その内容をここに援用する。
【背景技術】
【０００２】
ｋ近傍法は実装が簡素であるにもかかわらず分類や情報検索で有効であるために多くの分類システムや情報検索システムで用いられている。しかし、データセット内のデータが高次元空間に存在するとみなせる場合（例えばデータが多数の属性を持つベクトルとして表現される場合）、他のデータのｋ近傍に頻出するデータ（ハブと呼ばれる）が出現し、結果としてｋ近傍法の性能は低下する。このハブの現象は、Ｒａｄｏｖａｎｏｖｉｃら（非特許文献１参照）により、ごく最近発見されたデータの高次元性にまつわる現象である。一方、発明者らは「（グローバル）センタリング」、すなわち、原点をデータセットの平均（グローバルセントロイド）に移動することにより、ｋ近傍法におけるハブの影響を軽減できることを発表した（非特許文献２参照）。ハブはデータセットの平均の近くに位置するデータであり、センタリングはハブを軽減するのに有効である。
【先行技術文献】
【非特許文献】
【０００３】
【非特許文献１】Ｒａｄｏｖａｎｏｖｉｃ，Ｍ；Ｎａｎｏｐｏｕｌｏｓ，Ａ．；ａｎｄ Ｉｖａｎｏｖｉｃ，Ｍ．２０１０ａ，“Ｈｕｂｓ ｉｎ ｓｐａｃｅ：Ｐｏｐｕｌａｒ

nearest neighbors in high-dimensional data.” Journal of Machine Learning Research 11：2487－2531，2010年
【非特許文献２】鈴木郁美、原一夫、新保仁「ｋ近傍法でハブを軽減する類似度尺度」、情報処理学会研究報告、自然言語処理研究会、２０１２－ＮＬ－２０９、Ｎｏ．１１、ｐｐ．１－８、２０１２、２０１３年
【発明の概要】
【発明が解決しようとする課題】
【０００４】
しかしながら、次元ｄがさほど大きくなくても、データ数ｎが大きくなると、例えばｎ＝１０，０００、ｄ＝５００、異種のハブが存在することが解った。そして、かかるハブはセンタリングでは軽減できないという問題があった。
本発明は、次元数が大きいときだけでなく、データ数が大きいときに出現するハブを軽減できる情報処理装置、情報処理方法、プログラム、及び非一時記憶媒体を提供することを目的とする。
【課題を解決するための手段】
【０００５】
本発明の第１の態様に係るハブを軽減する類似度演算システム１は、例えば図７に示すように、検索先となる高次元及び／又は大規模データセット、及び、検索対象データとしてのクエリｑを設定するデータセット・クエリ設定部１１と、データセット・クエリ設定部１１にて設定されたクエリｑ及びデータセット内の各データｘについて、ベースとする類似度尺度を用いて、クエリｑと各データｘとの類似度Ｓｉｍ（ｘ；ｑ）を演算するベース類似度スコア演算部１２と、前記設定されたデータセット内の各データｘについて、データセット内の他のデータのｋ近傍への出現回数N_k（ｘ）を計算するハブ度演算部１３と、ハブ度演算部１３で各データｘに対して計算された出現回数N_k（ｘ）を用い、各データの近傍に入るデータの集合の要素がデータセット全体を一様にカバーせず局在化する度合いに応じて、各データの近傍に入るデータからなるローカルな部分集合の中心としてのローカルセントロイドc_κ（ｘ）を計算すると共に、各データについてローカルセントロイドc_κ（ｘ）を用いた類似度演算により各データのペナルティスコアとして用いるローカルアフィニティＬｏｃａｌＡｆｆｉｎｉｔｙ（ｘ）を計算することによりローカライズドセンタリングの前工程を実行するペナルティスコア演算部１４と、ベース類似度スコア演算部１２にて計算されたクエリと各データとの類似度Ｓｉｍ（ｘ；ｑ）とペナルティスコア演算部１４にて計算されたローカルアフィニティＬｏｃａｌＡｆｆｉｎｉｔｙ（ｘ）を用い、各データに対して総合スコアを計算することによりローカライズドセンタリングの後工程を実行する総合スコア演算部１５とを備える。
【０００６】
ここにおいて、高次元及び／又は大規模データセットとは、次元数が大きい（あるいは次元数が大きいとみなせる）及び／又はデータ数が大きいデータセットをいう。例えば、生命情報学分野では、機能未知の配列データの機能を、機能既知の配列データに対する相同性を頼りに予測しようとするが、検索先として使用される機能既知のＤＮＡの塩基配列あるいはタンパク質のアミノ酸配列のデータベースは、過去数十年間に世界中の研究成果を集めた膨大な数の配列データを格納する、大規模なデータセットである。また、ＤＮＡの塩基配列あるいはタンパク質のアミノ酸配列の相同性検索アルゴリズムとして広く使われるブラスト（ＢＬＡＳＴ：Ｂａｓｉｃ Ｌｏｃａｌ Ａｌｉｇｎｍｅｎｔ Ｓｅａｒｃｈ Ｔｏｏｌ）は、個々の配列データを特徴付ける属性として、配列内に出現しうるあらゆる部分配列（及びその組み合わせ）を用いるため、配列データを本質的に高次元データとみなして扱う。ここでの高次元及び／又は大規模データセットとは、ハブが発生するあらゆるデータセットを意味する。なぜなら、現在までに報告されているハブは、データセットが高次元である場合と、データセットが大規模である場合に限られて発生しているからである。敢えて数値化するとすれば例えば、高次元は人間の直感で理解が難しくなる４

次元以上をいい、大規模はデータ数１００以上をいい、１０００以上が好ましい。なお、後者のハブは、本発明者らにより発見されたハブである。

【０００７】

また、ここでのデータセットとは、検索対象データ（クエリ）に対する類似データの検索先となるデータ集合をいう。例えば、生命情報学分野における配列の相同性検索においては、データセットとして用いられるのは、過去に世界中の研究成果として蓄積された塩基配列データ又はアミノ酸配列データからなる配列データ集合群、又はそれらの部分集合である。

また、ここでの類似度尺度とは、２つのデータの類似性を測る尺度として使用できるものすべてを含む。典型的には、内積、コサイン、距離である。内積は２つのベクトルデータのスカラー積であり、コサインは長さ１に規格化された２つのベクトルデータの内積である。さらに、内積の一般化とみなせる（機械学習分野で主に呼ばれるところの各種の）カーネルも含む。距離の典型は、２つのベクトルデータ間のユークリッド距離（l^2ノルム）であるが、ユークリッド距離を一般化した距離（マンハッタン距離やl^pノルムなど）も含む。さらに、ドメインの知識を持つ人間が各タスクの目的に応じて適宜定めた類似度スコア計算方法（ＢＬＡＳＴなど）が出力する類似度も、ここでの類似度尺度に含まれる。

【０００８】

また、「ベースとする類似度尺度」は、ベース類似度スコア演算部にて類似度Ｓｉｍ（ｘ；ｑ）を計算するとき、及びペナルティスコア演算部にてローカルアフィニティＬｏｃａｌＡｆｆｉｎｉｔｙ（ｘ）を計算するときに使用される類似度尺度である。類似度Ｓｉｍ（ｘ；ｑ）を計算するとき及びローカルアフィニティＬｏｃａｌＡｆｆｉｎｉｔｙ（ｘ）を計算するときに使用される類似度尺度として、上記全ての類似度尺度を使用できる。また、高次元及び／又は大規模データセットを検索先とする場合、ベース類似度スコア演算部にて計算されるスコアである類似度Ｓｉｍ（ｘ；ｑ）はハブを発生させ易い。そこで、ベース類似度スコア演算部にて計算されるスコアをそのまま用いず、ハブに成り易い程度に応じて各データに対してペナルティスコアを計算し、ベースとするスコアからペナルティスコアを差し引いた総合スコアを用いる。すなわち、ハブに成り易いデータの総合スコアを小さくすることで、ハブに成り易いデータを他のデータのｋ近傍に入りにくくするものである。

【０００９】

また、ハブ度演算部１３は、データセット内の各データｘに対し、他のデータのｋ近傍に出現する回数$N_k(x)$を計算する。さらに、$N_k(x)$の分布の歪度を計算する。ｋ近傍になるとは、類似度が高い方からｋ番目以内になることをいう。また、$N_k(x)$が大きいほど、データｘがハブであることを意味する。さらに、$N_k(x)$の分布の歪度は、$S_{Nk}=E[(N_k-\mu_{Nk})^3/\sigma_{Nk}^3]$（$E[\ ]$は期待値を計算するオペレータ、$\mu_{Nk}$と$\sigma_{Nk}^2$はそれぞれ$N_k(x)$の分布の平均と分散である）と計算する。$N_k(x)$の分布の歪度が大きいほど、データセットにハブが存在することを意味する。

また、大規模なデータセット、すなわち、データ数ｎが大きいデータセットにおいては、各データｘについてその近傍に入るデータの集合を考えたとき、その集合がデータセット全体を一様にカバーせず、データセットに局在する部分集合を形成し易い、という事実がある。そこで、ローカルセグメントサイズκをパラメタとし、各データに対してそのκ近傍に入るデータをローカルな部分集合として定め、このローカルな部分集合の中心としてローカルセントロイド$c_\kappa(x)$を定義する。また、各データの近傍に入るデータの集合の要素がデータセット全体を一様にカバーせず局在化する度合いに応じて、ローカルセントロイド$c_\kappa(x)$を計算するのは、ハブ度演算部１３で各データｘに対して計算された出現回数$N_k(x)$に応じてペナルティを与えるためである。なお、ローカルセントロイドは各データにより異なるが必ず存在するとみなしてよい。したがって、ローカルセントロイドの有無を判断する必要はない。

具体的には、ローカルセントロイド$c_\kappa(x)$は例えば、式（２Ａ）で求められる（段

落００３１参照）。パラメタκにより隣接データの範囲が変わるので、ローカルセントロイドｃ$_{κ}$（ｘ）も変わり得る。ローカライズドセンタリングとは、ハブ度と相関が高くなるようなローカルなセントロイドを見つけ出し、ローカルなセントロイドとの類似度をペナルティとして与えることをいう。例えば、式（５）が使用される（段落００３７参照）。類似度尺度にはベースとする類似度尺度を使用できる。ここでは、ローカライズドセンタリングは、先にローカルアフィニティを計算し、その後総合スコアを計算することにより実行される。すなわち、ベース類似度スコア演算部にて計算されるスコア（ベースとするスコア）から、ハブに成り易い程度に応じて各データに対して付与するペナルティスコア（ローカルアフィニティ）を計算し、ベースとするスコアからペナルティスコアを差し引いた総合スコアを計算することにより実行される。

また、ローカルアフィニティとは、各データｘ∈Ｄについて、ｘとそのκ近傍（κＮＮすなわちκ－Ｎｅａｒｅｓｔ－Ｎｅｉｇｈｂｏｒ）に属するデータ間の平均類似度として定義される（段落００３１、式（２）参照）。また、ペナルティスコアとして用いるとは、ローカルセントロイドｃ$_{κ}$（ｘ）近傍のデータの類似度スコアを低くするために用いることをいい、ここでは、ローカルアフィニティをペナルティスコアとして差し引いている。

【００１０】

本態様のように構成すると、ローカルセントロイドという新しい概念に基づいて計算されるローカルアフィニティが、データのハブ度と高い相関を持つため、それをペナルティスコアとして用いることにより、次元数が大きいときだけでなく、データ数ｎが大きいときに出現するハブをも軽減できる。なお、既存手法において用いられたグローバルセントロイドでは、次元数ｄが大きいときに生じるハブは軽減できるがデータ数ｎが大きいときに生じるハブを軽減できなかった。

