

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第6183376号

(P6183376)

(45) 発行日 平成29年8月23日(2017.8.23)

(24) 登録日 平成29年8月4日(2017.8.4)

(51) Int.Cl.

F I

G 0 6 F 17/30 (2006.01)

G 0 6 F 17/30 4 1 4 A

請求項の数 21 (全 33 頁)

(21) 出願番号	特願2014-556333 (P2014-556333)	(73) 特許権者	000004237
(86) (22) 出願日	平成25年11月15日(2013.11.15)		日本電気株式会社
(86) 国際出願番号	PCT/JP2013/080873		東京都港区芝五丁目7番1号
(87) 国際公開番号	W02014/109127	(74) 代理人	100109313
(87) 国際公開日	平成26年7月17日(2014.7.17)		弁理士 机 昌彦
審査請求日	平成28年10月14日(2016.10.14)	(74) 代理人	100124154
(31) 優先権主張番号	特願2013-3970 (P2013-3970)		弁理士 下坂 直樹
(32) 優先日	平成25年1月11日(2013.1.11)	(72) 発明者	劉 健全
(33) 優先権主張国	日本国(JP)		東京都港区芝五丁目7番1号
			日本電気株式会社内
		審査官	笠田 和宏
			最終頁に続く

(54) 【発明の名称】 インデックス生成装置及び方法並びに検索装置及び検索方法

(57) 【特許請求の範囲】

【請求項1】

インデックス対象となるデータを示す少なくとも1つのエントリをそれぞれ含む複数のノードが階層化されたインデックスを生成するインデックス生成装置において、

類似度算出部により算出されるデータ間の類似度を用いて、下位ノードと、該下位ノードとリンク関係を持つエントリであって、該下位ノードのエントリにより示されるデータとの類似度が該リンク関係に付与される類似度閾値以上の代表データを示すエントリを含む上位ノードと、を含む前記インデックスを生成するインデックス生成部と、

を備えるインデックス生成装置。

【請求項2】

前記類似度閾値を階層が下がる程大きい値に決定する閾値決定部、

を更に備える請求項1に記載のインデックス生成装置。

【請求項3】

インデックス対象データを取得する対象取得部、

を更に備え、

前記インデックス生成部は、前記上位ノードのエントリにより示される前記代表データと前記対象取得部により取得されるインデックス対象データとの類似度を前記類似度算出部から取得し、該取得される類似度と前記類似度閾値とを比較し、該比較結果に基づいて、該インデックス対象データを示すエントリの挿入先を前記上位ノード又は前記下位ノードに決定する、

請求項 2 に記載のインデックス生成装置。

【請求項 4】

前記インデックス生成部は、

前記上位ノードに含まれる複数エントリの中から代表エントリを選択する代表選択部と

、

前記上位ノードのエントリ数が所定エントリ閾値を超える場合に、前記上位ノードに含まれる複数エントリの中から、前記代表選択部により選択される代表エントリにより示されるデータとの類似度が前記閾値決定部により決定される新たな類似度閾値以上のデータを示すエントリを選択し、該選択されたエントリを含む下位ノードを生成し、該上位ノードの該代表エントリに該下位ノードへのリンク関係を設定し、該リンク関係に該新たな類似度閾値を付与する溢れ調整部と、

10

を含む請求項 2 又は 3 に記載のインデックス生成装置。

【請求項 5】

前記代表選択部は、他の各エントリにより示される各データとの類似度の分散が最小となるデータを示すエントリ、又は、自身が最近傍となる他のエントリにより示されるデータの数が最大となるデータを示すエントリを前記代表エントリとして選択する、

請求項 4 に記載のインデックス生成装置。

【請求項 6】

前記閾値決定部は、既に決定されている類似度閾値と均等間隔の値、既に決定されている類似度閾値と最大類似度とを再帰的に 2 分割して得られる値、又は、前記代表データと前記上位ノードに含まれる他のエントリとの類似度から得られる値を新たな類似度閾値に決定する、

20

請求項 2 から 5 のいずれか 1 項に記載のインデックス生成装置。

【請求項 7】

前記代表選択部は、前記上位ノードに含まれる複数エントリで示される複数データの中から複数の代表データを選択し、

前記溢れ調整部は、前記上位ノードのエントリ数が所定エントリ閾値を超える場合に、前記上位ノードに含まれる複数エントリで示される複数データの中から、前記代表選択部により選択される複数の代表データとの複数の類似度の中の少なくとも 1 つが前記閾値決定部により決定される新たな類似度閾値以上のデータを選択し、該選択されたデータを示すエントリを含む下位ノードを生成し、該上位ノードには該複数の代表データを示すエントリを生成し、該エントリに該下位ノードへのリンク関係を設定し、該リンク関係に該新たな類似度閾値を付与する、

30

請求項 4 又は 5 に記載のインデックス生成装置。

【請求項 8】

前記代表選択部は、前記上位ノードに含まれる複数エントリで示される複数データの中から、類似度の値が最大又は最小となるデータの組み合わせを前記複数の代表データとして選択する請求項 7 に記載のインデックス生成装置。

【請求項 9】

請求項 1 から 8 のいずれか 1 項に記載のインデックス生成装置により生成される前記インデックスを用いる検索装置において、

40

検索対象データ及び類似度条件を取得する条件取得部と、

前記上位ノードのエントリにより示される少なくとも 1 つの代表データと前記検索対象データとの少なくとも 1 つの類似度を取得する類似度取得部と、

前記類似度取得部により取得される少なくとも 1 つの類似度と、前記少なくとも 1 つの代表データを示すエントリの前記リンク関係に付与された類似度閾値又は前記類似度条件とを比較し、該比較結果に基づいて、前記上位ノードに含まれる前記少なくとも 1 つの代表データを示すエントリ、及び、該エントリとリンク関係を持つ前記下位ノード以下の下位ノードのエントリを、前記検索対象データに関し前記類似度条件に適合する抽出対象から除外する検索処理部と、

50

を備える検索装置。

【請求項 1 0】

インデックス対象となるデータを示す少なくとも 1 つのエントリをそれぞれ含む複数のノードが階層化されたインデックスを生成するインデックス生成方法において、

下位ノードと、該下位ノードとリンク関係を持つエントリであって、該下位ノードのエントリにより示されるデータとの類似度が該リンク関係に付与される類似度閾値以上の代表データを示すエントリを含む上位ノードと、を含む前記インデックスを生成する、

ことを含むインデックス生成方法。

【請求項 1 1】

前記類似度閾値を階層が下がる程大きい値に決定する、

10

ことを更に含む請求項 1 0 に記載のインデックス生成方法。

【請求項 1 2】

インデックス対象データを取得し、

前記上位ノードのエントリにより示される前記代表データと前記取得されたインデックス対象データとの類似度を取得し、

前記取得された類似度と前記類似度閾値とを比較し、

前記比較結果に基づいて、前記インデックス対象データを示すエントリの挿入先を前記上位ノード又は前記下位ノードに決定する、

ことを更に含む請求項 1 1 に記載のインデックス生成方法。

20

【請求項 1 3】

前記上位ノードのエントリ数が所定エントリ閾値を超える場合に、前記上位ノードに含まれる複数エントリの中から、代表エントリを選択し、

前記選択された代表エントリにより示されるデータとの類似度が新たに決定される類似度閾値以上のデータを示すエントリを選択し、

前記選択されたエントリを含む下位ノードを生成し、

前記上位ノードの前記代表エントリに前記下位ノードへのリンク関係を設定し、

前記リンク関係に前記新たな類似度閾値を付与する、

ことを更に含む請求項 1 1 又は 1 2 に記載のインデックス生成方法。

【請求項 1 4】

前記代表エントリの選択は、他の各エントリにより示される各データとの類似度の分散が最小となるデータを示すエントリ、又は、自身が最近傍となる他のエントリにより示されるデータの数が最大となるデータを示すエントリを前記代表エントリとして選択する、

30

請求項 1 3 に記載のインデックス生成方法。

【請求項 1 5】

前記類似度閾値の決定は、既に決定されている類似度閾値と均等間隔の値、既に決定されている類似度閾値と最大類似度とを再帰的に 2 分割して得られる値、又は、前記代表データと前記上位ノードに含まれる他のエントリとの類似度から得られる値を新たな類似度閾値に決定する、

請求項 1 1 から 1 4 のいずれか 1 項に記載のインデックス生成方法。

【請求項 1 6】

40

前記代表エントリの選択は、前記上位ノードのエントリ数が所定エントリ閾値を超える場合に、前記上位ノードに含まれる複数エントリで示される複数データの中から複数の代表データを選択し、

前記エントリの選択は、前記選択された複数の代表データとの複数の類似度の中の少なくとも 1 つが新たに決定される類似度閾値以上のデータを選択し、

前記下位ノードの生成は、前記選択されたデータを示すエントリを含む下位ノードを生成し、

前記リンク関係の設定は、前記上位ノードには前記複数データを示すエントリを生成し、該エントリに前記下位ノードへのリンク関係を設定する、

請求項 1 3 又は 1 4 に記載のインデックス生成方法。

50

【請求項 17】

前記代表エントリの選択は、前記上位ノードに含まれる複数エントリで示される複数データの中から、類似度の値が最大又は最小となるデータの組み合わせを前記複数の代表データとして選択する請求項 16 に記載のインデックス生成方法。

【請求項 18】

請求項 10 から 17 のいずれか 1 項に記載のインデックス生成方法により生成される前記インデックスを用いる検索方法において、

検索対象データ及び類似度条件を取得し、

前記上位ノードのエントリにより示される少なくとも 1 つの代表データと前記検索対象データとの少なくとも 1 つの類似度を取得し、

前記取得された少なくとも 1 つの類似度と、前記少なくとも 1 つの代表データを示すエントリの前記リンク関係に付与された類似度閾値又は前記類似度条件とを比較し、

前記比較結果に基づいて、前記上位ノードに含まれる前記少なくとも 1 つの代表データを示すエントリ、及び、該エントリとリンク関係を持つ前記下位ノード以下の下位ノードのエントリを、前記検索対象データに関し前記類似度条件に適合する抽出対象から除外する、

ことを含む検索方法。

【請求項 19】

請求項 10 から 17 のいずれか 1 項に記載のインデックス生成方法を少なくとも 1 つのコンピュータに実行させるプログラム。

【請求項 20】

請求項 18 に記載の検索方法を少なくとも 1 つのコンピュータに実行させるプログラム。

【請求項 21】

インデックス対象となるデータを示す少なくとも 1 つのエントリをそれぞれ含む複数のノードが階層化されたインデックスのデータ構造において、

前記インデックスが、

下位ノードと、

前記下位ノードとリンク関係を持つエントリであって、前記下位ノードのエントリにより示されるデータとの類似度が該リンク関係に付与される類似度閾値以上の代表データを示すエントリを含む上位ノードと、

を含み、

コンピュータが、前記上位ノードのエントリにより示される前記代表データと検索対象データとの類似度を算出し、該類似度と前記上位ノードのエントリが持つ前記リンク関係に付与された前記類似度閾値とを比較し、該比較結果に基づいて該検索対象データに対応する検索結果を得るために参照する、