【００１１】

本発明の第２の態様に係るハブを軽減する類似度演算システム１は、第１の態様において、ベースとする類似度尺度として、内積、コサイン、距離、内積の一般化とみなせるカーネル、又は、ドメインの知識を持つ人間が各タスクの目的に応じて適宜定めた類似度スコア計算方法が出力する類似度尺度を用いる。

ここにおいて、内積の一般化とみなせるカーネルとは、機械学習分野で主に呼ばれるところの各種のカーネルをいう。また、ドメインの知識を持つ人間が各タスクの目的に応じて適宜定めた類似度スコア計算方法が出力する類似度にはＢＬＡＳＴなどが含まれる。

本態様のように構成すると、内積、コサイン、距離、カーネル、ＢＬＡＳＴスコア等は演算が簡素で扱いやすく、また、ローカルアフィニティをペナルティスコアとして用いるローカライズドセンタリングによりハブを軽減できる実績が得られている。なお、ローカライズドセンタリングによりハブが軽減しさえすれば、その他の類似度尺度も使用できる。

【００１２】

本発明の第３の態様に係るハブを軽減する類似度演算システム１は、第１ないし第３のいずれかの態様において、ペナルティスコア演算部１４は、ローカルアフィニティＬｏｃａｌＡｆｆｉｎｉｔｙ（ｘ）を計算する際のパラメタであるローカルセグメントサイズκを、出現回数Ｎ$_{k}$（ｘ）とローカルアフィニティＬｏｃａｌＡｆｆｉｎｉｔｙ（ｘ）との相関が最大になるように定める。

【００１３】

ここにおいて、ローカルアフィニティは、各データｘ∈Ｄについて、ｘとそのκ近傍（κはローカルセグメントサイズであり、ローカルアフィニティを定めるパラメタである）に属するデータ間の平均類似度として定義され、明細書中の式（２）で表される（段落００３１参照）。

本態様のように構成すると、ローカルセグメントサイズκを適切に定めることができる。

【００１４】

本発明の第４の態様に係るハブを軽減する類似度演算方法は、例えば図８に示すように、検索先となる高次元及び／又は大規模データセット、及び、検索対象データとしてのクエリの設定を行うデータセット・クエリ設定工程（Ｓ１０１）と、データセット・クエリ設定工程（Ｓ１０１）にて設定されたクエリとデータセット内の各データについて、ベースとする類似度尺度を用いてクエリと各データとの類似度を計算するベース類似度スコア演算工程（Ｓ１０２）と、設定されたデータセット内の各データについて、データセット内の他のデータのｋ近傍への出現回数N_k（ｘ）を計算するハブ度演算工程（Ｓ１０３）と、ハブ度演算工程（Ｓ１０３）で各データに対して計算された、出現回数N_k（ｘ）を用い、各データの近傍に入るデータの集合の要素がデータセット全体を一様にカバーせず局在化する度合いに応じて、各データの近傍に入るデータからなるローカルな部分集合の中心としてのローカルセントロイドc_κ（ｘ）を計算すると共に、各データについてローカルセントロイドc_κ（ｘ）を用いた類似度演算により各データのペナルティスコアとして用いるローカルアフィニティＬｏｃａｌＡｆｆｉｎｉｔｙ（ｘ）を計算することによりローカライズドセンタリングの前工程を実行するペナルティスコア演算工程（Ｓ１０６）と、ベース類似度スコア演算工程（Ｓ１０２）にて計算されたクエリと各データとの類似度Ｓｉｍ（ｘ；ｑ）とペナルティスコア演算工程（Ｓ１０６）にて計算されたローカルアフィニティＬｏｃａｌＡｆｆｉｎｉｔｙ（ｘ）を用い、各データに対して総合スコアを計算することによりローカライズドセンタリングの後工程を実行する総合スコア演算工程（Ｓ１０７）とを備える。

【００１５】

ここにおいて、ローカライズドセンタリングは前工程と後工程からなる。

本態様のように構成すると、（次元数ｄが大きいときに生じるハブは軽減できるがデータ数ｎが大きいときに生じるハブを軽減できない既存手法で用いられたグローバルセントロイドに対し）本発明が新たに導入するローカルセントロイドという概念に基づいて計算されるローカルアフィニティは各データのハブ度と高い相関を持つため、それをペナルティスコアとして用いることにより、次元数が大きいときだけでなく、データ数ｎが大きいときに出現するハブをも軽減できる。

【００１６】

本発明の第５の態様に係るハブを軽減する類似度演算方法は、第４の態様において、ベースとする類似度尺度として、内積、コサイン、距離、内積の一般化とみなせるカーネル、又は、ドメインの知識を持つ人間が各タスクの目的に応じて適宜定めた類似度スコア計算方法が出力する類似度これらに基づいて作成された類似度尺度を用いる。

このように構成すると、内積、コサイン、距離、カーネル、ＢＬＡＳＴスコア等は演算が簡素で扱いやすく、また、ローカルアフィニティをペナルティスコアとして用いるローカライズドセンタリングによりハブを軽減できる実績が得られている。なお、ローカライズドセンタリングによりハブが軽減する限りは、その他の類似度尺度も使用できる。

【００１７】

本発明の第６の態様に係るハブを軽減する類似度演算方法は、第４又は第５の態様において、ペナルティスコア演算工程（Ｓ１０６）は、ローカルアフィニティＬｏｃａｌＡｆｆｉｎｉｔｙ（ｘ）を計算する際のパラメタであるローカルセグメントサイズκを、出現回数N_k（ｘ）とローカルアフィニティＬｏｃａｌＡｆｆｉｎｉｔｙ（ｘ）との相関が最大になるように定める。

このように構成すると、ローカルセグメントサイズκを適切に定めることができる。

【００１８】

本発明の第７の態様に係るプログラムは、第４ないし第６のいずれかの態様におけるハブを軽減する類似度演算方法をコンピュータに実行させるためのプログラムである。

【００１９】

本発明の第８の態様に係る記録媒体は、第７の態様におけるプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【発明の効果】

【００２０】
本発明の一態様によれば、次元数ｄが大きいときに出現するハブのみならず、データ数ｎが大きいときに出現するハブを軽減できる情報処理装置、情報処理方法、プログラム、及び非一時記憶媒体を提供できる。

【図面の簡単な説明】

【００２１】
【図１】高次元人工データセット（ｄ＝４０００）におけるハブを説明するための第１図である。図１は、センタリング（グローバルセンタリング）前のN_{10}（ｋ＝１０のとき、各データが他のデータのｋ近傍に出現する回数）の分布を示す図である。
【図２】高次元人工データセット（ｄ＝４０００）におけるハブを説明するための第２図である。図２は、センタリング（グローバルセンタリング）後のN_{10}（ｋ＝１０のとき、各データが他のデータのｋ近傍に出現する回数）の分布を示す図である。
【図３】高次元人工データセット（ｄ＝４０００）におけるハブを説明するための第３図である。図３は、各データのN_{10}とグローバルアフィニティ（データセットの平均への類似度）との関係（グローバルセンタリング前）を示す散布図である。
【図４】大規模人工データセット（ｎ＝１００００）におけるハブを説明するための第１図である。図４は、センタリング（グローバルセンタリング）前のN_{10}の分布を示す図である。
【図５】大規模人工データセット（ｎ＝１００００）におけるハブを説明するための第２図である。図５は、センタリング（グローバルセンタリング）後のN_{10}の分布を示す図である。
【図６】大規模人工データセット（ｎ＝１００００）におけるハブを説明するための第３図である。図６は、各データのN_{10}とグローバルアフィニティ（データセットの平均への類似度）との関係（グローバルセンタリング前）を示す散布図である。
【図７】データ数ｎ及び次元数ｄを変数として、生成した様々な人工データセットを用いて作成した、N_{10}分布の歪度の等高線第１図である。図７は、センタリング（グローバルセンタリング）前の等高線図である。数値が高いほど、当該データセットにハブが多く存在することを示す。なお、等高線図における＋マークは、生成したデータセット（のデータ数ｎと次元数ｄ）に対応する。
【図８】データ数ｎ及び次元数ｄを変数として、生成した様々な人工データセットを用いて作成した、N_{10}分布の歪度の等高線第２図である。図８は、センタリング（グローバルセンタリング）後の等高線図である。数値が高いほど、当該データセットにハブが多く存在することを示す。
【図９】図７、８と同じ人工データセットを用いたときの、ローカライズドセンタリング後のN_{10}分布の歪度の等高線図である。
【図１０】図７、８と同じ人工データセットを用いたときの、グローバル対ローカルアフィニティ比を示す等高線図である。
【図１１】図４－６と同じ大規模人工データセット（ｎ＝１００００）を用いたときの、各データのＮ１０とローカルアフィニティ（ローカルセントロイドへの類似度）の相関（ローカライズドセンタリング前）を示す散布図である。
【図１２】図４－６と同じ大規模人工データセット（ｎ＝１００００）を用いたときの、ローカライズドセンタリング後のN_{10}分布を示す図である。
【図１３】図１－３と同じ高次元人工データセット（ｄ＝４０００）を用いたときの、各データのN_{10}とローカルアフィニティ（ローカルセントロイドへの類似度）の相関を示す散布図である。
【図１４】実施例１におけるハブを軽減する類似度演算システムの構成例を示す図である。
【図１５】実施例１におけるハブを軽減する類似度演算方法の処理フロー例を示す図である。
【図１６】実世界データセットに対する、データ数ｎとN_{10}分布の歪度との関係を示す

折れ線グラフである。（ｉ）はグローバルセンタリングやローカライズドセンタリングを施す前の折れ線グラフ（比較のベースライン）である。（ｉｉ）はグローバルセンタリング後、（ｉｉｉ）はローカライズドセンタリング後の折れ線グラフである。（ｉｖ）はグローバル対ローカルアフィニティ比の折れ線グラフである。使用データセットはＷｅｂＫＢである。

【図１７】実世界データセットに対する、データ数ｎとN_{10}分布の歪度との関係を示す折れ線グラフである。（ｉ）はグローバルセンタリングやローカライズドセンタリングを施す前の折れ線グラフ（比較のベースライン）である。（ｉｉ）はグローバルセンタリング後、（ｉｉｉ）はローカライズドセンタリング後の折れ線グラフである。（ｉｖ）はグローバル対ローカルアフィニティ比の折れ線グラフである。使用データセットはＲｅｕｔｅｒｓ－５２である。

【図１８】実世界データセットに対する、データ数ｎとN_{10}分布の歪度との関係を示す折れ線グラフである。（ｉ）はグローバルセンタリングやローカライズドセンタリングを施す前の折れ線グラフ（比較のベースライン）である。（ｉｉ）はグローバルセンタリング後、（ｉｉｉ）はローカライズドセンタリング後の折れ線グラフである。（ｉｖ）はグローバル対ローカルアフィニティ比の折れ線グラフである。使用データセットはＴＤＴ２－３０である。

【図１９】実世界データセットに対する、データ数ｎとN_{10}分布の歪度との関係を示す折れ線グラフである。（ｉ）はグローバルセンタリングやローカライズドセンタリングを施す前の折れ線グラフ（比較のベースライン）である。（ｉｉ）はグローバルセンタリング後、（ｉｉｉ）はローカライズドセンタリング後の折れ線グラフである。（ｉｖ）はグローバル対ローカルアフィニティ比の折れ線グラフである。使用データセットは２０Ｎｅｗｓｇｒｏｕｐｓである。

【図２０】図１６－１９で使用したのと同じ実世界データセットを用いて行った、ｋ近傍法による多クラス分類の精度、及び、N_{10}分布の歪度を示す表形式の第１図である。