前記インデックスのデータ構造。

【発明の詳細な説明】**【技術分野】****【0001】**

本発明は、データの類似検索技術に関する。

【背景技術】**【0002】**

下記特許文献 1 から 10 に示されるように、現在、様々な類似検索手法が提案されている。このような類似検索は、画像の特徴量データのような多次元データや高次元データを対象に行われることが多い。例えば、特許文献 1、3 及び 6 は、類似画像検索方法を提案する。特許文献 2 は、一方のデータから他方のデータに辿るためのリンクがデータ間に設定されたデータベースを用いて、類似データの検索を行う手法を提案する。特許文献 5 は、任意の画像集合を階層的に分類する手法を提案する。特許文献 7 は、高次元の特徴ベクトルの集合から、クエリ特徴ベクトルに類似した特徴ベクトルを検索する手法を提案する

10

20

30

40

50

。特許文献9は、ハッシュ関数を用いて各学習パターンをハッシュ値に対応するバケットに分類し、入力パターンのハッシュ値に対応するバケットに属する学習パターンの中から、入力パターンに最も類似する学習パターンを探索する手法を提案する。特許文献10は、複数の特徴量をベクトルで表現可能な多次元のデータに対し、条件を指定して所望のデータを抽出するデータマッチング方法を提案する。なお、以降、「高次元」と「多次元」とは、特別に区別されることなく用いられる。

【0003】

このような類似検索では、類似度関数等を用いて、対象データ間の類似度が算出される。例えば、画像の特徴量データは、多次元の数値ベクトルで表わされ、比較対象の特徴量データ間の類似度が類似度関数により算出される。特許文献4は、データベース内の全ての特徴量に関し、他の特徴量との類似度を計算し、類似度の高い順に上位 $f(x)$ 件分のID情報を、類似度順付で格納しておき、この格納内容を検索することにより、類似特徴量を検索する手法を提案する。

10

【0004】

また、対象データに関し索引が構築され、この索引を用いて類似検索を行うことで、検索の高速化が図られる。多次元データの索引生成手法として、R木(R-tree)が知られている(非特許文献1参照)。また、特許文献8は、特徴ベクトル空間を複数の近似領域に分割し、各近似領域を近似領域の密疎に応じて階層化されたインデクシングツリーを生成する手法を提案する。なお、下記非特許文献2、3及び4については後述する。

【先行技術文献】

20

【特許文献】

【0005】

【特許文献1】特許第4545641号公報

【特許文献2】特開2011-090352号公報

【特許文献3】特開2012-079186号公報

【特許文献4】特開2000-035965号公報

【特許文献5】特開2001-160057号公報

【特許文献6】特許第4906900号公報

【特許文献7】特開2011-257970号公報

【特許文献8】特開2002-163272号公報

30

【特許文献9】特開2009-020769号公報

【特許文献10】特開2004-046612号公報

【非特許文献】

【0006】

【非特許文献1】Antonin Guttman著「R-Trees: A Dynamic Index Structure for Spatial Searching」、SIGMOD Conference出版、1984年、pp.47-57

【非特許文献2】Wei Dong、Moses Charikar、Kai Li、「Efficient k-nearest neighbor graph construction for generic similarity measures」、WWW 2011、577-586

【非特許文献3】Stanley Milgram、「The Small World Problem」、Psychology Today、May 1967、pp.60-67

40

【非特許文献4】J.Travers and S.Milgram、「An experimental study of the small world problem」、Sociometry 32、425、1969

【発明の概要】

【発明が解決しようとする課題】

【0007】

しかしながら、上述のような類似検索手法は、対象データのデータ構造に強く依存しているため、対象データの次元数が所定数を超える場合、対象データの次元数や、各次元におけるデータ型等のようなデータ内部構造が未知の場合等には、対応できない。例えば、R木のような空間索引を用いた手法は、距離空間における三角不等式等の幾何学的性質を用いて類似検索処理を行うため、次元数や各次元のデータ内部構造が未知の場合、空間索

50

引を構築することができず、ひいては、類似検索処理を行うことができない。ここで、対象データのデータ構造とは、対象データの全体のデータフォーマット、対象データの次元数、各次元のデータの内部構造などを意味する。

【0008】

本発明は、このような事情に鑑みてなされたものであり、対象データのデータ構造に依存しない汎用的な類似検索技術を提供することにある。

【課題を解決するための手段】

【0009】

本発明の各側面では、上述した課題を解決するために、それぞれ以下の構成を採用する。

10

【0010】

第1の側面は、インデックス対象となるデータを示す少なくとも1つのエントリをそれぞれ含む複数のノードが階層化されたインデックスを生成するインデックス生成装置に関する。第1の側面に係るインデックス生成装置は、類似度算出部により算出されるデータ間の類似度を用いて、下位ノードと、下位ノードとリンク関係を持つエントリであって、下位ノードのエントリにより示されるデータとの類似度がリンク関係に付与される類似度閾値以上の代表データを示すエントリを含む上位ノードと、を含むインデックスを生成するインデックス生成部と、上記類似度閾値を階層が下がる程大きい値に決定する閾値決定部と、を有する。

【0011】

20

第2の側面は、インデックス対象となるデータを示す少なくとも1つのエントリをそれぞれ含む複数のノードが階層化されたインデックスを生成するインデックス生成方法に関する。第2の側面に係るインデックス生成方法は、下位ノードと、下位ノードとリンク関係を持つエントリであって、下位ノードのエントリにより示されるデータとの類似度がリンク関係に付与される類似度閾値以上の代表データを示すエントリを含む上位ノードと、を含むインデックスを生成し、上記類似度閾値を階層が下がる程大きい値に決定する、ことを含む。

【0012】

第3の側面は、上述の第1の側面又は第2の側面により生成されるインデックスを用いる検索装置に関する。第3の側面に係る検索装置は、検索対象データ及び類似度条件を取得する条件取得部と、上位ノードのエントリにより示される少なくとも1つの代表データと検索対象データとの少なくとも1つの類似度を取得する類似度取得部と、類似度取得部により取得される少なくとも1つの類似度と、少なくとも1つの代表データを示すエントリのリンク関係に付与された類似度閾値又は類似度条件とを比較し、比較結果に基づいて、上位ノードに含まれる少なくとも1つの代表データを示すエントリ、及び、そのエントリとリンク関係を持つ下位ノード以下の下位ノードのエントリを、検索対象データに関し類似度条件に適合する抽出対象から除外する検索処理部と、を有する。

30

【0013】

第4の側面は、上述の第1の側面又は第2の側面により生成されるインデックスを用いる検索方法に関する。第4の側面に係る検索方法は、検索対象データ及び類似度条件を取得し、上位ノードのエントリにより示される少なくとも1つの代表データと検索対象データとの少なくとも1つの類似度を取得し、取得された少なくとも1つの類似度と、少なくとも1つの代表データを示すエントリのリンク関係に付与された類似度閾値又は類似度条件とを比較し、この比較結果に基づいて、上位ノードに含まれる少なくとも1つの代表データを示すエントリ、及び、このエントリとリンク関係を持つ下位ノード以下の下位ノードのエントリを、検索対象データに関し類似度条件に適合する抽出対象から除外する、ことを含む。

40

【0014】

第5の側面は、インデックス対象となるデータを示す少なくとも1つのエントリをそれぞれ含む複数のノードが階層化されたインデックスのデータ構造に関する。第5の側面に

50

係るインデックスのデータ構造は、当該インデックスが、下位ノードと、下位ノードとリンク関係を持つエントリであって、下位ノードのエントリにより示されるデータとの類似度がリンク関係に付与される類似度閾値以上の代表データを示すエントリを含む上位ノードと、を含み、上記類似度閾値が階層が下がる程大きい値に設定されており、コンピュータが、上位ノードのエントリにより示される代表データと検索対象データとの類似度を算出し、この類似度と上位ノードのエントリが持つリンク関係に付与された類似度閾値とを比較し、この比較結果に基づいて検索対象データに対応する検索結果を得るために参照する。

【0015】

なお、本発明の他の側面としては、上記第2又は第4の側面の方法を少なくとも1つのコンピュータに実行させるプログラムであってもよいし、このようなプログラムを記録したコンピュータが読み取り可能な記録媒体であってもよい。また、上記第5の側面のデータ構造を持つインデックスを記録したコンピュータが読み取り可能な記録媒体であってもよい。この記録媒体は、非一時的な有形の媒体を含む。

【発明の効果】

【0016】

上記各側面によれば、対象データのデータ構造に依存しない汎用的な類似検索技術を提供することができる。

【図面の簡単な説明】

【0017】

上述した目的、およびその他の目的、特徴および利点は、以下に述べる好適な実施の形態、およびそれに付随する以下の図面によってさらに明らかになる。

【0018】

【図1】本発明の実施の形態に係るインデックス生成装置の構成例を概念的に示す図である。

【図2】本実施形態で生成されるインデックスの一部の例を概念的に示す図である。

【図3】第1実施形態における高次元データ検索装置のハードウェア構成例を概念的に示す図である。

【図4】第1実施形態における高次元データ検索装置の処理構成例を概念的に示す図である。

【図5】インデックス生成処理に関する高次元データ検索装置の動作例を示すフローチャートである。

【図6】データ削除処理に関する高次元データ検索装置の動作例を示すフローチャートである。

【図7】類似検索処理に関する高次元データ検索装置の動作例を示すフローチャートである。

【図8】第2実施形態における類似度木の一部の例を概念的に示す図である。

【図9】実施例1におけるインデックス生成過程を概念的に示す図である。

【図10】実施例1におけるインデックス生成過程を概念的に示す図である。

【図11】実施例1におけるインデックス生成過程を概念的に示す図である。

【図12】実施例1におけるインデックス生成過程を概念的に示す図である。

【図13】実施例1におけるインデックス生成過程を概念的に示す図である。

【図14】実施例1におけるインデックス生成過程を概念的に示す図である。

【図15】実施例1におけるインデックス生成過程を概念的に示す図である。

【図16】実施例1におけるインデックス生成過程を概念的に示す図である。

【図17A】実施例1におけるインデックス生成過程を概念的に示す図である。

【図17B】実施例1におけるインデックス生成過程を概念的に示す図である。

【図18】実施例1におけるインデックス生成過程を概念的に示す図である。

【図19】実施例2におけるインデックス生成過程を概念的に示す図である。

【図20】実施例2におけるインデックス生成過程を概念的に示す図である。

10

20

30

40

50

【発明を実施するための形態】

【0019】

以下、本発明の実施の形態について説明する。なお、以下に挙げる実施形態は例示であり、本発明は以下の実施形態の構成に限定されない。

【0020】

図1は、本発明の実施の形態に係るインデックス生成装置100の構成例を概念的に示す図である。インデックス生成装置100は、インデックス対象となるデータを示す少なくとも1つのエントリをそれぞれ含む複数のノードが階層化されたインデックスを生成する。ここで、「ノード」とは、生成されるインデックスのデータ構造を構成する1つのデータ要素を意味し、例えば、配列や構造体やクラス等により実現される。また、「インデックス対象となるデータ」とは、インデックスが付与されるデータを意味する。本実施形態では、インデックス対象となるデータのデータ構造は何ら制限されない。即ち、当該データの次元数は制限されず、かつ、当該データのデータ構造は未知であってもよい。また、本実施形態では、ノードに含まれるエントリとデータとの関連付け形態は、制限されない。エントリにデータが含まれていてもよいし、エントリにはデータの格納領域を指すポインタが含まれていてもよい。