【図２１】図１６－１９で使用したのと同じ実世界データセットを用いて行った、ｋ近傍法による多クラス分類の精度、及び、N_{10}分布の歪度を示す表形式の第２図である。

【図２２】図１６－１９で使用したのと同じ実世界データセットを用いて行った、ｋ近傍法による多クラス分類の精度、及び、N_{10}分布の歪度を示す表形式の第３図である。

【図２３】図１６－１９で使用したのと同じ実世界データセットを用いて行った、ｋ近傍法による多クラス分類の精度、及び、N_{10}分布の歪度を示す表形式の第４図である。

【発明を実施するための形態】

【００２２】

図面を参照して以下に本発明の実施の形態について説明する。なお、各図において、互いに同一又は相当する部分には同一符号を付し、重複した説明は省略する。

【００２３】

〔高次元データセット中でのハブ〕

ハブは高次元空間での最近傍検索に係る現象として知られている。$D \subset R^d$をｄ次元空間上のデータセットとする。N_k（ｘ）は、データ$x \in D$がＤに含まれる他データのｋ近傍に含まれる回数を示すものとする。次元ｄが増加すると、N_kの分布の形状は右に長い尾を引くように変化する。すなわち、少数のデータが予期しない高いN_k値を持つ。このようなデータをハブという。ここで、テキスト文書集合を模した人工データセット（各データが一つのテキスト文書に対応する）を用いて、ハブの出現を例示する。テキスト文書集合を模したデータセットＤを生成するに際し、テキスト文書の表現として通常用いられるバッグオブワーズ「ｂａｇ－ｏｆ－ｗｏｒｄｓ」モデルによって、非負値を要素として持つｄ次元ベクトルを、各テキスト文書を模したベクトルデータとして生成する。ｎ＝｜Ｄ｜を生成するデータ数とするとき、まずｎ個のベクトルを生成し、これらをゼロベクトルとして初期化する。次に、各次元ｉ＝１，・・・，ｄについて、対数正規分布ＬｏｇＮｏｒｍ（５，１）（５，１はそれぞれ、平均、分散に係るパラメタ）に従う実数を生成し、端数を丸めた整数n_iを得る。そして、データセットＤのｎ個のベクトルから一様分布

に従ってn_i個のベクトルを選び、選ばれたベクトルの第ｉ成分を０でない値に置換する。具体的には、〔０，１〕の一様分布に従う値（０～１の間の実数）に置換する。最後に、全てのベクトルを長さが１になるように正規化する。ここでは、一例として、類似度尺度として内積を使用するが、全てのベクトルの長さが１に等しいため、内積はコサインに等しい。なお、上述したように、類似度尺度としては、内積、コサイン、距離等の別の尺度をベースとしてもよく、いずれの尺度を用いるかは、ユーザにより選択可能としてもよい。

【００２４】

図１－３は高次元データセット（ｄ＝４０００）におけるハブを説明するための図である。図１、２はそれぞれセンタリング（グローバルセンタリング）前後のN_{10}（ｋ＝１０のとき、各々のデータが他のデータのｋ近傍に出現する回数）の分布を示す図である。図において、ハブ、すなわち、極端に大きなN_{10}値を持つデータの存在を確認できる。

データセットにどの程度ハブが存在するかを、Ｒａｄｏｖａｎｏｖｉｃら（Ｒａｄｏｖａｎｏｖｉｃ，Ｍ；Ｎａｎｏｐｏｕｌｏｓ，Ａ．；ａｎｄ Ｉｖａｎｏｖｉｃ，Ｍ．２０１０ａ，Ｈｕｂｓ ｉｎ ｓｐａｃｅ：Ｐｏｐｕｌａｒ ｎｅａｒｅｓｔ ｎｅｉｇｈｂｏｒｓ ｉｎ ｈｉｇｈ－ｄｉｍｅｎｓｉｏｎａｌ ｄａｔａ．Ｊｏｕｒｎａｌ ｏｆ Ｍａｃｈｉｎｅ Ｌｅａｒｎｉｎｇ Ｒｅｓｅａｒｃｈ １１：２４８７－２５３１）に従って、N_k分布の歪度 $S_{Nk}=E〔(N_k-\mu_{Nk})^3/\sigma_{Nk}^3〕$によって評価する。ここにE〔 〕は期待値を計算するオペレータ、μ_{Nk}とσ_{Nk}はそれぞれN_k分布の平均と分散である。歪度は分布の対称性の程度を測るための標準尺度である。その値はガウス分布のような対称な分布では０であり、長い右又は左の尾を持つ分布については正又は負である。図１の人工データセットにおいては、S_{N10}は大きな正値４．４５である。

図３はN_{10}とデータセットのグローバルセントロイド（データセットの平均）への類似度（すなわち、後述するグローバルアフィニティ）との関係（グローバルセンタリング前）を示す散布図である。図が示すように、それらの間に強い相関が存在する。このように、高次元データセットでは、グローバルセントロイドに類似するデータがハブになる。更なる例としては上記Ｒａｄｏｖａｎｏｖｉｃらを参照されたい。

【００２５】

〔ハブ軽減法としてのセンタリング〕

グローバルセンタリング（単にセンタリングともいう）とは、特徴空間の原点をデータセットのグローバルセントロイド（平均）に移動する変換をいう。グローバルセントロイドｃは次式で表される。

$$c=(1/|D|)\Sigma_{x\in D}x$$

センタリングは、データセットの観測バイアスを除去する古典的技術であるが、最近、発明者らによって、ハブを除く効果を持つことが確認された（Ｓｕｚｕｋｉ，Ｉ．；Ｈａｒａ，Ｋ．；Ｓｈｉｍｂｏ，Ｍ．；Ｓａｅｒｅｎｓ，Ｍ．；ａｎｄ Ｆｕｋｕｍｉｚｕ，Ｋ．２０１３．“Ｃｅｎｔｅｒｉｎｇ ｓｉｍｉｌａｒｉｔｙ ｍｅａｓｕｒｅｓ ｔｏ ｒｅｄｕｃｅ ｈｕｂｓ．” Ｉｎ Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ ２０１３ Ｃｏｎｆｅｒｅｎｃｅ ｏｎ Ｅｍｐｉｒｉｃａｌ Ｍｅｔｈｏｄｓ ｉｎ Ｎａｔｕｒａｌ Ｌａｎｇｕａｇｅ Ｐｒｏｃｅｓｓｉｎｇ，６１３－６２３．）センタリングの重要な意義は、データとグローバルセントロイドｃ間の類似度（内積）を一定（実際にはゼロ）にすることである。具体的には、与えられたクエリ$q\in R^d$とデータ$x\in D$の内積

$$Sim(x,q)\equiv<x,q>$$

は、センタリング後、次の様に変化する。

$$Sim^{CENT}(x,q)\equiv<x-c,q-c>$$

ここで、ｑ＝ｃと置換すると、任意の$x\in R^d$に対して、

$$Sim^{CENT}(x,c)=0$$

となる。

【００２６】
このことは、グローバルセントロイドに特別に高い類似度を持つデータはもはや存在しないことを意味する。ここで期待されるのは、図１と図３で使用したデータセット（グローバルセントロイドに類似するデータがハブとなっていた）に対する、センタリングによるハブ軽減の効果である。期待されたとおり、このデータセットについてのN_{10}分布の歪度S_{N10}は、センタリング前の４．４５から、センタリング後に０．２７に減少する。実際、センタリング後の図２で見られるN_{10}分布はほぼ対称である。

【００２７】
〔大規模データセット中でのハブ〕
しかしながら、（グローバル）センタリングではハブが軽減されない場合がある。さほど高次元ではないが、データ数が大きい次元数ｄ＝５００、データ数ｎ＝１００００として前記と同様に人工データセットＤを生成し、この人工データセットに出現するハブについて以下に説明する。

【００２８】
図４－６は上記データセットにおけるハブを説明するための図である。図４、５はそれぞれ（グローバル）センタリング前後のN_{10}（ｋ＝１０のとき、各々のデータが他のデータのｋ近傍に出現する回数）の分布を示す図である。図６はN_{10}とグローバルセントロイドへの類似度（すなわち、後述するグローバルアフィニティ）との関係（センタリング前）を示す散布図である。
このデータセットではN_{10}分布の歪度S_{N10}はセンタリング前に５．８８、センタリング後に５．８２である。つまり、センタリングによりハブはさほど軽減されない。図４及び図５に示されるように、N_{10}分布の形状はセンタリング前後でほぼ同一である。さらに、図６の散布図に示されるように、このデータセットにおいてはセンタリング法が成功するために必要とされる条件（大きなN_k値を持つデータ、すなわち、ハブはグローバルセントロイドと高い類似度を持つ）を満たさない。

【００２９】
データセットを、データ数ｎを１００～１００００、次元数ｄを１００～４０００の範囲で変えて生成し、各データセットについて、N_{10}分布の歪度を計算する。各ｎとｄの組み合わせについて、１０個のデータセットを生成し、平均の歪度を計算する。

【００３０】
図７、８は、データ数ｎ及び次元数ｄを変数としたN_{10}分布の歪度の等高線図である。図７はセンタリング前、図８はセンタリング後の等高線図である。高い値はハブの存在を意味する。＋マークは生成したデータセット（のデータ数ｎと次元数ｄ）に対応する。図７を見ると、ハブは左上と右下の領域のデータセットで発生していることが分かる。他方、図８によれば、左上の領域のハブはセンタリング後に消え、右下の領域のハブはセンタリングしても消えず、センタリング前と殆ど同程度残ることが分かる。
何故、右下の領域のデータセットでは、（グローバル）センタリングによってハブが軽減されないのかを調査するために、各データの特徴を調べるための２つの量、ローカルアフィニティとグローバルアフィニティを導入する。

【００３１】
ローカルアフィニティ　ＬｏｃａｌＡｆｆｉｎｉｔｙ（ｘ）は、各データｘ∈Ｄについて、ｘとそのκ近傍（κＮＮすなわちκ－Ｎｅａｒｅｓｔ－Ｎｅｉｇｈｂｏｒ）に属するデータ間の平均類似度として定義され、次のように計算される。

$$\mathrm{LocalAffinity}(x) \equiv (1/\kappa)\,\Sigma_{x' \in \kappa NN(x)} <x, x'> = <x, c_\kappa(x)> \qquad (2)$$

ここに、ローカルセグメントサイズκは、各データｘに対してローカルセントロイドを計算するために使用する隣接データの範囲を定めるパラメタである。
そして、次式が定める$c_\kappa(x)$を、データｘのローカルセントロイドと呼ぶ（式において、ｋＮＮではなくκＮＮが使われることに注意）。

$$c_\kappa(x) = (1/\kappa)\,\Sigma_{x' \in \kappa NN(x)} x' \qquad (2A)$$

【００３２】
グローバルアフィニティ ＧｌｏｂａｌＡｆｆｉｎｉｔｙ（ｘ）は、各データｘ∈Ｄについて、ｘとＤ中の全てのデータ間の平均類似度として定義される。

ＧｌｏｂａｌＡｆｆｉｎｉｔｙ（ｘ）
≡（１／｜Ｄ｜）$\Sigma_{x' \in D}$＜ｘ，ｘ’＞＝＜ｘ，ｃ＞ （３）

ここに、データセットのグローバルセントロイド（データセットの平均）は、
ｃ＝（１／｜Ｄ｜）$\Sigma_{x' \in D}$ｘ’である。

上式から解るように、ｘのグローバルアフィニティは単純にｘとｃの類似度である。もし、Ｄが非負値を要素として持つベクトルデータの集合であるなら、全てのｘ∈Ｄについて、