10

【0021】

図1に示されるように、インデックス生成装置100は、類似度算出部105により算出されるデータ間の類似度を用いて、下位ノードと、下位ノードとリンク関係を持つエントリであって、下位ノードのエントリにより示されるデータとの類似度がそのリンク関係に付与される類似度閾値以上の代表データを示すエントリを含む上位ノードと、を含むインデックスを生成するインデックス生成部101と、上記類似度閾値を階層が下がる程大きい値に決定する閾値決定部102と、を有する。本実施形態における類似度とは、近似する度合いを意味し、類似度の値が大きい程、近似する度合いが増し、逆に、類似度の値が小さい程、近似する度合いが低下する。

20

【0022】

また、本実施形態では、上位ノードに含まれるエントリによる下位ノードとのリンク関係の持ち方は制限されない。例えば、エントリが下位ノードを指すポインタを有していてもよい。また、エントリが下位ノードの識別IDを有し、この下位ノードの識別IDと下位ノードを指すポインタとの対応関係が保持されていてもよい。更に、エントリがリンクの識別IDを有し、このリンクの識別IDと下位ノードを指すポインタとの対応関係が保持されていてもよい。

30

【0023】

インデックス生成装置100は、例えば、後述する詳細実施形態における高次元データ検索装置1と同様のハードウェア構成を有し、その高次元データ検索装置1と同様にプログラムが処理されることで、上述の各処理部が実現される。類似度算出部105は、インデックス生成装置100により実現されてもよいし、他のコンピュータにより実現されてもよい。

【0024】

また、本発明の実施の形態に係るインデックス生成方法は、インデックス対象となるデータを示す少なくとも1つのエントリをそれぞれ含む複数のノードが階層化されたインデックスを生成する。本インデックス生成方法は、インデックス生成装置100等のようなコンピュータにより実行され、下位ノードと、下位ノードとリンク関係を持つエントリであって、下位ノードのエントリにより示されるデータとの類似度が当該リンク関係に付与される類似度閾値以上の代表データを示すエントリを含む上位ノードと、を含むインデックスを生成し、上記類似度閾値を階層が下がる程大きい値に決定する、ことを含む。但し、本インデックス生成方法に含まれる各工程は、順不同に逐次的に実行されてもよいし、同時に実行されてもよい。

40

【0025】

図2は、本実施形態で生成されるインデックスの一部の例を概念的に示す図である。図

50

2に示されるように、本実施形態では、下位ノードL1と、下位ノードL1とリンク関係R1を持つエントリE2であって、下位ノードL1のエントリE1(1)、E1(2)、E1(3)及びE1(n)により示されるデータ(f1、f3、f9等)との類似度がそのリンク関係R1に付与される類似度閾値T1以上の代表データf8を示すエントリE2を含む上位ノードL2と、を含むインデックスを生成する。更に、類似度閾値T1は、階層が下がる程大きい値に決定される。従って、図2の例によれば、類似度閾値T1は、それより上位のリンク関係に付与される類似度閾値T2よりも大きい値に決定される。なお、図2の例では、代表データf8を示すエントリは、上位ノードL2のみに含まれるが、上位ノードL2及び下位ノードL1の両方にそれぞれ含まれていてもよい。

【0026】

10

以下、本実施形態により具現化されている技術的思想と共に、本実施形態の作用及び効果について説明する。

【0027】

本発明者は、データAとデータBとが類似し、かつ、データBとデータCとが類似する場合、データAとデータCとが類似するという規則(非特許文献2)に着眼し、この規則を類似検索のためのインデックスのデータ構造に採用するという着想を得た。この規則は、必ずしも真とはならないが、真となる可能性が高い。この点の実世界での例として、いわゆるSMALL-WORLD理論が知られている(上記非特許文献3及び4参照)。即ち、この理論は、実世界において、人物Aと人物Bとが友人であり、かつ、人物Bと人物Cとが友人である場合、人物Aと人物Cとが友人である可能性が高いことを示す。

20

【0028】

ここで、インデックス対象となる3つのデータをfa、fb及びfcで表し、インデックス対象となるデータfiの集合をFで表し、検索対象データをfqで表す。ここで、データfbは集合Fに属し、データfbは集合Fに含まれる全データの代表であると仮定する。この場合、上記規則は、以下の式で表すことができる。下記式におけるsim()は、2つのデータ間の類似度を計算する関数であり、データとデータの集合との類似度を計算することもできる。例えば、sim()は、類似度算出部105に相当する。また、δは、類似度の閾値を示す。

【数1】

$$(1) \text{sim}(f_a, f_b) \geq \delta \wedge \text{sim}(f_b, f_c) \geq \delta \Rightarrow \text{sim}(f_a, f_c) \geq \delta$$

30

$$(2) \text{sim}(f_q, f_b) \geq \delta \wedge \text{sim}(f_b, f_c) \geq \delta \Rightarrow \text{sim}(f_q, f_c) \geq \delta$$

$$(3) \text{sim}(f_q, f_b) \geq \delta \wedge \text{sim}(f_b, F) \geq \delta \Rightarrow \text{sim}(f_q, F) \geq \delta$$

【0029】

上記式(1)は、上述の規則を表す。即ち、上記式(1)は、データfaとデータfbとの類似度が閾値δ以上であり、かつ、データfbとデータfcとの類似度が閾値δ以上である場合、データfaとデータfcとの類似度は閾値δ以上であることを示す。ここで、上記式(1)のデータfaを検索対象データfqに置き換えると、上記式(2)が成立する。即ち、データfqとデータfbとの類似度が閾値δ以上であり、かつ、データfbとデータfcとの類似度が閾値δ以上である場合、データfqとデータfcとの類似度は閾値δ以上であると判断することができる。

40

【0030】

更に、上記式(2)のデータfcをデータ集合Fに置き換えると、上記式(3)が成立する。即ち、データfqとデータfbとの類似度が閾値δ以上であり、かつ、データfbとデータ集合Fとの類似度が閾値δ以上である場合、データfqとデータ集合Fとの類似度は閾値δ以上であると判断することができる。

【0031】

50

上述したように、データ f_b は集合 F に含まれる全データの代表である。よって、上記式 (3) によれば、データ f_b とデータ集合 F に含まれる各データ f_i との類似度が予めそれぞれ算出されている状態では、検索対象データ f_q と代表のデータ f_b との類似度のみ算出することで、検索対象データ f_q とデータ集合 F 内の各データ f_i との類似性が計算なく判断され得る。本発明者は、このような規則を本実施形態におけるインデックスのデータ構造に採用した。

【0032】

図2によれば、下位ノード L_1 がデータ集合 F に相当し、代表データ f_8 が集合 F の代表のデータ f_b に相当する。下位ノード L_1 には、代表データ f_8 との類似度が類似度閾値 T_1 以上のデータを示すエントリが設定される。これにより、本実施形態によれば、代表データ f_8 と検索対象データ f_q との類似度のみを算出することで、検索対象データ f_q と下位ノード L_1 で示されるデータ集合との類似度を、計算することなく、判断することができる。例えば、代表データ f_8 と検索対象データ f_q との類似度が類似度閾値 T_1 未満の場合には、検索対象データ f_q と下位ノード L_1 により示される全データとの類似度も、類似度計算することなく、類似度閾値 T_1 未満と判断することができる。

10

【0033】

従って、本実施形態により生成されるインデックスを用いることにより、検索時の計算量（類似度計算回数）を減らすことができる。

【0034】

更に、図2の例によれば、類似度閾値 T_1 は、それより上位のリンク関係に付与される類似度閾値 T_2 よりも大きい値に決定される。これにより、上位ノード L_2 に含まれる各エントリで示される各データと、更なる上位ノード L_3 に含まれるその代表データ f_{13} との類似度が類似度閾値 T_2 以上かつ類似度閾値 T_1 未満となる。結果として、本実施形態で生成されたインデックスでは、階層が下がる程、高い類似度を持つデータ集合が示される。

20

【0035】

このように、本実施形態では、対象データのデータ構造に依存せず、データ間の類似度に基づいて階層化された木構造のインデックスが生成される。更に、本実施形態におけるインデックス生成は、対象データの次元数にも依存しない。これは、本実施形態がデータの次元に対応する幾何学的性質に依存することなく、データ間の類似度に基づいてインデックスを生成しているからである。

30

【0036】

従って、本実施形態により生成されるインデックスを用いることにより、対象データのデータ構造が未知である場合や対象データが高次元データである場合等においても、類似検索をすることができる。即ち、本実施形態によれば、対象データのデータ構造に依存しない汎用的な類似検索技術を提供することができる。

【0037】

更に、本実施形態によれば、類似度の関係を用いた上述の規則をインデックスのデータ構造に採用しているため、上述のように類似度計算量を削減することができ、ひいては、対象データの次元数の増加に伴う検索効率の低下を防ぐこともできる。この点、背景技術で述べた提案手法によれば、データの次元数の増加に伴い、空間索引内でのノード間のオーバーラップが大量に発生するため、検索時に全データをスキャンすることになり、検索効率が急激に低下していた。

40

【0038】

更に、本実施形態では、データ間の類似度に基づいてインデックスを生成するため、上述の類似度算出部 105 の類似度算出手法はブラックボックスでもよい。よって、本実施形態によれば、対象データのデータ構造だけでなく、類似度算出手法にも依存しない汎用的でかつ効率的な類似検索技術を提供することができる。

【0039】

以下、上述の実施形態について更に詳細を説明する。以下には、詳細実施形態として、

50

第1実施形態及び第2実施形態を例示する。以下の各実施形態は、上述のインデックス生成装置及びインデックス生成方法を高次元データ検索装置に適用した場合の例である。なお、上述のインデックス生成装置及びインデックス生成方法は、高次元データを扱う検索装置への適用に限定されるものではなく、類似度を算出し得る様々なデータの検索装置に適用可能である。

【0040】

〔第1実施形態〕

〔装置構成〕

図3は、第1実施形態における高次元データ検索装置（以降、単に検索装置と表記する）1のハードウェア構成例を概念的に示す図である。第1実施形態における検索装置1は、図3に示されるように、ハードウェア構成として、相互にバス15により接続される、CPU（Central Processing Unit）10、メモリ11、入出力インタフェース（I/F）12、通信装置13等を有する。メモリ11は、RAM（Random Access Memory）、ROM（Read Only Memory）、ハードディスク、可搬型記憶（記録）媒体等である。入出力I/F12は、キーボード、マウス等のようなユーザ操作の入力を受け付ける入力装置（図示せず）、表示装置やプリンタ等のようなユーザに情報を提供する出力装置（図示せず）などと接続可能である。通信装置13は、他のノードと通信を行う。なお、検索装置1は、入力装置や出力装置を持たなくてもよく、検索装置1のハードウェア構成は制限されない。

【0041】

検索装置1は、インデックス対象データを取得し、取得されるインデックス対象データに基づいて上述の実施形態で示されるデータ構造を持つインデックスを生成する。また、検索装置1は、検索対象データ及び類似度条件を取得し、上述のように生成されたインデックスを用いて、検索対象データに関し類似度条件に適合するデータを検索する。本実施形態では、検索装置1により生成されるインデックスを類似度木と表記する場合もある。類似度木とは、類似度に基づく階層構造を持つインデックスを意味する。検索装置1の一部が上述のインデックス生成装置100に相当する。

【0042】

〔処理構成〕

図4は、第1実施形態における検索装置1の処理構成例を概念的に示す図である。第1実施形態における検索装置1は、対象取得部20、インデックス生成部21、閾値決定部22、検索部30、データベース35等を有する。これら各処理部は、例えば、CPU10によりメモリ11に格納されるプログラムが実行されることにより実現される。また、当該プログラムは、例えば、CD（Compact Disc）、メモ리카ード等のような可搬型記録媒体やネットワーク上の他のコンピュータから入出力I/F12を介してインストールされ、メモリ11に格納されてもよい。

【0043】

対象取得部20は、映像等のような高次元の特徴量データをインデックス対象データとして取得する。特徴量データは、入力画面等に基づいて入力装置をユーザが操作することにより入力された情報であってもよいし、可搬型記録媒体、他のコンピュータ等から入出力I/F12又は通信装置13を経由して取得された情報であってもよい。

【0044】

インデックス生成部21は、上述のインデックス生成装置100におけるインデックス生成部101と同様に、対象取得部20により取得される特徴量データのインデックスとなる類似度木を生成及び更新する。生成又は更新された類似度木のデータは、データベース35に格納される。なお、当該類似度木のデータは、入出力I/F12を介して可搬型記録媒体に格納されてもよいし、通信装置13を介して他のコンピュータに送信されてもよい。

【0045】

本実施形態は、類似度木を構成する各ノードが次のようなデータ構造を持つ形態を採用

する。下位ノードとリンク関係を持つ上位ノードのエントリは、その下位ノードを指すポインタと、そのリンク関係に付与される類似度閾値と、そのエントリにより示される特徴量データとを含む。その上位ノードにおける、下位ノードとリンク関係を持たないエントリは、そのエントリにより示される特徴量データを含む。また、下位ノードへのポインタを含むエントリを1つも持たないノードに含まれる各エントリは、各エントリにより示される特徴量データのデータベース35内の格納領域を指すポインタをそれぞれ含む。また、下位ノードへのポインタが設定されたエントリが持つ特徴量データは、その下位ノードに含まれるエントリにも設定されるか、又は、その下位ノードに含まれるエントリにその特徴量データへのポインタが設定される。

【0046】

10

以降、下位ノードへのポインタを含むエントリを1つも持たないノードは、葉ノードと表記され、下位ノードへのポインタを含むエントリを1つでも持つノードは、中間ノードと表記される場合がある。また、中間ノードの中で最上位のノードは、ルートノードと表記される場合もある。よって、図2によれば、ノードL3はルートノード又は中間ノードと表記され、ノードL2は中間ノードと表記され、ノードL1は葉ノードと表記され得る。

【0047】

また、上位ノードと、その上位ノードに含まれるエントリのポインタでリンクされる下位ノードとの関係は、親子関係と表記される場合もあり、その親子関係を持つ上位ノード及び下位ノードは、親ノード及び下位ノードとそれぞれ表記される場合もある。よって、図2によれば、ノードL2とノードL1との関係では、ノードL2は親ノードと表記され、ノードL1は子ノードと表記され得る。

20

【0048】

インデックス生成部21は、対象取得部20により取得された対象の特徴量データを類似度木に反映する際、当該対象の特徴量データと、下位ノードへのポインタを含む中間ノードのエントリにより示される特徴量データとの類似度を類似度算出部25から取得し、その類似度とそのエントリに含まれる類似度閾値とを比較し、この比較結果に基づいて、当該対象の特徴量データを含むエントリの挿入先をその中間ノード又はその下位ノードに決定する。具体的には、インデックス生成部21は、類似度が類似度閾値以上であれば、当該対象の特徴量データを含むエントリの挿入先を下位ノードに決定し、類似度が類似度閾値未満であれば、当該対象の特徴量データを含むエントリの挿入先を中間ノードに決定する。

30

【0049】

また、インデックス生成部21は、代表選択部23及び溢れ調整部24を含む。

【0050】

代表選択部23は、特徴量データの追加時又は削除時に、必要に応じて、或るノードに含まれる複数エントリの中から代表エントリを選択する。これは、データ集合Fの中から代表のデータfbを選択することに相当する。よって、代表エントリとは、代表として選択された特徴量データを示すエントリを意味する。また、代表エントリに含まれる特徴量データは代表特徴量データとも表記される。代表エントリを選択するエントリの母集団は、対象ノードに含まれる全エントリであってもよいし、対象ノードに含まれる全エントリの中の下位ノードへのポインタを含まないエントリのみであってもよい。母集団を対象ノードに含まれる全エントリとする場合、既に存在する子ノードの再構成が必要になる場合がある。

40

【0051】

例えば、代表選択部23は、以下に示す2つの選択方法のいずれか1つを用いて、代表エントリを選択する。但し、本実施形態は、代表エントリ選択方法を以下のような方法に制限しない。代表選択部23は、以下に示す2つの選択方法以外で、任意のエントリを代表エントリとしてランダムに選択してもよい。

【0052】

50

第1の選択方法では、代表選択部23は、他の各エントリにより示される各特微量データとの類似度の分散が最小となる特微量データを示すエントリを代表エントリとして選択する。この方法は、以下の式により表わすことができる。代表選択部23は、下記式に表わされるように、集合Fに属する各特微量データ f_i に関し、特微量データ f_i と他の特微量データ f_j との全ての類似度をそれぞれ算出し、類似度分布の分散値 σ が最小となる特微量データ f_i を、集合Fの代表として選出する。但し、第1の選択方法において、代表エントリとして選択されるエントリは、各データとの類似度の分散が最小となるデータを示すエントリのみには制限されるわけではない。代表エントリは、各データとの類似度の分散が小さいデータを示すエントリから優先的に選択されるとしてもよい。例えば、当該分散が2番目に小さいデータを示すエントリが代表エントリに選択されてもよい。

10

【数2】

$$\operatorname{argmin}_{f_i \in F} \{ \sigma_{f_i} \mid \sigma \text{ が } \operatorname{sim}(f_i, f_j) \text{ の分散}, f_j \in F \}$$

【0053】

第2の選択方法では、代表選択部23は、自身が最近傍となる他のエントリにより示されるデータの数が最大となるデータを示すエントリを代表エントリとして選択する。この方法は、以下の式により表わすことができる。代表選択部23は、下記式に表わされるように、集合Fに属する各特微量データ f_i に関し、特微量データ f_i が最近傍となる他の特微量データの数 RNN (Reverse Nearest Neighbors) (f_i) が最大となる特微量データ f_i を、集合Fの代表として選出する。但し、第2の選択方法において、代表エントリとして選択されるエントリは、当該データの数が最大となるデータを示すエントリのみには制限されるわけではない。代表エントリは、当該データの数が大きいデータを示すエントリから優先的に選択されるとしてもよい。例えば、当該データの数が2番目に大きいデータを示すエントリが代表エントリに選択されてもよい。

20

【数3】

$$\operatorname{argmax}_{f_i \in F} \{ |RNN(f_i)| \}$$

30

【0054】

溢れ調整部24は、対象取得部20により取得された特微量データを含むエントリが挿入されたノードのエントリ数が所定エントリ閾値を超える場合に、そのノードを子ノードと親ノードとに分割する。このようにノードを親ノードと子ノードとに分割することを溢れ調整とも表記する。このとき、溢れ調整部24は、エントリ数が所定エントリ閾値を超えたノードを対象に代表選択部23に代表エントリを選択させ、更に、閾値決定部22に新たな類似度閾値を決定させる。溢れ調整部24は、代表選択部23により選択された代表エントリに含まれる特微量データとの類似度が閾値決定部22により決定された新たな類似度閾値以上の特微量データを含むエントリを選択し、この選択されたエントリを含む下位ノード(子ノード)を生成し、元のノード(親ノード)の代表エントリにその下位ノードへのポインタとその新たな類似度閾値とを設定する。ここで、代表エントリは、子ノード及び親ノードの両方に含められてもよいし、親ノードのみに含められてもよい。

40

【0055】

閾値決定部22は、上述のインデックス生成装置100における閾値決定部102と同様に、類似度閾値を階層が下がる程大きい値に決定する。例えば、閾値決定部22は、以下に示す3つの決定方法のいずれか1つを用いて、新たな類似度閾値を決定する。但し、本実施形態は、類似度決定方法を以下のような方法に制限しない。

【0056】

50

第1の決定方法では、新たな類似度閾値が、既に決定されている類似度閾値と均等間隔の値に決定される。言い換えれば、第1の決定方法では、0から1までの類似度範囲を等間隔に分割することにより、各類似度閾値がそれぞれ決定される。例えば、0.1の等間隔で[0, 1]の類似度範囲区間を10等分割し、各類似度閾値はそれぞれの分割点の数値に決定される。この場合、類似度閾値は、0.1、0.2、0.3、...、0.9、1.0といった値に決定される。

【0057】

第2の決定方法では、新たな類似度閾値が、既に決定されている類似度閾値と最大類似度とを再帰的に2分割して得られる値に決定される。言い換えれば、第2の決定方法では、0から1までの類似度範囲を再帰的にLogスケールに分割することにより、各類似度閾値がそれぞれ決定される。例えば、1番目の類似度閾値が0.2に設定され、2番目の類似度閾値が、0.2から1.0までの範囲を2等分割することにより、その分割点0.6に決定される。このように2等分割が繰り返されることで、3番目以降の各類似度閾値は、0.8、0.9、0.95、0.975、0.9875などに決定される。

【0058】

第3の決定方法では、新たな類似度閾値が、代表の特徴量データと、代表の特徴量データが属するデータ集合の他の特徴量データとの類似度から得られる値に決定される。例えば、算出される複数の類似度の中の中央値となる類似度が、新たな類似度閾値に決定される。具体的には、データ集合に含まれる5つの特徴量データと代表の特徴量データとの類似度がそれぞれ、0.5、0.3、0.6、0.9、0.95である場合、新たな類似度閾値が、中央値0.6に決定される。また、算出される複数の類似度の中の最大値と最小値との間を2等分割し、その分割点が新たな類似度閾値に決定されてもよい。

【0059】

類似度算出部25は、2つの特徴量データ間の類似度を算出する。本実施形態では、類似度算出部25による類似度算出手法は何ら制限されない。類似度算出部25は、外部から提供される内部処理がブラックボックス化されたAPI (Application Programming Interface) やライブラリ等を用いて、当該類似度を算出してよい。

【0060】

データベース35は、インデックス対象となる特徴量データ群、インデックス生成部21により生成された類似度木の情報、この類似度木に関連する情報(類似度木の生成のために算出された類似度等)を格納する。インデックス対象となるデータ群が、表形式データを形成する或る属性のデータである場合には、データベース35は、その表形式データを格納してもよい。データベース35は、メモリ11に含まれるハードディスク等のような外部メモリ上に実現される。また、類似度木の生成及び検索を高速化するために、データベース35に格納される類似度木の葉ノード以外のノードの情報は、メモリ11に含まれる主記憶にバッファリングされる。