ＬｏｃａｌＡｆｆｉｎｉｔｙ（ｘ）≧ＧｌｏｂａｌＡｆｆｉｎｉｔｙ（ｘ）≧０

である。ここで、例えばテキスト文書は通常、（ｔｆ－ｉｄｆ重み付け）ワードカウントベクトル、すなわち非負値を要素として持つベクトルとして表現されるため、上記不等式が成り立つ。さらに、κ＝｜Ｄ｜なら、次式が成立する。

ＬｏｃａｌＡｆｆｉｎｉｔｙ（ｘ）＝ＧｌｏｂａｌＡｆｆｉｎｉｔｙ（ｘ）

【００３３】
次に、グローバル対ローカルアフィニティ比

ＬｏｃａｌＡｆｆｉｎｉｔｙ（ｘ）／ＧｌｏｂａｌＡｆｆｉｎｉｔｙ（ｘ）

について説明する。

各データｘが非負ベクトルとして表されるデータセットＤにおいては、あらゆるデータｘ∈Ｄに対して、この比は〔０，１〕に入る。この比は、Ｄ中でデータｘがどの程度ローカル化しているかの指標となる。すなわち、（あるκ≪｜Ｄ｜に対して）比が０に近いなら、ｘは近傍データと特別に高い類似度を持つ、つまりｘはローカル化されている局在データの集合を持っている。比がκに関わりなく１に近いなら、ｘは特別に高い類似度を持つデータをＤ中に持たない、つまりｘはローカル化されていない。

【００３４】
図９、１０はローカライズドセンタリングに係る図（その１）である。データ数と次元数を変えた（図７－８に使用したのと同じ）様々なデータセットを用いて作成された。図９はローカライズドセンタリング後の$Ｎ_{10}$分布の歪度を示す等高線図である。図１０はグローバル対ローカルアフィニティ比を示す等高線図である。

ローカライズドセンタリングとは、ハブ度と相関が高くなるようなローカルなセントロイドを見つけ出し、ローカルなセントロイドとの類似度をペナルティとして与えることをいう。ローカライズドセンタリングは、例えば、先にローカルアフィニティを計算し、その後総合スコアを計算することにより実行される。

グローバル対ローカルアフィニティ比は個々のデータに対して定義されるので、データセットに対するグローバル対ローカルアフィニティ比として、データセット中の全データのグローバル対ローカルアフィニティ比の平均を用いた（ローカルセグメントサイズはκ＝２０に固定）。これを図８と比較すると、（グローバル）センタリング後に歪度が軽減されず大きいままである右下の領域は、図１０ではグローバル対ローカルアフィニティ比が小さい領域に対応する。このことは、この領域のデータセットにおいてはデータがローカル化していることを示す。このようなデータセットにおいては、ローカルアフィニティがグローバルアフィニティよりハブを軽減するために関連性が高いということを示唆する。

【００３５】
図１１、１２はローカライズドセンタリングに係る図（その２）である。図４－６で使用した大規模データセット（ｎ＝１００００）について、図１１は$Ｎ_{10}$とローカルアフィニティの相関（ローカライズドセンタリング前）を示す散布図である。図１２はローカライズドセンタリングを適用した後の$Ｎ_{10}$の分布を示す図である。

図１１が示すように、$Ｎ_{10}$とローカルアフィニティ（すなわち、ローカルセントロイドへの類似度）との間に強い相関が観察された。これは図６で示した$Ｎ_{10}$とグローバル

アフィニティ（すなわちグローバルセントロイドへの類似度）との間の弱い相関と対照的である。そして、図１２が示すように、ローカライズドセンタリング後にN_{10}の分布の歪度が小さくなること（ハブが軽減されること）が観察された。

【００３６】

図１３は、図１－３で用いた高次元データセットについて、N_{10}と、ローカルアフィニティの相関を示す散布図である。図１３を見ると、高次元データセット（ｄ＝４０００）においてもN_{10}はローカルアフィニティと強く相関する。これらの結果は、ローカルアフィニティは、大規模データセットだけでなく高次元データセットにおいても、ハブを軽減するための指標となり得ることを示す。

【００３７】

〔提案方法：ローカライズドセンタリング〕

以上の所見から、ローカルアフィニティはハブを軽減するための指標となり得る。そこで、ローカルアフィニティに基づくハブ軽減の新たな方法「ローカライズドセンタリング」を提案する。それは、高次元データセットと大規模データセットの両方に効果がある。

ローカライズドセンタリングを、グローバルセンタリングによる式（１）とのアナロジーで説明する。データセットＤに属するデータｘとクエリｑとの類似度を計算する上で、ｘに依存しない項は、異なるｘに対して一定である。よって、式（１）から、次式が導かれる。

$$Sim^{CENT}(x;q) \equiv \langle x-c, q-c\rangle = \langle x, q\rangle - \langle x, c\rangle + constant \quad (4)$$

言い換えれば、センタリング法では、もともとの類似度スコア＜ｘ，ｑ＞から、グローバルアフィニティ＜ｘ，ｃ＞が、ペナルティ項として差し引かれる。同様にローカライズドセンタリングと称する提案方法では、もともとの類似度スコアからｘのローカルアフィニティをペナルティ項として差し引く。

$$Sim^{LCENT}(x;q) \equiv \langle x, q\rangle - \langle x, c_\kappa(x)\rangle \quad (5)$$

【００３８】

式（５）はローカルセグメントサイズを定めるパラメタκを含む。パラメタκは、例えば、N_k（ｘ）とローカルアフィニティ＜ｘ，c_κ（ｘ）＞との相関が最大になるように定める。もしκ＝｜Ｄ｜であれば、提案方法はグローバルセンタリングと同一になる。式（５）への変換後、任意のデータｘ∈Ｄはそのローカルセントロイドc_κ（ｘ）と一定の類似度（実際はゼロ）を持つ。その理由は、式（５）に$q = c_\kappa(x)$を代入すると、次式が導かれるからである。

$$Sim^{LCENT}(x;c_\kappa(x)) = 0$$

言い換えれば、変換後にはローカルセントロイドとの類似度が特別高くなるデータは存在しない。ローカルセントロイドと高い類似度を持つデータがハブになるという、前記の所見を考慮すれば、この変換がハブを軽減することが期待される。実際、この期待は図１２のデータセットに対しては成立している。

【００３９】

ローカライズドセンタリングはデータ数ｎと次元数ｄを変えた他のデータセットでもハブを軽減する。図９にローカライズドセンタリング後のN_{10}分布の歪度の等高線図を示す。図より、ローカライズドセンタリングは両タイプのハブ、すなわち、大規模データセットに生じるハブ（等高線図の右下の領域に対応する）と、高次元データセットに生じるハブ（左上の領域に対応する）の、どちらのハブも軽減できることが分かる。高次元データセットに対しても、ローカライズドセンタリング（式（５））はグローバルセンタリング（式（４））と同様に効果的である。このようなデータセットでは、図１０の左上領域に示したように、グローバル対ローカルアフィニティ比は１に近いからである。このことは、この領域のデータセットの殆どのデータについて、＜ｘ，c_κ（ｘ）＞　＜ｘ，ｃ＞を意味し、したがって、式（５）は式（４）と殆ど同じになる。

式（５）はさらに拡張できる。式（５）の右辺の第２項はペナルティ項として説明できる、すなわち、ペナルティ項はｘがどの程度のハブであるかを表し、それに応じて類似度

スコアを小さくするために使用される。式（５）の拡張は、たとえば式（６）あるいは（７）のように、ペナルティの程度をコントロールするためのパラメタγの導入により行う。

$$Sim_{\gamma}^{LCENT}(x;q) \equiv <x, q> - <x, c_{\kappa}(x)>^{\gamma} \quad (6)$$

$$Sim_{\gamma}^{LCENT}(x;q) \equiv <x, q> - \gamma <x, c_{\kappa}(x)> \quad (7)$$

パラメタγは、N_k分布の歪度がもっとも小さくなるように選択すればよい。

【実施例１】

【００４０】

〔システム構成〕

図１４に本実施例における類似度演算システム１（情報処理装置の一例）のシステム構成を例示する。類似度演算システム１は、データセット・クエリ設定部１１、ベース類似度スコア演算部１２、ハブ度演算部１３、ペナルティスコア演算部１４、総合スコア演算部１５、演算結果整理・分類部１６、制御部１７、表示部１８、入出力部１９、記憶部２０を備える。

データセット・クエリ設定部１１は検索対象データ（クエリ）とデータセットの設定を行う。本実施例で扱うデータセットは、例えば、テキスト文書データセットや、ＤＮＡの塩基配列あるいはタンパク質のアミノ酸配列のデータセット等で、高次元又は／及び大規模データを有する。操作者が例えば、コンピュータ１０の表示部１８の設定画面のクエリ入力欄とデータセット入力欄にクエリとデータセット名を入出力部１９のマウスやキーボードを用いて入力し、類似度演算システム１が類似度演算に用いるクエリとデータセットであると認識するように、検索対象データ（クエリ）とデータセットを設定する。

ベース類似度スコア演算部１２（第１算出部の一例）は、データセット・クエリ設定部１１にて設定されたクエリと設定されたデータセット内の各データについて、ベースとする類似度尺度を用いてクエリとの類似度を演算する。ベースとする類似度尺度として、内積、コサイン、距離等を使用できる。内積、コサイン、距離と異なり、データが明にベクトル表現されていることを前提としない類似度尺度であっても使用できる。例えば、塩基配列間の類似度尺度であるＢＬＡＳＴスコアや、文字列間の類似度尺度となるストリングカーネル、グラフ間の類似度尺度となるグラフカーネルなどを、ベースとする類似度尺度として使用できる。

【００４１】

ハブ度演算部１３はハブ度として出現回数及び歪度を演算する。すなわち、データセット内の各データｘについて、他データのｋ近傍（ベースとする類似度尺度により決定される）への出現回数$N_k(x)$を演算する出現回数演算部１３１、出現回数の分布の歪度を演算する歪度演算部１３２を有する。分布の歪度は、$S_{Nk}=E[(N_k-\mu_{Nk})^3/\sigma_{Nk}^3]$（$E[\ ]$は期待値を計算するオペレータ、$\mu_{Nk}$と$\sigma_{Nk}$はそれぞれＮｋ分布の平均と分散である)で表される。

【００４２】

ペナルティスコア演算部１４（特定部、第２算出部の一例）は、各データｘのローカルセントロイド$c_{\kappa}(x)=(1/\kappa)\Sigma_{x'\in\kappa NN(x)}x'$の計算に用いる隣接データの範囲であるローカルセグメントサイズを定めるパラメタκを決定するローカルセグメントサイズ決定部１４１、各データｘのローカルセントロイドを計算するローカルセントロイド演算部１４２を有する。ローカルセグメントサイズは、例えば、ローカルアフィニティ、すなわち、ローカルセントロイドとの類似度$<x, c_{\kappa}(x)>$と、他データのｋ近傍に含まれる回数$N_k(x)$との相関が最大となるようにパラメタκを定める。隣接データの範囲とは出現回数$N_k(x)$の大きいデータの近傍データが局在化されている範囲で、ローカルセグメントサイズ内をいう。なお、κが固定されていれば、ローカルセントロイドは各データに対して１つだけ存在する。なお、ローカルセントロイドは各データにより異なるが必ず存在するとみなしてよい。したがって、ローカルセントロイドの有無を判断する必要はない。このように、ペナルティスコア演算部１４は、各データｘごとに隣接データを特定し、各データの基準座標ともいえるローカルセントロイドを算出する。

ペナルティスコア演算部１４は、さらに、各データｘに対し、ローカルセントロイド演算部１４２で演算されたローカルセントロイドc_{κ}（ｘ）を用い、ローカルアフィニティ、すなわち、ローカルセントロイドとの類似度＜ｘ，c_{κ}（ｘ）＞を、各データｘのペナルティスコアとして演算する、ローカルアフィニティ演算部１４３を有する。ローカルアフィニティの演算はローカライズドセンタリングの前工程を形成する。

総合スコア演算部１５（第３算出部の一例）は、各データｘに対し、ベース類似度スコア演算部１２においてベースとする類似度尺度を用いて演算されたクエリとの類似度から、ローカルアフィニティ演算部１４３で演算されたペナルティスコアを差し引きすることにより、総合スコアを演算する。このように、総合スコア演算部１５は、ベース類似度スコア演算部１２が演算する類似度と、ペナルティスコア演算部１４が算出するローカルセントロイドとに基づいて、各データｘごとに総合スコアを算出する。総合スコアの演算はローカライズドセンタリングの後工程を形成する。

【００４３】

演算結果整理・分類部１６（抽出部の一例）は、各データｘに対して総合スコア演算部１５で計算した総合スコアを、大きい順に整理する総合スコア順整理部１６１と、この整理結果を用いてｋ近傍法により分類する分類部１６２を有する。分類は、総合スコアが大きいｋ個のデータの分類ラベルを用いた（スコア重み付き）多数決により行う。つまり、分類部１６２は、クエリについて、総合スコアが大きい順にデータセットからｋ個のデータを抽出し、分類を行う。これにより、類似度演算システム１は、クエリに対応する分類ラベルを特定することができる。

【００４４】

制御部１７は類似度演算システム１及びその各部１１～１６を制御して、類似度演算システム１としての機能を発揮させる。

以上のデータセット・クエリ設定部１１から制御部１７はコンピュータ１０内に実現可能であり、本実施例でもそのようにしている。ただし、ベース類似度スコア演算部１２では、ＢＬＡＳＴのような汎用のアルゴリズムが提供する類似度演算を利用しても良い。この場合でも、ベース類似度スコア演算部１２は演算結果を取り込み、記憶部２０に記憶させる機能を保有する。

表示部１８は、類似度演算システム１各部１１～１６の入力結果、演算結果、記憶部２０の記憶内容、入出力操作時の案内画面等を表示する。入出力部１９は、キーボード、マウス、マイクロフォン等の入力機器、プリンタ、スピーカー等の出力機器を有し、操作者はこれら入出力機器を用いて適宜入出力できる。

【００４５】

記憶部２０は、類似度演算システム１各部１１～１６の入力結果、演算結果、演算途中の情報、類似度演算システム１を動作させるのに必要なプログラム及び情報を記憶する。記憶部２０は、データセット・クエリ情報記憶部２１、ベース類似度スコア情報記憶部２２、出現回数・歪度情報記憶部２３、ローカルセグメントサイズ情報記憶部２４、ローカルセントロイド情報記憶部２５、ペナルティスコア情報記憶部２６、総合スコア情報記憶部２７、分類情報記憶部２８、プログラム記憶部２９及び一時記憶部３０を有する。

【００４６】

データセット・クエリ情報記憶部２１は、検索先データセット、検索対象データ（クエリ）に関して、データセット・クエリ設定部１１による設定結果を記憶する。ベース類似度スコア情報記憶部２２は、ベース類似度として用いる類似度の計算アルゴリズム（内積、コサイン、距離等の計算アルゴリズムや、ＢＬＡＳＴのような汎用の類似度計算アルゴリズムを含む）や、ベース類似度スコア演算部１２による類似度演算結果等を記憶する。出現回数・歪度情報記憶部２３は、ハブ度演算部１３での演算結果、すなわち、出現回数演算部１３１により演算された各データが他のデータのｋ近傍に出現する回数（各データのハブ度）、及び、歪度演算部１３２により演算された出現回数分布の歪度（データセットのハブ度）を記憶する。

【００４７】

ローカルセグメントサイズ情報記憶部２４は、ローカルセグメントサイズ決定部１４１で決定されたローカルセグメントサイズκを、ローカルセントロイド情報記憶部２５は、ローカルセントロイド演算部１４２で演算された各データのローカルセントロイドに関する情報を記憶する。ペナルティスコア情報記憶部２６は、ローカルアフィニティ演算部１４３で演算された各データのペナルティスコアを記憶する。総合スコア情報記憶部２７は、総合スコア順整理部１６１による総合スコア順の整理結果等を記憶する。分類情報記憶部２８は、分類部１６２によるｋ近傍法を用いたクエリの分類結果を記憶する。プログラム記憶部２９は、類似度演算システム１及びその各部１１～１６のデータの流れ、演算、分類を実行させるためのプログラム、類似度演算システム１及びその各部１１～１６を制御するために必要なプログラム、類似度演算システム１の起動に必要なプログラム等を記憶する。一時記憶部３０は、演算途中の情報等を一時的に記憶する。

【００４８】

〔処理フロー〕

図１５に本実施例における類似度演算の処理フロー例を示す。まず、データセット・クエリ設定部１１にて高次元又は／及び大規模データについて、検索対象クエリと検索先データセットを設定する（データセット・クエリ設定工程：Ｓ１０１）。例えば、生命情報学分野における配列の相同性検索では、世界中の研究成果として蓄積された機能既知のＤＮＡの塩基配列あるいはタンパク質のアミノ酸配列のデータセット、及び、検索対象クエリとなる機能未知の塩基配列やアミノ酸配列が設定され、データセット・クエリ情報記憶部２１に記憶される。次に、データセット・クエリ設定工程（Ｓ１０１）にて設定されたクエリとデータセット内の各データについて、ベース類似度スコア演算部１２にてベースとする類似度尺度を用いてクエリとデータの類似度を演算する（ベース類似度スコア演算工程：Ｓ１０２）。ベースとする類似度尺度として、生命情報学分野における配列の相同性検索では、ＢＬＡＳＴアルゴリズムが出力するスコアが用いられるが、一般には内積、距離、コサイン等が使用される。演算されたベース類似度スコアはベース類似度スコア情報記憶部２２に記憶される。

【００４９】

次に、ハブ度演算部１３にて、各データのハブ度、及び、データセットのハブ度を演算する（ハブ度演算工程：Ｓ１０３）。すなわち、出現回数演算部１３１では、各データｘのハブ度を、データセット内の他のデータのｋ近傍への出現回数N_k（ｘ）を演算する。歪度演算部１３２では、N_k（ｘ）の分布の歪度を演算する。出現回数演算部１３１及び歪度演算部１３２により作成された出現回数分布図及び歪度（図１、２、４、５、１２参照）が参照される。ハブが存在すれば、出現回数分布図に大きいN_k（ｘ）を持つデータが現われ、大きい歪度が観測される。出現回数N_k（ｘ）及び歪度は、出現回数・歪度情報記憶部２３に記憶される。

【００５０】

次に、ローカルセグメントサイズ決定部１４１にて、ローカルセグメントサイズκを決定し（ローカルセグメントサイズ決定工程：Ｓ１０４）、κの値はローカルセグメントサイズ情報記憶部２４に記憶する。そして、決定されたκを用い、ローカルセントロイド演算部１４２にて、各データｘのローカルセントロイドc_κ（ｘ）を演算し（ローカルセントロイド演算工程：Ｓ１０５）、c_κ（ｘ）はローカルセントロイド情報記憶部２５に記憶する。さらに、演算されたc_κ（ｘ）を用い、ローカルアフィニティ演算部１４３にて、各データｘのローカルアフィニティをc_κ（ｘ）とｘの類似度として計算し（ペナルティスコア演算工程：Ｓ１０６）、これを各データのペナルティスコアとしてペナルティスコア情報記憶部２６に記憶する。ローカルアフィニティの演算はローカライズドセンタリングの前工程を形成する。

【００５１】

次に、総合スコア演算部１５にて、ベース類似度スコア情報記憶部２２及びペナルティスコア情報記憶部２６に記憶された各データに対するベース類似度スコア及びペナルティスコアを用い、各データに対する総合スコアを演算する（総合スコア演算工程：Ｓ１０７

）。ここでは、ベース類似度スコアからペナルティスコアを差し引くことにより、総合スコアを得る。総合スコアの演算はローカライズドセンタリングの後工程を形成する。ローカライズドセンタリングは前工程と後工程からなる。データセットのデータは、総合スコア演算工程（Ｓ１０７）での演算結果に基づいて、演算結果整理・分類部１６にて整理・分類される（演算結果整理工程：Ｓ１０８）。典型的には、総合スコアの大きい順に整理（ランク付け）され、総合スコア情報記憶部２７にランキングが記憶される。さらに、検索対象クエリを分類する場合には、ｋ近傍法を用いて（総合スコアが大きいｋ個のデータの分類ラベルの投票により）検索対象クエリの分類ラベルの予測を行う（分類工程：Ｓ１０９）。予測結果は、分類情報記憶部２８に記憶される。

【００５２】

〔実世界データを用いた実験〕

本発明による手法（ローカライズドセンタリング）は、（人工データではなく）実世界データセットに対してハブを軽減する効果を持つかどうか、さらに、ハブを軽減する結果としてタスクの精度を向上するかどうかを調べるために、多クラス分類タスクのベンチマークとして用いられる実世界のテキスト文書データセットに対してローカライズドセンタリングを適用する。使用するテキスト文書データセットは、図１６はＷｅｂＫＢ、図１７はＲｅｕｔｅｒｓ－５２、図１８はＴＤＴ２－３０、図１９は２０Ｎｅｗｓｇｒｏｕｐｓである。実験を通して、テキスト文書の分類タスクにおいて標準的なデータ表現であるｔｆ－ｉｄｆ重み付きｂａｇ－ｏｆ－ｗｏｒｄｓベクトル表現を用い、標準的な類似度尺度（ベースライン）としてコサインを用いる。

【００５３】

〔ハブの軽減〕

最初の実験では、ローカライズドセンタリングが大規模データセットに生じるハブを軽減するかどうかを調べた。データセット全体からｎ個のデータをランダムに選択し、データ数ｎのサブセットを生成した。データ数ｎを１００から４０００の範囲で動かし、各ｎについてサブセットの生成を１０回行った。各サブセットに対してN_{10}分布の歪度を計算し、各ｎについて平均歪度（１０回の平均）を計算した。そして、グローバルセンタリング（センタリング）後及びローカライズドセンタリング後の平均歪度を比較した。

【００５４】

図１６－１９は、データ数ｎとN_{10}分布の平均歪度との関係を示す図である。（ｉ）はグローバルセンタリング又はローカライズドセンタリングを施す前の折れ線グラフ（比較のベースライン）である。（ｉｉ）はグローバルセンタリング後、（ｉｉｉ）はローカライズドセンタリングの折れ線グラフである。使用した４つのデータセットを通して、同じ傾向が観察された。すなわち、データ数ｎの増加につれて、ベースラインではN_{10}分布の平均歪度は増加し、ハブが増大する。一方、グローバルセンタリングによって、平均歪度はやはり増加するが、平均歪度はベースラインよりやや小さい。他方、ローカライズドセンタリングでは、データ数ｎに関わらず、平均歪度はほぼ同じ小さい値のままで変化しない。まとめると、グローバルセンタリングはデータ数ｎが小さい（約５００より小さい）ときに効果が見られるが、データ数ｎが大きいときには否である。他方、ローカライズドセンタリングは、データ数ｎが大きくてもハブの発生を低いレベルに維持する。

さらに、式（２）及び式（３）を用い、各データセットについてグローバル対ローカルアフィニティ比を計算し、図１６－１９にその折れ線グラフを（ｉｖ）として示した。すると、人工データセットで観察されたのと同じ傾向が明確に観察された。すなわち、グローバル対ローカルアフィニティ比が小さいとき、グローバルセンタリングはハブを軽減しないが、ローカライズドセンタリングはハブを軽減する。