【0061】

検索部30は、インデックス生成部21により生成される類似度木を用いて、データベース35に格納される特徴量データ群の中から、検索対象データに関し類似度条件に適合する特徴量データを検索する。検索結果のデータは、入出力I/F12を介して表示装置や印刷装置に出力されてもよいし、入出力I/F12を介して可搬型記録媒体に格納されてもよいし、通信装置13を介して他のコンピュータに送信されてもよい。

【0062】

検索部30は、図4に示されるように、条件取得部31、類似度算出部32及び検索処理部33を含む。

【0063】

条件取得部31は、検索対象の特徴量データ及び類似度条件を取得する。検索対象の特徴量データは、入力画面等に基づいて入力装置をユーザが操作することにより入力された情報であってもよいし、可搬型記録媒体、他のコンピュータ等から入出力I/F12又は通信装置13を経由して取得された情報であってもよい。

【0064】

類似度算出部32は、類似度算出部25と同様である。検索装置1において、類似度算出部25と類似度算出部32とは1つの処理部として共有されてもよい。

【0065】

検索処理部33は、類似度算出部32により取得される類似度と、類似度閾値、又は、条件取得部31により取得された類似度条件とを比較し、この比較結果に基づいて、その類似度閾値を持つエントリ、及び、その類似度閾値を含むエントリとリンクされる子ノード以下の下位ノードのエントリを、検索対象の特徴量データに関する類似度条件に適合する抽出対象（検索結果）から除外する。但し、検索処理部33は、辿り着いたノードに含まれる全エントリをチェックする幅優先探索法を用いてもよいし、同一ノード内の他のエントリよりもリンク先の子ノードのエントリを優先的にチェックする深さ優先探索法を用いてもよい。

10

【0066】

〔動作例〕

まず、第1実施形態におけるインデックス生成方法について図5及び図6を用いて説明する。以下の説明では、検索装置1が各方法の実行主体となるが、検索装置1に含まれる上述の各処理部が実行主体となってもよい。

【0067】

図5は、インデックス生成処理に関する検索装置1の動作例を示すフローチャートである。ここでは、子ノードへのポインタ及び類似度閾値が設定された少なくとも1つのエントリを含むルートノードが、既に生成されていると仮定する。また、図5には示されていないが、インデックス生成処理を開始する前には、検索装置1は、類似度木の生成に関する各種パラメータを既に保持している。パラメータには、エントリ閾値、代表エントリの選択方法及び類似度閾値の決定方法が含まれる。また、パラメータには、類似度木の階層の深さの閾値が含まれていてもよい。

20

【0068】

検索装置1は、インデックス対象となる特徴量データ f_i を取得する（S50）。

続いて、検索装置1は、対象ノードをルートノードに設定する（S51）。対象ノードとは、その特徴量データ f_i を示すエントリの挿入候補となるノードを意味する。このように、検索装置1は、上位ノードから下位ノードへと順にエントリの挿入候補を選んでいく。

30

【0069】

検索装置1は、対象ノードに含まれる、子ノードへのポインタを含む各エントリについて、そのエントリに含まれる特徴量データ f_e と特徴量データ f_i との類似度 S をそれぞれ算出し、類似度 S が類似度閾値以上となる特徴量データ f_e を含むエントリを特定する（S52）。検索装置1は、対象ノードに、類似度 S が類似度閾値以上となる特徴量データ f_e を含むエントリが存在する場合（S52；YES）、その特定されたエントリの中で類似度 S が最大となるエントリのポインタが示す子ノードを対象ノードに設定する（S53）。

【0070】

検索装置1は、対象ノードが葉ノードでない場合（S54；NO）、工程（S52）以降を再度実行する。検索装置1は、対象ノードが葉ノードである場合（S54；YES）、又は、対象ノードに、該当するエントリが存在しない場合（S52；NO）、特徴量データ f_i を含むエントリをその対象ノードに追加する（S55）。

40

【0071】

検索装置1は、エントリが追加された対象ノードのエントリ数が所定エントリ閾値を超えるか否かを判定する（S56）。検索装置1は、対象ノードのエントリ数が所定エントリ閾値を超える場合（S56；YES）、次のように、その対象ノードに含まれるエントリを親ノードと子ノードとに分割する。

【0072】

50

検索装置 1 は、対象ノードのエントリの中から代表エントリを選択する（S 5 7）。ここでの代表エントリの選択方法には、例えば、上述のような 2 つの選択方法のいずれか 1 つが利用される。また、代表エントリを選択するエントリの母集団は、対象ノードに含まれる全エントリであってもよいし、対象ノードに含まれる全エントリの中の下位ノードへのポインタを含まないエントリのみであってもよい。

【0073】

検索装置 1 は、選択された代表エントリに設定するための類似度閾値を決定する（S 5 8）。ここでの類似度閾値の決定方法には、例えば、上述のような 3 つの決定方法のいずれか 1 つが利用される。

【0074】

検索装置 1 は、代表エントリに含まれる特徴量データと、代表エントリを選択した母集団に含まれる代表エントリ以外の各エントリに含まれる各特徴量データとの類似度をそれぞれ算出する（S 5 9）。

【0075】

続いて、検索装置 1 は、類似度閾値以上の類似度を示す特徴量データを持つエントリを含む子ノードを生成する（S 6 0）。このとき、検索装置 1 は、対象ノードから、その子ノードに追加されたエントリを削除する。また、代表エントリを選択したエントリの母集団が、対象ノードに含まれる全エントリであり、かつ、選択された代表エントリが既に子ノードへのポインタを持っていた場合には、検索装置 1 は、新たに子ノードを生成することなく、既に存在する子ノードに、類似度閾値以上の類似度を示す特徴量データを持つエントリを追加してもよい。

【0076】

検索装置 1 は、対象ノードの代表エントリに、（S 6 0）で生成された子ノードへのポインタと（S 5 8）で決定された類似度閾値とを設定する（S 6 1）。

【0077】

検索装置 1 は、対象ノードのエントリ数が所定エントリ閾値を超えない場合（S 5 6；NO）、又は、（S 6 1）の完了後、特徴量データ f_i と共に、類似度木データをデータベース 3 5 に格納する（S 6 2）。ここで、検索装置 1 は、類似度木データのうち更新された箇所のデータのみをデータベース 3 5 に反映してもよい。また、類似度木データの格納前に、検索装置 1 は、類似度木の葉ノードのエントリに設定されている特徴量データ f_i をデータベース 3 5 に格納される特徴量データ f_i へのポインタに置き換える。

【0078】

図 6 は、データ削除処理に関する検索装置 1 の動作例を示すフローチャートである。

【0079】

検索装置 1 は、削除対象となる特徴量データ f_d （以降、削除データと表記する）を取得すると、その削除データ f_d を示すエントリを特定する（S 7 1）。このとき、検索装置 1 は、葉ノードの各エントリにより示される特徴量データを、そのエントリに含まれるポインタを用いて、データベース 3 5 から取得する。削除データ f_d の特定方法は、図 7 を用いて後述する類似検索方法と同様でよい。ため、ここでは説明を省略する。但し、この場合には、類似度条件 δ は 1（一致）に設定される。

【0080】

検索装置 1 は、特定されたエントリの中の葉ノードのエントリを削除する（S 7 2）。削除データ f_d が或るデータ集合の代表に設定されている場合、特定されたエントリの中には中間ノードも含まれる。また、検索装置 1 は、削除データ f_d をデータベース 3 5 から削除してもよい。

【0081】

続いて、検索装置 1 は、削除されたエントリを含む葉ノードへのポインタを含む親ノードのエントリを対象エントリに設定する（S 7 3）。

【0082】

検索装置 1 は、この対象エントリを更新する必要があるか否かを判定する（S 7 4）。

10

20

30

40

50

例えば、対象エントリが削除データ f d を含む場合、上述の (S 5 7) で実行される代表エントリの選択方法が子ノードで示される特徴量データ集合の値分布に依存する場合などには、検索装置 1 は、対象エントリを更新する必要があると判定する (S 7 4 ; Y E S) 。

【0083】

例えば、一方、代表エントリの選択方法が任意のエントリの選択である場合で、かつ、対象エントリが削除データ f d を含まない場合には、検索装置 1 は、対象エントリを更新する必要があると判定する (S 7 4 ; N O) 。対象エントリを更新する必要があると判定すると (S 7 4 ; N O) 、検索装置 1 は、データ削除処理を終了する。

【0084】

検索装置 1 は、対象エントリを更新する必要があると判定すると (S 7 4 ; Y E S) 、代表エントリを再選択する (S 7 5) 。ここでの代表エントリを選択する母集団は、例えば、対象エントリの子ノードに含まれるエントリとされる。但し、当該母集団には、その対象エントリを含むノードの他のエントリが更に加えられてもよい。

【0085】

検索装置 1 は、対象エントリに含まれる元の特徴量データを再選択された代表エントリの特徴量データに更新する (S 7 6) 。

【0086】

検索装置 1 は、更に、その対象エントリに設定されている類似度閾値を更新する必要があるか否かを判定する (S 7 7) 。例えば、上述の (S 5 8) で実行される類似度閾値の決定方法が、上述の第 3 の決定方法のように、代表エントリを選択する母集団に関する類似度に依存する場合には、検索装置 1 は、類似度閾値を更新する必要があると判定する (S 7 7 ; Y E S) 。検索装置 1 は、類似度閾値を更新する必要があると判定すると (S 7 7 ; Y E S) 、新たな類似度閾値を決定し、対象エントリの元の類似度閾値をこの新たな類似度閾値に更新する (S 7 8) 。

【0087】

検索装置 1 は、対象エントリに設定されている特徴量データ及び類似度閾値に基づいて、対象エントリを含むノード及びその対象エントリの現在の子ノードを再構成する必要があるか否かを判定する (S 7 9) 。具体的には、検索装置 1 は、対象エントリの特徴量データと同一ノードの他のエントリの特徴量データとの類似度を対象エントリに設定される類似度閾値と比較することにより、対象ノードに含まれる他のエントリを現在の子ノードに移動させるか否かを判定する。例えば、当該類似度閾値以上となる類似度を示すエントリは、子ノードに移動させるべきと判定される。

【0088】

検索装置 1 は、対象エントリを含むノードの他のエントリを子ノードへ移動させることで、対象エントリを含むノード及び対象エントリの子ノードを再構成する (S 8 0) 。

【0089】

続いて、検索装置 1 は、対象エントリを含むノードの親ノードが存在するか否かを判定する (S 8 1) 。具体的には、検索装置 1 は、対象エントリを含むノードへのポインタを含むエントリを持つ中間ノードが存在するか否かを判定する。

【0090】

検索装置 1 は、対象エントリを含むノードの親ノードが存在する場合 (S 8 1 ; Y E S) 、現対象エントリを含むノードへのポインタを含む親ノードのエントリを新たな対象エントリに設定する (S 8 2) 。検索装置 1 は、この新たな対象エントリについて、工程 (S 7 4) 以降を実行する。ここで、検索装置 1 は、類似度木データのうち更新された箇所のデータをデータベース 3 5 に反映する。

【0091】

次に、第 1 実施形態における類似検索方法について図 7 を用いて説明する。図 7 は、類似検索処理に関する検索装置 1 の動作例を示すフローチャートである。

【0092】

検索装置 1 は、検索対象の特徴量データ（以降、クエリデータ f_q と表記）及び類似度条件 δ を取得する（S 9 0）。ここで取得されたデータは、クエリデータ f_q と類似度条件 δ 以上の類似度を示す特徴量データをデータベース 3 5 から抽出することを要求する。