【００５５】

〔ｋ近傍分類の精度〕

このように、ローカライズドセンタリングによりハブが軽減することは確認がされたが、次の実験では、このハブ軽減がｋ近傍分類精度を改善するかどうかを調べた。タスクはテストデータとしてのテキスト文書の分類ラベルを訓練データとしてのテキスト文書の分

類ラベルを用いて予測することである。リーブワンナウトｌｅａｖｅ－ｏｎｅ－ｏｕｔ交差検証法（全データから一つをテストデータとして、残りを訓練データとして用い、訓練データを使ってテストデータの分類ラベルを予測することを、全データに亘ってデータ数だけ繰り返し、予測の正解率を評価値とする）によりパフォーマンスを評価した。

【００５６】

ベースラインとなるコサイン（ＣＯＳ）、及び、コサインを変形した５つの類似度尺度を試した。グローバルセンタリング（標準的なセンタリング）（ＣＥＮＴ）、ローカライズドセンタリング（ＬＣＥＮＴ）、コミュートタイムカーネル（ＣＴ）、ミューチュアルプロキシミティ（ＭＰ）及ローカルスケーリング（ＬＳ）である。コミュートタイムカーネルは、グラフラプラシアンに基づくグラフノード間の類似度を定めるためにＳａｅｒｅｎｓらにより発表され（Ｓａｅｒｅｎｓ，Ｍ．；Ｆｏｕｓｓ，Ｆ．；Ｙｅｎ，Ｌ．；ａｎｄ Ｄｕｐｏｎｔ，Ｐ．２００４．“Ｔｈｅ ｐｒｉｎｃｉｐａｌ ｃｏｍｐｏｎｅｎｔｓ ａｎａｌｙｓｉｓ ｏｆ ｇｒａｐｈ，ａｎｄ ｉｔｓ ｒｅｌａｔｉｏｎｓｈｉｐｓ ｔｏ ｓｐｅｃｔｒａｌ ｃｌｕｓｔｅｒｉｎｇ．” Ｉｎ Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ １５ｔｈ Ｅｕｒｏｐｅａｎ Ｃｏｎｆｅｒｅｎｃｅ ｏｎ Ｍａｃｈｉｎｅ Ｌｅａｒｎｉｎｇ（ＥＣＭＬ’０４），３７１－３８３．２００４年）、後に発明者らによりハブ軽減に効果があることが示された（Ｓｕｚｕｋｉ，Ｉ．；Ｈａｒａ，Ｋ．；Ｓｈｉｍｂｏ，Ｍ．；Ｍａｔｓｕｍｏｔｏ，Ｙ．；ａｎｄ Ｓａｅｒｅｎｓ，Ｍ．２０１２．“Ｉｎｖｅｓｔｉｇａｔｉｎｇ ｔｈｅ ｅｆｆｅｃｔｉｖｅｎｅｓｓ ｏｆ Ｌａｐｌａｃｉａｎ－ｂａｓｅｄ ｋｅｒｎｅｌｓ ｉｎ ｈｕｂ ｒｅｄｕｃｔｉｏｎ．” Ｉｎ Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ Ｔｗｅｎｔｙ－Ｓｉｘｔｈ ＡＡＡＩ Ｃｏｎｆｅｒｅｎｃｅ ｏｎ Ａｒｔｉｆｉｃｉａｌ Ｉｎｔｅｌｌｉｇｅｎｃｅ．２０１２年）。ミューチュアルプロキシミティ（Ｓｃｈｎｉｔｚｅｒ，Ｄ．；Ｆｌｅｘｅｒ，Ａ．；Ｓｃｈｅｄｌ，Ｍ．；ａｎｄ Ｗｉｄｍｅｒ，Ｇ．２０１２．“Ｌｏｃａｌ ａｎｄ ｇｌｏｂａｌ ｓｃａｌｉｎｇ ｒｅｄｕｃｅ ｈｕｂｓ ｉｎ ｓｐａｃｅ．” Ｊｏｕｒｎａｌ ｏｆ Ｍａｃｈｉｎｅ Ｌｅａｒｎｉｎｇ Ｒｅｓｅａｒｃｈ １３（１）：２８７１－２９０２．２０１２年）とローカルスケーリング（Ｚｅｌｎｉｋ－Ｍａｎｏｒ，Ｌ．ａｎｄ Ｐｅｒｏｎａ，Ｐ．２００５．“Ｓｅｌｆ－ｔｕｎｉｎｇ ｓｐｅｃｔｒａｌ ｃｌｕｓｔｅｒｉｎｇ．” Ｉｎ Ａｄｖａｎｃｅｓ ｉｎ Ｎｅｕｒａｌ Ｉｎｆｏｒｍａｔｉｏｎ Ｐｒｏｃｅｓｓｉｎｇ Ｓｙｓｔｅｍｓ １７．ＭＩＴ Ｐｒｅｓｓ．１６０１－１６０８．２００５年）はデータ間の近傍関係を対称化しようと試みるものである。

【００５７】

図２０－２３に実験結果を示した。ＬＣＥＮＴあるいはＬＳが精度の項目において最良のパフォーマンスを達成した。ＣＥＮＴはコサインに対して歪度と精度をわずかに改善するのみに留まったが、対照的にＬＣＥＮＴはコサインに対して歪度と精度を大きく改善した。

【００５８】

以上により、本実施例によれば、次元数が大きいときだけでなく、データ数が大きいときに出現するハブを軽減できる類似度演算システム及び類似度演算方法を提供できる。

【実施例２】

【００５９】

本実施例では、同一データセットについて検索を行う場合、ペナルティスコア情報記憶部２６に記憶されているペナルティスコアを用いることにより、検索処理を早められる。あるいは、前処理としてオフラインに各データに対するペナルティスコアを計算しておき、ペナルティスコア情報記憶部２６に記憶しておくことで、検索処理を早められる。その他のシステム構成及び処理フローは実施例１と同様であり、実施例１と同様に次元数が大きいときだけでなく、データ数が大きいときに出現するハブを軽減できる類似度演算システム及び類似度演算方法を提供できるという効果を奏する。

【実施例３】

【００６０】
本実施例では、実施例１で説明した総合スコアを示す式（６）、（７）のペナルティスコアを調整する例について説明する。

$$Sim_{\gamma}^{LCENT}(x;q) \equiv <x, q> - <x, c_{\kappa}(x)>^{\gamma} \quad (6)$$

$$Sim_{\gamma}^{LCENT}(x;q) \equiv <x, q> - \gamma <x, c_{\kappa}(x)> \quad (7)$$

式（６）、（７）の右側の第２項はペナルティスコアである。パラメタγはN_k分布の歪度を最も軽減するように調整され得る。その他のシステム構成及び処理フローは実施例１と同様であり、実施例１と同様に次元数が大きいときだけでなく、データ数が大きいときに出現するハブを軽減できる類似度演算システム及び類似度演算方法を提供できるという効果を奏する。

【実施例４】
【００６１】
本実施例では、ベースとする類似度尺度として、これまで説明に用いてきた内積ではなく，ユークリッド距離（l^2ノルム）を用いた場合の、ハブ軽減方法について説明する。ここでは、クエリｑと各データｘとの類似度Ｓｉｍ（ｘ；ｑ）をユークリッド距離（ｓｑｒｔ（||ｘ－ｑ||^2））の２乗にマイナス符号を付した値として定義する。すなわち、

$$Sim(x;q) \equiv -||x-q||^2 \quad (8)$$

である。
これを類似度尺度として用いて発生するハブは、ローカルセントロイド$c_{\kappa}(x)$を用い、次のように類似度スコアを補正することによって、軽減できる。

$$Sim^{LCENT}(x;q) \equiv -||x-q||^2 + ||x-c_{\kappa}(x)||^2 \quad (9)$$

さらに、類似度スコアを調整する右辺第２項に、次のようにパラメタγを付加してもよい。

$$Sim_{\gamma}^{LCENT}(x;q) \equiv -||x-q||^2 + (||x-c_{\kappa}(x)||^2)^{\gamma} \quad (10)$$

$$Sim_{\gamma}^{LCENT}(x;q) \equiv -||x-q||^2 + \gamma||x-c_{\kappa}(x)||^2 \quad (11)$$

【実施例５】
【００６２】
本実施例では、各データのベクトル表現が与えられず、機械学習分野で呼ばれるところのカーネルやＢＬＡＳＴスコアなどのようにデータ間の類似度だけが、ベースとする類似度尺度として与えられる場合の、ハブ軽減方法について説明する。まず、クエリｑと各データｘとの類似度が、ＢＬＡＳＴスコアにより定義されているとする。すなわち、

$$Sim(x;q) \equiv BLASTスコア(x;q) \quad (12)$$

とする。ここで、ＢＬＡＳＴスコアを類似度尺度として用いて発生するハブは、次のようにスコアを補正することによって、軽減できる。

$$Sim^{LCENT}(x;q) \equiv BLASTスコア(x;q) - (1/\kappa)\Sigma_{x' \in \kappa NN(x)} x' BLASTスコア(x;x') \quad (13)$$

上記ではＢＬＡＳＴスコアを例に用いて説明したが、ベースとする類似度尺度としてその他の類似度（カーネルなど）を用いる場合は、上記ＢＬＡＳＴスコアの部分を、用いる類似度（カーネルなど）が定めるスコアに置き換えればよい。すなわち、

$$Sim^{LCENT}(x;q) \equiv Sim(x;q) - (1/\kappa)\Sigma_{x' \in \kappa NN(x)} x' Sim(x;x') \quad (14)$$

とすれば、任意の類似度尺度Ｓｉｍ（ｘ；ｑ）に対して、ハブを軽減できる。ここで、κ＝ｎ＝|Ｄ|でも良く、さらに、実施例３、４と同様に、ペナルティスコアを調整するためのパラメタγを導入しても良い。

【実施例６】

【００６３】

本実施例では、ローカルセグメントサイズκの値を、データセット内で一定としない場合について説明する。例えば、ローカルセグメントサイズκの値は、データセット内のデータｘごとに異なっていてよい。この場合、κの値は、データｘの近傍におけるデータの密集度に基づいて定めてもよい。具体的には、データｘのハブ度、すなわち、データセット内の他のデータのｋ近傍への出現回数N_k（ｘ）が大きくなる程、κの値も大きくなるようにしてもよい。これにより、データセット内に複数のデータのクラスター（局所集合）が存在し、各クラスターを構成するデータ数が大きく異なる場合であっても、各データｘに対応するローカルセントロイドc_κ（ｘ）を、データｘが属するクラスターの中心付近に設定することができる。したがって、類似度演算システム１は、ハブの発生を抑制し、ｋ近傍法の精度を向上させることができる。

【００６４】

次に、変形例に係る類似度演算システム１について説明する。

上述した実施例では、一例として、類似度演算システム１がローカルセントロイドc_κ（ｘ）に基づいて、ペナルティスコアを算出する場合について説明したがこれには限られない。類似度演算システム１は、ローカルセントロイドc_κ（ｘ）に基づいてボーナススコアを算出してもよい。類似度尺度として内積を用いる場合、内積は大きければ大きい程、２つのデータが類似することを表す。したがって、内積を用いる場合には、ローカルアフィニティＬｏｃａｌＡｆｆｉｎｉｔｙ（ｘ）はペナルティとして算出される。他方、類似度尺度として、距離を用いる場合、距離は小さければ小さい程、２つのデータが類似することを表す。したがって、距離を用いる場合には、ローカルアフィニティＬｏｃａｌＡｆｆｉｎｉｔｙ（ｘ）はボーナスとして算出される。このように、ローカルアフィニティＬｏｃａｌＡｆｆｉｎｉｔｙ（ｘ）に係るペナルティ、ボーナスの概念は、類似度尺度等に応じて異なる。つまり、ローカルアフィニティＬｏｃａｌＡｆｆｉｎｉｔｙ（ｘ）とは、スコアの補正に用いられるパラメタであり、その物理的な意味付けは、システムの構成に応じて異なるパラメタである。