【0 0 9 3】

検索装置 1 は、対象ノードをルートノードに設定する（S 9 1）。対象ノードとは、クエリデータ f_q との比較候補となるノードを意味する。このように、検索装置 1 は、上位ノードから下位ノードへと順に比較候補を選んでいく。

【0 0 9 4】

検索装置 1 は、対象ノードの各エントリの各特徴量データ f_e とクエリデータ f_q との類似度 S をそれぞれ算出する（S 9 2）。 10

【0 0 9 5】

検索装置 1 は、対象ノードが葉ノードである場合（S 9 3 ; YES）、対象ノードに含まれるエントリの中から、類似度条件 δ 未満の類似度 S を示す特徴量データ f_e を示すエントリを抽出対象から除外する（S 9 4）。

【0 0 9 6】

一方、検索装置 1 は、対象ノードが葉ノードでない場合（S 9 3 ; NO）、対象ノードに含まれるエントリの中から、類似度閾値と類似度条件 δ との小さい方の値未満の類似度 S を示す特徴量データ f_e を示すエントリを抽出対象から除外する（S 9 5）。

【0 0 9 7】

検索装置 1 は、対象ノードに、子ノードを持つエントリが除外されず残っているか否かを判定する（S 9 6）。検索装置 1 は、子ノードを持つエントリが残っている場合（S 9 6 ; YES）、現対象ノードの子ノードを新たな対象ノードに設定する（S 9 7）。検索装置 1 は、この新たな対象ノードについて、（S 9 2）以降を実行する。 20

【0 0 9 8】

検索装置 1 は、抽出対象に残ったエントリの特徴量データ f_e を検索結果に設定する（S 9 8）。

【0 0 9 9】

上述の類似検索方法は、幅優先探索法を用いた方法であったが、深さ優先探索法が用いられてもよい。深さ優先探索法が用いられる類似検索方法では、対象ノードが中間ノードである場合、（S 9 2）では、対象ノードの 1 つのエントリの特徴量データ f_e とクエリデータ f_q との類似度 S が算出される。更に、（S 9 4）の実行後、クエリデータ f_q との比較を行っていないエントリを有する最上位の中間ノードを新たな対象ノードに設定し、（S 9 2）以降が更に実行される。 30

【0 1 0 0】

本実施形態におけるインデックス生成方法及び類似検索方法は、図 5、図 6 及び図 7 に示される例に制限されない。例えば、図 5 の例では、特徴量データ f_i が取得される度に、その特徴量データ f_i 及び類似度木データがデータベース 3 5 に格納されたが、特徴量データ f_i 及び類似度木データのデータベース 3 5 への格納タイミングは、任意である。また、（S 5 9）等の類似度を算出する工程は、毎回、類似度計算をすることなく、過去に算出されている類似度を保持部から取得するようにしてもよい。 40

【0 1 0 1】

〔第 1 実施形態の作用及び効果〕

上述のように、第 1 実施形態では、インデックス対象となる特徴量データの形態に依存することなく、特徴量データ間の類似度及び類似度に関する上述の規則に基づいて、階層化された類似度木が生成される。そして、この類似度木に基づいて、検索対象の特徴量データに関し類似度条件を満たす特徴量データが抽出される。また、第 1 実施形態では、類似度算出部 2 5 及び類似度算出部 3 2 における類似度算出手法自体は、何ら制限されない。このように、第 1 実施形態では、インデックス対象データの幾何学的性質やデータ構造に依存せず、インデックスである類似度木が生成される。更に、第 1 実施形態では、類似度算出手法に依存することなく、当該類似度木の生成及び類似検索が実行される。 50

【0102】

従って、第1実施形態によれば、インデックス対象のデータ構造が未知である場合や特徴量データが高次元データである場合等においても、適切に類似検索をすることができる。即ち、第1実施形態によれば、インデックス対象のデータ構造に依存しない汎用的な類似検索技術を提供することができる。

【0103】

また、第1実施形態では、類似度木に新たな特徴量データを反映させる（追加する）際に、その新たな特徴量データと、下位ノードへのポインタを含む中間ノードのエントリにより示される特徴量データとの類似度が算出され、その類似度とそのエントリに含まれる類似度閾値とが比較され、この比較結果に基づいて、当該新たな特徴量データを含むエントリの挿入先がその中間ノード又はその下位ノードに決定される。更に、第1実施形態では、ノード内のエントリ数が制限（所定エントリ閾値）を超えた場合には、溢れ調整されることにより、そのノードが親ノードと子ノードとに分割される。その溢れ調整では、そのノード内のエントリから代表エントリが選択され、その代表エントリに設定される新たな類似度閾値が決定され、代表エントリの特徴量データと他のエントリの特徴量データとの類似度とその新たな類似度閾値との関係から、各エントリの配置が子ノード又は親ノードに決定される。

10

【0104】

このように、第1実施形態では、各ノードの特徴量データとの類似度に基づいて、上位階層から下位階層へ順番に各ノードがチェックされることにより、取得されたインデックス対象の特徴量データに対応するエントリが適切なノードに挿入される。更に、各ノードのエントリ数に応じて、各ノードが階層的に分割される。

20

【0105】

従って、第1実施形態によれば、インデックス対象のデータに基づいて、インデックス対象のデータ規模によらず、動的に、類似度木を生成及び更新することができる。一方、背景技術で述べた提案技術には、動的なインデックスの生成及び更新をサポートするものが少ない。R木技術のようにそれをサポートしている既存技術であったとしても、大規模データには対応し難い。

【0106】

また、第1実施形態では、検索対象の特徴量データ（クエリデータ）及び類似度条件が取得され、類似度木において上位ノードから下位ノードへ順にチェックされることにより、クエリデータに関する類似度条件を満たす特徴量データが特定され、データベース35から抽出される。このとき、中間ノードでは、エントリにより示される特徴量データとクエリデータとの類似度が算出され、この類似度と、そのエントリに含まれる類似度閾値又は類似度条件とが比較され、この比較結果に基づいて、中間ノードのそのエントリ、及び、そのエントリの子ノード以下の下位ノードのエントリが抽出対象から除外される。

30

【0107】

このように、第1実施形態では、類似度木の上位ノードから順に各エントリがチェックされ、当該類似度条件を満たさないエントリ及びそのエントリの子ノードを含む子孫ノードのエントリが抽出対象から逐次除外される。従って、第1実施形態によれば、類似度木に基づいてチェックするノードを制限することができるため、高い検索効率を実現することができる。更に、第1実施形態によれば、上述のように、類似度木は特徴量データの次元数に依存しないため、特徴量データの次元数が増加することに伴う検索効率の低減を防ぐことができる。

40

【0108】

また、第1実施形態では、ノードの溢れ調整時などに代表エントリを選出する方法として、上述のような2つの選択方法が例示された。代表エントリは、子ノードにより示される特徴量データの集合の代表であり、上述の規則に示されるように、類似度閾値と共に、その子ノードに含めるべき特徴量データを定める判断材料となる。よって、代表エントリは、既に類似度木に登録されている特徴量データに関する親ノードと子ノードとのエント

50

り数のバランスを取るためには、それら特徴量データの集合の中心近辺に位置することが望ましい。上記2つの選択方法によれば、ノードの溢れ調整直後の親ノードと子ノードとのエントリ数のバランスを取ることができる。但し、上記2つの選択方法以外でランダムに選択する方法が採用された場合には、代表エントリ選出の処理負荷を削減することができると共に、ノードの溢れ調整後（ノード分割後）、類似度木生成が進むにつれ、親ノードと子ノードとのエントリ数のバランスは取られる。

【0109】

また、第1実施形態では、ノードの溢れ調整時などに類似度閾値を決定する方法として、上述のような3つの決定方法が例示された。類似度閾値は、子ノードにより示される特徴量データの集合における、代表の特徴量データからの距離の閾値に相当し、上述の規則に示されるように、代表データと共に、その子ノードに含めるべき特徴量データを定める判断材料となる。上述の例における、新たな類似度閾値が、既に登録されている特徴量データの代表データとの類似度の分布から決定される方法によれば、既に類似度木に登録されている特徴量データに関する親ノードと子ノードとのエントリ数のバランスを取ることができる。但し、他の類似度閾値の決定方法においても、類似度閾値決定の処理負荷を削減することができると共に、ノードの溢れ調整後（ノード分割後）、類似度木生成が進むにつれ、親ノードと子ノードとのエントリ数のバランスは取られる。

【0110】

[第2実施形態]

上述の第1実施形態では、子ノードへのポインタを含むエントリが、子ノードで示される特徴量データ集合の代表となる1つの特徴量データを含んでいた。第2実施形態では、代表となる複数の特徴量データが、子ノードへのポインタと共に当該エントリに含まれる。代表となる特徴量データの数の上限は、例えば、子ノードで示される特徴量データの数より小さい値に設定される。以下、第2実施形態における検索装置1について、第1実施形態と異なる内容を中心に説明する。以下の説明では、第1実施形態と同様の内容については適宜省略する。

【0111】

図8は、第2実施形態における類似度木の一部の例を概念的に示す図である。図8に示されるように、第2実施形態では、下位ノードL1へのポインタ（リンク関係R1）を含む上位（中間）ノードL2のエントリE2は、類似度閾値T1と、下位ノードL1に含まれる複数エントリE1（n）で示される複数の特徴量データ（f1、f3、f9等）の代表となる複数の特徴量データf8及びf11を含む。

【0112】

代表選択部23は、特徴量データの追加時又は削除時に、必要に応じて、或るノードに含まれる複数の特徴量データの中から代表となる複数の特徴量データを選択する。中間ノードには、複数の特徴量データ及び子ノードへのポインタを含むエントリと、子ノードへのポインタを含まず1つの特徴量データを含むエントリとが混在する場合がある。この場合、代表選択部23は、子ノードへのポインタを含まないエントリの中から代表となる複数の特徴量データを選択するようにしてもよい。また、代表選択部23は、その中間ノードに含まれる全エントリの全特徴量データの中から代表となる複数の特徴量データを選択するようにしてもよい。

【0113】

例えば、代表選択部23は、以下に示す2つの選択方法のいずれか1つを用いて、代表となる複数の特徴量データを選択する。但し、本実施形態は、代表となる複数の特徴量データの選択方法を以下のような方法に制限しない。代表選択部23は、代表となる複数の特徴量データをランダムに選択するようにしてもよい。

【0114】

第1の選択方法では、代表選択部23は、類似度の値が最大となる特徴量データの組み合わせを代表として選択する。2つの特徴量データが代表として選択される場合、代表選択部23は、類似度の値が最大となる特徴量データのペアを代表として選択する。

【0115】

第2の選択方法では、代表選択部23は、類似度の値が最小となる特徴量データの組み合わせを代表として選択する。2つの特徴量データが代表として選択される場合、代表選択部23は、類似度の値が最小となる特徴量データのペアを代表として選択する。

【0116】

閾値決定部22は、第1実施形態における3つの決定方法に加えて、以下に示す第4の決定方法を用いてもよい。第4の決定方法では、新たな類似度閾値が、複数の代表の特徴量データの中心点となる特徴量データと他の特徴量データとの類似度に決定される。

【0117】

インデックス生成部21は、新たな特徴量データを類似度木に反映する場合に、当該新たな特徴量データと、下位ノードへのポインタを含む中間ノードのエントリにより示される複数の特徴量データとの各類似度をそれぞれ取得し、各類似度と類似度閾値とをそれぞれ比較し、各比較結果に基づいて、当該新たな特徴量データを含むエントリの挿入先をその中間ノード又はその下位ノードに決定する。具体的には、インデックス生成部21は、全ての類似度が類似度閾値以上であれば、当該新たな特徴量データを含むエントリの挿入先を下位ノードに決定し、いずれか1つの類似度でも類似度閾値未満であれば、当該対象の特徴量データを含むエントリの挿入先を中間ノードに決定する。但し、インデックス生成部21は、いずれか1つの類似度でも類似度閾値以上であれば、当該新たな特徴量データを含むエントリの挿入先を下位ノードに決定し、全ての類似度が類似度閾値未満であれば、当該対象の特徴量データを含むエントリの挿入先を中間ノードに決定するようにしてもよい。

【0118】

検索処理部33は、検索対象の特徴量データとエントリに含まれる複数の特徴量データとに関し取得される各類似度と、類似度閾値、又は、条件取得部31により取得された類似度条件とを比較し、この比較結果に基づいて、その類似度閾値を持つエントリ、及び、その類似度閾値を含むエントリとリンクされる子ノード以下の下位ノードのエントリを、検索対象の特徴量データに関する類似度条件に適合する抽出対象（検索結果）から除外する。このとき、全ての類似度が、類似度閾値と類似度条件との小さい方の値未満であれば、その類似度閾値を含むエントリは抽出対象から除外される。

【0119】

〔動作例〕

以下、第2実施形態におけるインデックス生成方法について図5及び図6を用いて説明する。第2実施形態では、(S52)、(S53)、(S57)、(S59)、(S60)及び(S61)において、複数の特徴量データが処理対象とされる点において、第1実施形態と異なる。

【0120】

具体的には、(S52)では、検索装置1は、対象ノードに含まれる各エントリについて、そのエントリに含まれる複数の特徴量データ f_e と特徴量データ f_i との各類似度 S をそれぞれ算出し、類似度閾値以上の複数の類似度 S を示す複数の特徴量データ f_e を示すエントリを特定する。但し、検索装置1は、類似度閾値以上の類似度 S を示す1つの特徴量データを含む複数の特徴量データ f_e を示すエントリが特定されてもよい。(S53)では、検索装置1は、その特定されたエントリの中で類似度 S が最大となるエントリのポインタが示す子ノードを対象ノードに設定する。

【0121】

具体的には、(S57)では、検索装置1は、対象ノードのエントリの中から複数の特徴量データを代表として選択する。(S59)では、検索装置1は、(S57)で代表として選択された複数の特徴量データと、それら代表を選択した母集団に含まれるそれら代表以外の各特徴量データとの類似度をそれぞれ算出する。(S60)において、検索装置1は、代表として選択された複数の特徴量データとの類似度が類似度閾値以上を示す特徴量データを持つエントリを含む子ノードを生成する。但し、検索装置1は、代表として選

択された複数の特徴量データの中の1つとの類似度が類似度閾値以上を示す特徴量データを持つエントリを子ノードに含めてもよい。

【0122】

(S61)では、検索装置1は、代表として選択された複数の特徴量データを含む1つのエントリを生成し、それら代表を含んでいたエントリを削除し、その生成されたエントリに(S60)で生成された子ノードへのポインタと(S58)で決定された類似度閾値とを設定する。

【0123】

第2実施形態では、(S75)、(S76)及び(S79)において、複数の特徴量データが処理対象とされる点において、第1実施形態と異なる。

10

【0124】

(S75)では、検索装置1は、代表となる複数の特徴量データを再選択する。

(S76)では、検索装置1は、再選択された複数の特徴量データを対象エントリに設定する。

【0125】

(S79)では、検索装置1は、対象エントリの複数の特徴量データと同一ノードの他のエントリの特徴量データとの各類似度を対象エントリに設定される類似度閾値とそれぞれ比較することにより、対象ノードに含まれる他のエントリを現在の子ノードに移動させるか否かを判定する。例えば、対象エントリの複数の特徴量データと当該類似度閾値以上となる類似度を示す特徴量データを含むエントリは、子ノードに移動させるべきと判定される。但し、対象エントリの複数の特徴量データの中の1つと当該類似度閾値以上となる類似度を示す特徴量データを含むエントリは、子ノードに移動させると判定されてもよい。

20

【0126】

次に、第2実施形態における類似検索方法について図7を用いて説明する。第2実施形態では、(S92)及び(S95)において、複数の特徴量データが処理対象とされる点において、第1実施形態と異なる。

【0127】

(S92)では、検索装置1は、対象ノードが中間ノードである場合、対象ノードの各エントリの複数の特徴量データ f_e とクエリデータ f_q との各類似度 S をそれぞれ算出する。(S95)では、検索装置1は、対象ノードに含まれるエントリの中から、類似度閾値と類似度条件 δ との小さい方の値未満の複数の類似度 S を示す複数の特徴量データ f_e を示すエントリを抽出対象から除外する。

30

【0128】

〔第2実施形態の作用及び効果〕

第2実施形態では、子ノードへのポインタを含む中間ノードのエントリは、複数の特徴量データを示す。即ち、第2実施形態では、子ノードで示される特徴量データ集合の代表となる複数の特徴量データが選出され、当該複数の特徴量データが親ノードのエントリに設定される。そして、子ノードで示される特徴量データの集合は、親ノードのエントリに含まれる全ての特徴量データ又はいずれか1つの特徴量データとの類似度がその類似度閾値以上を示す。

40

【0129】

このように、第2実施形態によれば、親ノードのエントリで示される複数の特徴量データが用いられることにより、子ノードで示される特徴量データ集合との類似度が判定されるため、親ノードのエントリにより、子ノードで示されるデータ集合の類似度範囲を、第1実施形態に比べより厳密に定義することができる。

【0130】

また、類似度の値が最大となる特徴量データの組み合わせを示す複数エントリを複数代表エントリとして選択する方法によれば、当該複数エントリで示される複数の特徴量データの分散が大きい場合でも、親ノードのエントリにより、子ノードで示されるデータ集合

50

の類似度範囲を厳密に定義することができる。一方、類似度の値が最小となる特徴量データの組み合わせを示す複数エントリを複数代表エントリとして選択する方法によれば、当該複数エントリで示される複数の特徴量データが密集する場合でも、親ノードのエントリにより、子ノードで示されるデータ集合の類似度範囲を厳密に定義することができる。

【0131】

このように、親ノードのエントリにより、子ノードで示されるデータ集合の類似度範囲を厳密に定義することができるため、第2実施形態によれば、類似度木のバランスを取り易くなり、ひいては、検索効率を向上させることができる。

【0132】

以下に実施例を挙げ、上述の各実施形態を更に詳細に説明する。本発明は以下の実施例から何ら限定を受けない。

【実施例1】

【0133】

実施例1は、上述の第1実施形態に対応する。以下、実施例1における検索装置1の動作及びインデックス生成方法について、図5に沿って説明する。

【0134】

図9から図18は、実施例1におけるインデックス生成過程を概念的に示す図である。実施例1では、図9に示されるように、子ノードへのポインタ、類似度閾値(0.1)及び特徴量データf1が設定されたエントリを持つルートノードが既に生成されていると仮定する。また、エントリ閾値が3に設定されていると仮定する。また、以降、特徴量データを単にデータと表記する。

【0135】

検索装置1は、インデックス対象となるデータf5を取得する(S50)。その後、検索装置1は、データf1とデータf5との類似度(0.8)が類似度閾値(0.1)より大きいため、ルートノードの子ノードであるノードN1を対象ノードに設定する(S53)。ここで、ノードN1は葉ノードであるため(S54; YES)、検索装置1は、データf5を含むエントリを対象ノードN1に追加する(S55)。

【0136】

検索装置1は、対象ノードのエントリ数(4)が所定エントリ閾値(3)を超えているため(S56; YES)、溢れ調整を行う。溢れ調整において、検索装置1は、対象ノードN1の代表エントリを選択する(S57)。図10の例によれば、データf2を含むエントリが代表エントリに選択される。更に、検索装置1は、新たな類似度閾値(0.8)を決定する(S58)。

【0137】

検索装置1は、代表エントリのデータf2と、他の各エントリのデータf3、f4及びf5の各々との類似度をそれぞれ算出する(S59)。図10の例に示されるような各類似度がそれぞれ算出される。これにより、検索装置1は、図11に示されるように、類似度閾値(0.8)以上の類似度を示すデータf3及びf4を持つ2つのエントリを含む子ノードN2を生成する(S60)。更に、検索装置1は、対象ノードN1の代表エントリに子ノードN2へのポインタと類似度閾値(0.8)とを設定する(S61)。検索装置1は、データf5をデータベース35に格納すると共に、更新された類似度木データをデータベース35に格納する(S62)。

【0138】

続いて、図11に示されるように、検索装置1は、新たなデータf6を取得する(S50)。検索装置1は、データf1とデータf6との類似度(0.9)が類似度閾値(0.1)より大きいため、ルートノードの子ノードであるノードN1を対象ノードに設定する(S53)。ここで、ノードN1は中間ノードであるため(S54; NO)、検索装置1は、対象ノードN1に含まれる、子ノードへのポインタを含むエントリについて、そのエントリに含まれるデータf2とデータf6との類似度を算出する。図11の例では、データf2とデータf6との類似度は0.85であるため、検索装置1は、類似度閾値(0.