【００６５】

また、上述した実施例では、一例として、データセット、クエリを構成するデータとして塩基配列データやアミノ酸配列データを用いる場合について説明したが、これには限られない。例えば、上述した配列データの他にも、マイクロアレイデータ等の生命科学分野のデータが用いられてもよい。また、生命科学分野に限られず、音声データ、画像データ、テキストデータ等の任意の種類のデータが用いられてよい。

【００６６】

なお、ローカルセグメントサイズκの値は、任意の方法で設定されてよい。例えば、塩基配列データやアミノ酸配列データは、データ空間において、比較的、データ数の小さなクラスターが大量に発生し、これらがハブ発生の原因となる場合がある。そこで、この場合には、ローカルセグメントサイズκの値は、例えば、１０以下の比較的小さな値としてもよい。このように、データ空間においてデータ数の小さなクラスターが大量に発生している場合には、ローカルセグメントサイズκとして、相対的に小さな値を設定する。これにより、類似度演算システム１は、演算量を低減しつつ、ハブを軽減することができる。これに対して、データ空間においてデータ数の大きなクラスターが多く発生している場合には、ローカルセグメントサイズκの値を、相対的に大きな値としてもよい。これにより、データｘがクラスターの外縁部に位置していたとしても、データｘごとのローカルセントロイドc_κ（ｘ）をクラスターの中心に近づけることができるため、効率よくハブを軽減することができる。また、ローカルセグメントサイズκの値は、ｋの値に基づいて定められてもよい。例えば、ローカルセグメントサイズκの値は、ｋの値の２倍程度としてもよい。また、ローカルセグメントサイズκの値は、システムごとに予め定められていてもよい。また、ローカルセグメントサイズκの値は、データｘのハブ度、すなわち、データセット内の他のデータのｋ近傍への出現回数N_k（ｘ）に関する分布の歪度が最小になる

数としてもよい。

【００６７】

また、上述した実施例では、一例として、ｋ近傍法を用いる場合について説明したが、これには限られない。例えば、ｋ近傍法を用いる代わりにイプシロン近傍法を用いてもよい。この場合、隣接データの範囲は、データセット内のデータのうち、データｘとの類似度が所定の値よりも高いデータとしてもよい。

【００６８】

また、本発明の一態様は、以上の実施例のフローチャート等に記載の類似度演算方法を含むコンピュータに実行させるためのプログラムとしても実現可能である。プログラムは情報検索装置の内蔵記憶部に蓄積して使用してもよく、外付けの記憶装置に蓄積して使用してもよく、インターネットからダウンロードして使用しても良い。また、当該プログラムを記録した記録媒体としても実現可能である。

【００６９】

以上、本発明の一態様に係る実施の形態について説明したが、実施の形態は以上の例に限られるものではなく、本発明の趣旨を逸脱しない範囲で、種々の変更を加え得ることは明白である。例えば、上述の各実施例において説明した各構成は、特定の機能を発揮するのに不要である場合には、省略することができる。また、上述した各実施例において説明した各構成は、複数の装置に分散して備えられても良いし、１つの構成としてまとめられてもよい。

例えば、データセットがインターネット検索におけるＷｅｂ文書集合のように、データセットの次元数は明らかではないがデータ数が膨大であることが明白である場合、本発明の適用によりハブを軽減できる効果が認められるならば、本発明が提供する類似度演算システム及び類似度演算方法は有効である。また、以上の実施例では検索対象がクエリと同一である場合について説明したが、クエリが検索対象と関連性が深いキーワードであっても良い。また、データセットが膨大なときには、データセットが複数の記憶装置に分散して記憶されても良い。また、システム構成では、ローカルセントロイド情報記憶部２５とペナルティスコア情報記憶部２６をまとめて１つにしても良い。また、処理フローではデータセット・クエリ設定工程（Ｓ１０１）とベース類似度スコア演算工程（Ｓ１０２）とを同時に行っても良い。また、ｋ近傍法のパラメタｋは目的、状況に応じて適宜定めることができる。

【産業上の利用可能性】

【００７０】

本発明は類似度演算及び情報分類に利用される。

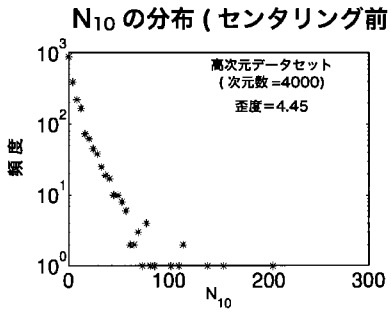
【符号の説明】

【００７１】

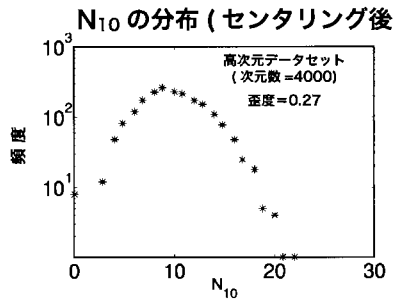
１　類似度演算システム
１０　コンピュータ
１１　データセット・クエリ設定部
１２　ベース類似度スコア演算部
１３　ハブ度演算部
１４　ペナルティスコア演算部
１５　総合スコア演算部
１６　演算結果整理・分類部
１７　制御部
１８　表示部
１９　入出力部
２０　記憶部
２１　データセット・クエリ情報記憶部
２２　ベース類似度スコア情報記憶部
２３　出現回数・歪度情報記憶部

２４　ローカルセグメントサイズ情報記憶部
２５　ローカルセントロイド情報記憶部
２６　ペナルティスコア情報記憶部
２７　総合スコア情報記憶部
２８　分類情報記憶部
２９　プログラム記憶部
３０　一時記憶部
１３１　出現回数演算部
１３２　歪度演算部
１４１　ローカルセグメントサイズ決定部
１４２　ローカルセントロイド演算部
１４３　ローカルアフィニティ演算部
１６１　総合スコア順整理部
１６２　分類部

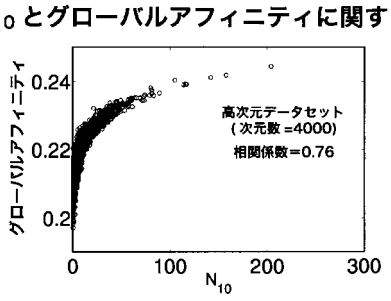
【図１】



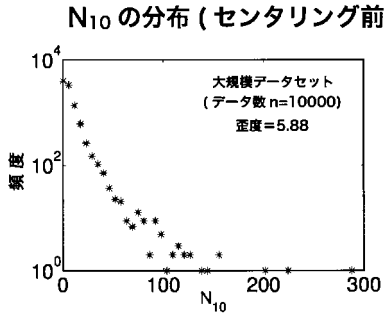
【図２】



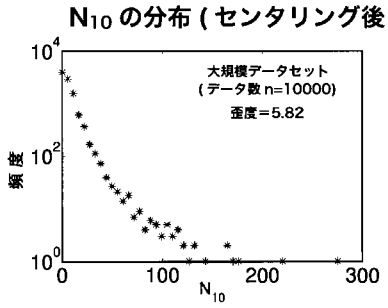
【図３】



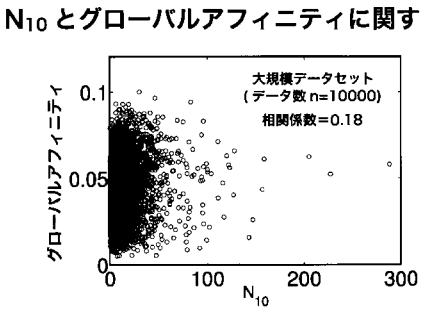
【図４】



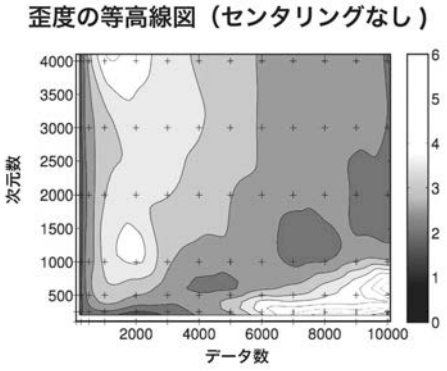
【図５】



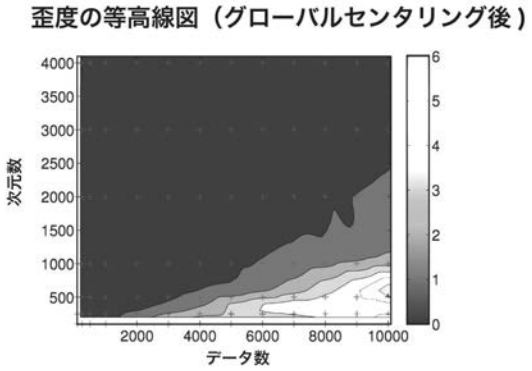
【図６】



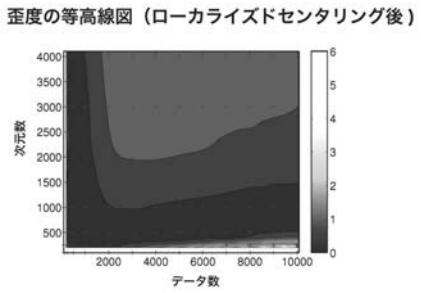
【図７】



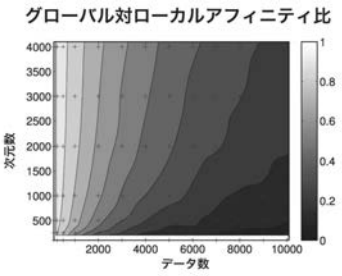
【図８】



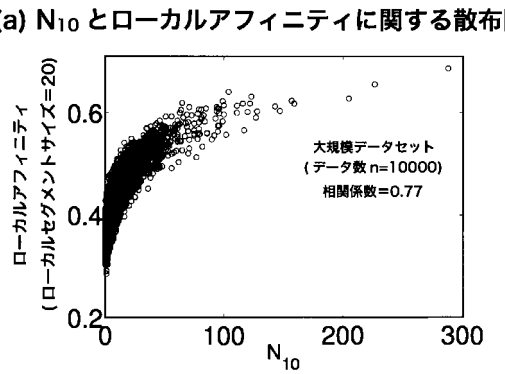
【図９】



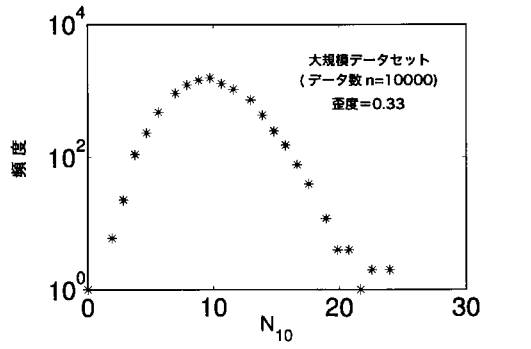
【図１０】



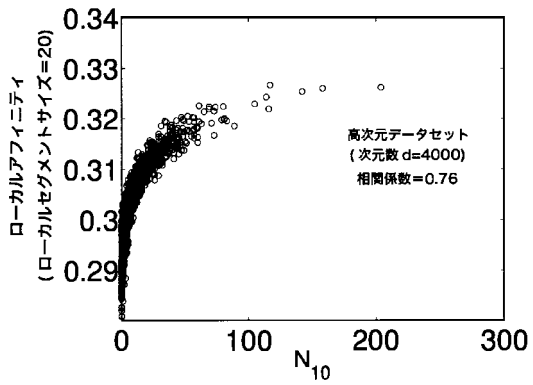
【図１１】



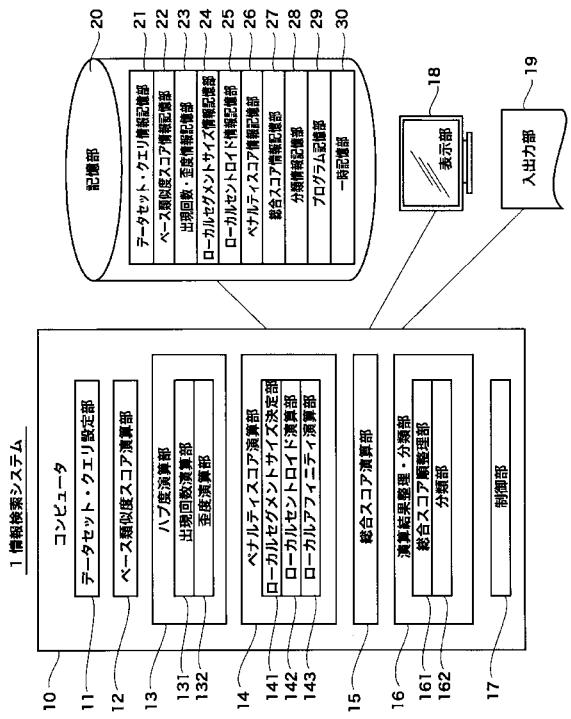
【図１２】



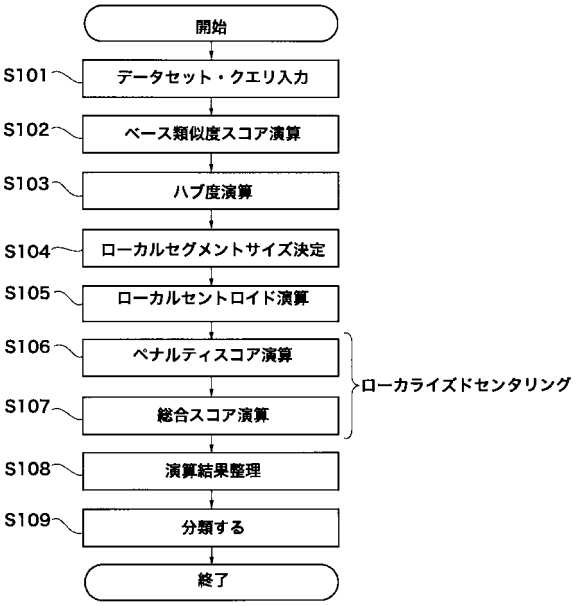
【図１３】



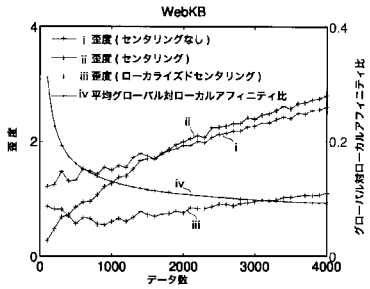
【図１４】



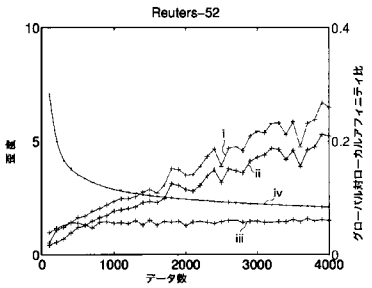
【図１５】



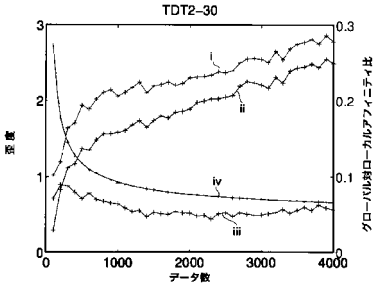
【図１６】



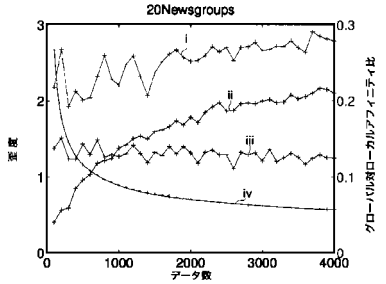
【図１７】



【図１８】



【図１９】



【図２０】

WebKB（データ数 4168, 分類クラス数 4）												
	k 近傍法分類の正解率						Nk 分布の歪度					
k	COS	CENT	LCENT	CT	MP	LS	COS	CENT	LCENT	CT	MP	LS
10	0.753	0.756	0.761	0.743	0.761	0.764	2.64	2.85	1.13	3.52	0.74	1.15
20	0.769	0.766	0.774	0.757	0.771	0.777	2.01	2.12	0.87	2.51	0.69	0.65
30	0.770	0.776	0.780	0.764	0.780	0.786	1.74	1.70	0.73	2.06	0.65	0.47
40	0.777	0.778	0.781	0.772	0.778	0.785	1.63	1.42	0.66	1.78	0.52	0.34
50	0.776	0.780	0.785	0.781	0.783	0.798	1.57	1.25	0.63	1.60	0.39	0.22

【図２１】

Reuters-52（データ数 9100, 分類クラス数 52）												
	k 近傍法分類の正解率						Nk 分布の歪度					
k	COS	CENT	LCENT	CT	MP	LS	COS	CENT	LCENT	CT	MP	LS
10	0.872	0.885	0.901	0.858	0.898	0.895	14.82	11.04	1.76	6.93	0.82	2.36
20	0.876	0.894	0.913	0.885	0.908	0.904	7.92	6.42	1.58	4.92	0.77	1.76
30	0.875	0.896	0.913	0.889	0.909	0.904	5.74	4.64	1.61	4.04	0.78	1.37
40	0.869	0.896	0.913	0.889	0.903	0.901	4.68	3.77	1.63	3.55	0.81	1.08
50	0.866	0.894	0.910	0.892	0.900	0.897	4.02	3.27	1.63	3.22	0.80	0.90

【図２２】

TDT2-30（データ数 9394, 分類クラス数 30）												
	k 近傍法分類の正解率						Nk 分布の歪度					
k	COS	CENT	LCENT	CT	MP	LS	COS	CENT	LCENT	CT	MP	LS
10	0.964	0.963	0.964	0.953	0.962	0.961	3.18	2.85	0.91	4.20	0.47	0.80
20	0.965	0.963	0.964	0.955	0.962	0.962	2.81	2.12	0.81	3.09	0.38	0.77
30	0.966	0.964	0.966	0.958	0.963	0.963	2.42	1.70	0.45	2.62	0.31	0.71
40	0.965	0.963	0.966	0.959	0.963	0.963	2.16	1.42	0.45	2.34	0.25	0.64
50	0.965	0.963	0.967	0.960	0.963	0.965	1.98	1.25	0.46	2.15	0.22	0.61

【図２３】

20Newsgroups（データ数 18,820, 分類クラス数 20）												
	k 近傍法分類の正解率						Nk 分布の歪度					
k	COS	CENT	LCENT	CT	MP	LS	COS	CENT	LCENT	CT	MP	LS
10	0.859	0.860	0.877	0.838	0.865	0.974	2.99	2.85	0.81	5.53	0.49	0.77
20	0.845	0.845	0.861	0.841	0.852	0.862	2.69	2.12	0.89	4.09	0.49	12.91
30	0.834	0.836	0.849	0.837	0.846	0.855	2.47	1.70	1.05	3.42	0.43	6.71
40	0.831	0.832	0.841	0.833	0.840	0.849	2.37	1.42	1.27	2.97	0.39	4.54
50	0.825	0.827	0.835	0.833	0.835	0.844	2.27	1.25	1.36	2.67	0.36	3.06

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2016/051909

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

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

B. FIELDS SEARCHED

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

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Jitsuyo Shinan Koho 1922-1996 Jitsuyo Shinan Toroku Koho 1996-2016
Kokai Jitsuyo Shinan Koho 1971-2016 Toroku Jitsuyo Shinan Koho 1994-2016

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.
A	Ikumi SUZUKI et al., "k Kinboho de Hub o Keigen suru Ruijido Shakudo", IPSJ SIG Notes 2012 (Heisei 24) Nendo (4), 04 January 2013 (04.01.2013) (received date), pages 1 to 8	1-10
A	Ikumi SUZUKI et al., "Effectiveness of Laplacian-based kernels from the hubness point of view", IEICE Technical Report, 02 November 2011 (02.11.2011), vol.111, no.275, pages 257 to 262	1-10
A	Kohei OZAKI et al., "Semi-supervised Word Sense Disambiguation using Degree Bounded Graph Construction", IPSJ SIG Notes Heisei 22 Nendo (4), 04 January 2011 (04.01.2011) (received date), pages 1 to 8	1-10

☐ 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	Date of mailing of the international search report
06 April 2016 (06.04.16)	19 April 2016 (19.04.16)
Name and mailing address of the ISA/ Japan Patent Office 3-4-3,Kasumigaseki,Chiyoda-ku, Tokyo 100-8915,Japan	Authorized officer Telephone No.

Form PCT/ISA/210 (second sheet) (January 2015)

国際調査報告	国際出願番号　ＰＣＴ／ＪＰ２０１６／０５１９０９

A.　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F17/30(2006.01)i

B.　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　G06F17/30

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０１６年
日本国実用新案登録公報	１９９６－２０１６年
日本国登録実用新案公報	１９９４－２０１６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
A	鈴木　郁美、外２名，ｋ近傍法でハブを軽減する類似度尺度，情報処理学会研究報告　２０１２（平成２４）年度（４），2013.01.04（受入日），ｐｐ．１－８	1-10
A	鈴木　郁美、外３名，「ハブの出現しやすさ」から見たラプラシアンベースカーネル，電子情報通信学会技術研究報告，2011.11.02，第１１１巻，第２７５号，ｐｐ．２５７－２６２	1-10

☒ 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	国際調査報告の発送日
０６．０４．２０１６	１９．０４．２０１６

国際調査機関の名称及びあて先	特許庁審査官（権限のある職員）	5M	４７７４
日本国特許庁（ＩＳＡ／ＪＰ） 郵便番号１００－８９１５ 東京都千代田区霞が関三丁目４番３号	齊藤　貴孝 電話番号　０３－３５８１－１１０１　内線　３５９９		

国際調査報告

国際出願番号 ＰＣＴ／ＪＰ２０１６／０５１９０９

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	小嵜 耕平、外３名，ハブを作らないグラフ構築法を用いた半教師あり語義曖昧性解消，情報処理学会研究報告 平成２２年度（４），2011.01.04（受入日），ｐｐ．１－８	1-10

様式ＰＣＴ／ＩＳＡ／２１０（第２ページの続き）（２０１５年１月）

フロントページの続き

(81)指定国 AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(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),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CZ,DE,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IR,IS,JP,KE,KG,KN,KP,KR,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,SM,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US

(72)発明者 原 一夫
静岡県三島市谷田１１１１番地 大学共同利用機関法人情報・システム研究機構 国立遺伝学研究所内
(72)発明者 鈴木 郁美
静岡県三島市谷田１１１１番地 大学共同利用機関法人情報・システム研究機構 国立遺伝学研究所内