10

20

30

40

50

8) 以上の類似度を示すデータ f 2 を含むエントリを特定する (S 5 2)。

【0139】

検索装置 1 は、その特定されたエントリのポインタが示す子ノード N 2 を対象ノードに設定し (S 5 3)、図 1 2 に示されるように、データ f 6 を含むエントリをその対象ノード N 2 に追加する (S 5 5)。ここで、対象ノード N 2 のエントリ数 (3) は、所定エントリ閾値 (3) を超えていないため (S 5 6; NO)、検索装置 1 は、データ f 6 及び更新された類似度木データをデータベース 3 5 に格納する (S 6 2)。

【0140】

続いて、図 1 3 に示される場面について説明する。ここでは、データ f 7 がノード N 1 に追加された状態で、検索装置 1 が新たなデータ f 8 を取得する (S 5 0)。検索装置 1 は、データ f 1 とデータ f 8 との類似度 (0. 7) が類似度閾値 (0. 1) より大きいため、ルートノードの子ノードであるノード N 1 を対象ノードに設定する (S 5 3)。

【0141】

ここで、ノード N 1 は中間ノードであるため (S 5 4; NO)、検索装置 1 は、対象ノード N 1 に含まれる、子ノードへのポインタを含むエントリについて、そのエントリに含まれるデータ f 2 とデータ f 8 との類似度を算出する。図 1 3 の例では、データ f 2 とデータ f 8 との類似度は 0. 75 であり、類似度閾値 (0. 8) 以上の類似度を示すデータを含むエントリが対象ノード N 1 に存在しない。検索装置 1 は、該当するエントリが対象ノード N 1 に存在しないため (S 5 2; NO)、図 1 3 に示されるように、データ f 8 を含むエントリを対象ノード N 1 に追加する (S 5 5)。

【0142】

検索装置 1 は、対象ノード N 1 のエントリ数 (4) が所定エントリ閾値 (3) を超えているため (S 5 6; YES)、溢れ調整を行う。図 1 3 の例では、検索装置 1 は、対象ノード N 1 に含まれる全エントリを母集団として、その母集団の中から代表エントリを選択する (S 5 7)。図 1 3 の例によれば、データ f 5 を含むエントリが代表エントリに選択される。更に、検索装置 1 は、新たな類似度閾値 (0. 8) を決定する (S 5 8)。

【0143】

検索装置 1 は、代表エントリのデータ f 5 と、他の各エントリのデータ f 2、f 7 及び f 8 の各々との類似度をそれぞれ算出する (S 5 9)。図 1 4 の例に示されるような各類似度がそれぞれ算出される。これにより、検索装置 1 は、図 1 4 及び図 1 5 に示されるように、類似度閾値 (0. 8) 以上の類似度を示すデータ f 8 を持つ 1 つのエントリを含む子ノード N 3 を生成する (S 6 0)。更に、検索装置 1 は、対象ノード N 1 におけるデータ f 5 を含む代表エントリに、子ノード N 3 へのポインタと類似度閾値 (0. 8) とを設定する (S 6 1)。このとき、検索装置 1 は、ノード N 2 へのポインタを含むエントリについてノード N 1 に残ることを決めたため、そのエントリに含まれるそのポインタや類似度閾値 (0. 8) をそのままとする。これにより、更新された類似度木は、図 1 5 に示されるようになる。検索装置 1 は、データ f 8 をデータベース 3 5 に格納すると共に、更新された類似度木データをデータベース 3 5 に格納する (S 6 2)。

【0144】

図 1 6 及び図 1 7 A は、図 1 4 及び図 1 5 に示される例とは異なるインデックス生成例を示す図である。図 1 4 の例では、代表エントリがデータ f 5 を含むエントリに決められ、新たな類似度閾値が 0. 8 に決められた。図 1 6 は、新たな類似度閾値が 0. 75 に決められる例を示す。この場合、図 1 6 の例に示されるような各類似度がそれぞれ算出される。

【0145】

この場合、検索装置 1 は、類似度閾値 (0. 75) 以上の類似度を示すデータ f 2 及び f 8 を持つ 2 つのエントリを子ノードへ移すことを決める。しかし、この場合、データ f 2 を含むエントリは、子ノード N 2 とリンクされている。そこで、検索装置 1 は、図 1 7 に示されるように、データ f 2 を含むエントリに設定されている子ノード N 2 へのポインタと類似度閾値 (0. 8) とをそのまま維持しつつ、データ f 2 及び f 8 を含む 2 つのエ

10

20

30

40

50

ントリを含む子ノードN4を生成し(S60)、親ノードとなるノードN1から、データf2及びf8を含む2つのエントリを削除する。更に、検索装置1は、対象ノードN1におけるデータf5を含む代表エントリに、子ノードN4へのポインタと類似度閾値(0.75)とを設定する(S61)。結果、更新された類似度木は、図17Aに示されるようになる。図17Aに示されるように、既に設定されているリンク関係を維持することで、ノードの溢れ調整を簡易処理で実現することができる。

【0146】

図17Bは、図17Aに示される例とは異なるインデックス生成例を示す図である。検索装置1は、ノードの溢れ調整において、既に子ノードとリンクされているエントリを子ノードへ移すことを決めた場合、図17Bに示されるように、既に設定されているリンク関係を解除するようにしてもよい。この場合、検索装置1は、図17Bに示されるように、データf2及びf8を含む2つのエントリと共に、子ノードN2に含まれる3つのエントリ(データf3、f4及びf6を含む)を含む子ノードN4を生成し(S60)、親ノードとなるノードN1から、データf2及びf8を含む2つのエントリを削除する。

【0147】

図16、図17A及び図17Bの例において、ノードN1の溢れ調整時に、代表エントリに含まれるデータf5と、ノードN2に含まれていたデータf3、f4及びf6の各々との類似度は、計算されなくてもよい。これは、上述の規則に基づいている。即ち、データf2は、データf3、f4及びf6の各々と類似度閾値(0.8)以上の類似度を示す代表データであり、かつ、代表データf2とデータf5とが新たな類似度閾値(0.75)以上の類似度を示す場合、代表データf2のデータ集合(f3、f4及びf6)とデータf5とについても新たな類似度閾値(0.75)以上の類似度を示す可能性が高いからである。

【0148】

図18は、図15に示される類似度木が既に生成されている状態で、新たなデータf9が取得された場合のインデックス生成過程を概念的に示す図である。この場合、検索装置1は、対象ノードN1に含まれる、子ノードへのポインタを含む2つのエントリについて、そのエントリに含まれるデータf2及びf5と、データf9との各類似度をそれぞれ算出する。図18の例では、データf2とデータf9との類似度は0.85であり、その類似度がデータf2のエントリの類似度閾値(0.8)より大きい。更に、データf5とデータf9との類似度は0.86であり、その類似度がデータf5のエントリの類似度閾値(0.8)より大きい。即ち、(S52)に該当するエントリが複数存在する(S52; YES)。

【0149】

この場合、検索装置1は、(S52)の条件を満たすエントリであって、かつ、データf9との類似度が最大となるエントリ、即ち、データf5を含むエントリを特定する。検索装置1は、データf5を含むエントリのポインタが示す子ノードN3を対象ノードに設定する(S53)。検索装置1は、データf9を含むエントリを対象ノードN3に追加する(S55)。

【実施例2】

【0150】

実施例2は、上述の第2実施形態に対応する。以下、実施例2における検索装置1の動作及びインデックス生成方法について、図5に沿って説明する。実施例2では、子ノードへのポインタを含むエントリが、子ノードで示される特徴量データ集合の代表となる2つの特徴量データを含む例が示される。また、子ノードへのポインタを含むエントリの2つの特徴量データとの各類似度がそれぞれ類似度閾値以上となるデータが、その子ノードへの追加対象とされる例が示される。

【0151】

図19及び図20は、実施例2におけるインデックス生成過程を概念的に示す図である。図19に示されるように、子ノードへのポインタ、類似度閾値(0.2)及びデータf

10

20

30

40

50

1 及び f 2 が設定されたエントリを持つルートノードが既に生成されており、ノード N 1 には、データ f 3、f 4 及び f 5 を含む 3 つのエントリが設定されている。

【0152】

この状態で、検索装置 1 は、データ f 6 を取得する (S 5 0)。検索装置 1 は、データ f 1 及び f 2 の各々と、データ f 6 との各類似度 (0. 9) (0. 8 5) が類似度閾値 (0. 2) よりそれぞれ大きいため、ルートノードの子ノードであるノード N 1 を対象ノードに設定する (S 5 3)。

【0153】

ここで、ノード N 1 は葉ノードであるため (S 5 4 ; N O)、検索装置 1 は、データ f 6 を含むエントリを対象ノード N 1 に追加する (S 5 5)。検索装置 1 は、対象ノード N 1 のエントリ数 (4) が所定エントリ閾値 (3) を超えているため (S 5 6 ; Y E S)、溢れ調整を行う。

【0154】

この溢れ調整において、検索装置 1 は、図 1 9 に示されるように、対象ノード N 1 に含まれる全エントリの全データの中から 2 つのデータを代表として選択する (S 5 7)。ここでは、上述の第 1 の選択方法が用いられることにより、検索装置 1 は、類似度の値が最大となる 2 つのデータを代表として選択する。図 1 9 の例では、データ f 3 及びデータ f 4 が代表に選択される。更に、検索装置 1 は、新たな類似度閾値 (0. 6) を決定する (S 5 8)。

【0155】

検索装置 1 は、代表として選択されたデータ f 3 及び f 4 の各々と、他のエントリのデータ f 5 及び f 6 の各々と、類似度をそれぞれ算出する (S 5 9)。ここでは、図 2 0 の例に示されるような各類似度がそれぞれ算出される。これにより、検索装置 1 は、図 2 0 に示されるように、データ f 3 及び f 4 の各々と類似度閾値 (0. 6) 以上の類似度を示すデータ f 5 を持つ 1 つのエントリを含む子ノード N 2 を生成する (S 6 0)。図 2 0 の例では、データ f 6 とデータ f 4 との類似度 (0. 6) は、類似度閾値 (0. 6) 以上となるが、データ f 6 とデータ f 3 との類似度 (0. 5) が類似度閾値 (0. 6) 未満となるため、データ f 6 は、ノード N 1 に残される。

【0156】

検索装置 1 は、対象ノード N 1 におけるデータ f 3 及び f 4 を含む 1 つのエントリを生成し、そのエントリに、子ノード N 2 へのポインタと類似度閾値 (0. 6) とを設定する (S 6 1)。これにより、更新された類似度木は、図 2 0 に示されるようになる。

【0157】

[変形例]

上述の各実施形態及び各実施例では、代表となる特徴量データを示すエントリは、親ノードのみに含まれるが、親ノードと子ノードとの両方にそれぞれ含まれていてもよい。

【0158】

また、上述の各実施形態では、検索装置 1 は、1 つの装置 (コンピュータ) として実現されたが、複数の装置として実現されてもよい。上述の各実施形態は、例えば、類似度木を生成するインデックス生成装置と、類似検索を行う検索装置とから構成されてもよい。この場合、インデックス生成装置は、対象取得部 2 0、インデックス生成部 2 1、閾値決定部 2 2 及び類似度算出部 2 5 を含み、検索装置は、検索部 3 0 及びデータベース 3 5 を含む。更に、各実施形態は、インデックス生成装置、検索装置及びデータベース装置から構成されてもよい。

【0159】

また、上述の説明で用いた複数のフローチャートでは、複数の工程 (処理) が順番に記載されているが、各実施形態で実行される工程の実行順序は、その記載の順番に制限されない。各実施形態では、図示される工程の順番を内容的に支障のない範囲で変更することができる。また、上述の各実施形態及び各変形例は、内容が相反しない範囲で組み合わせることができる。

10

20

30

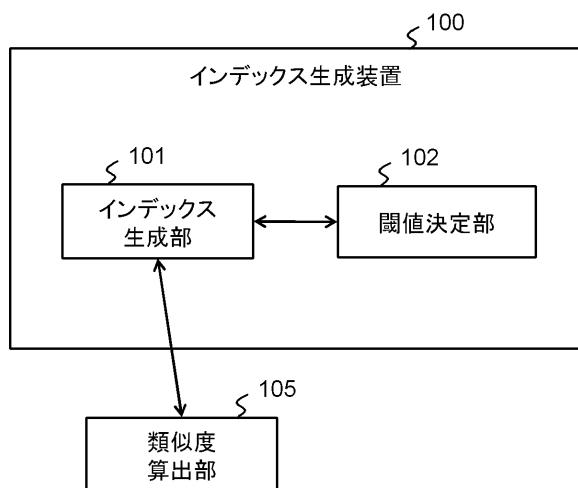
40

50

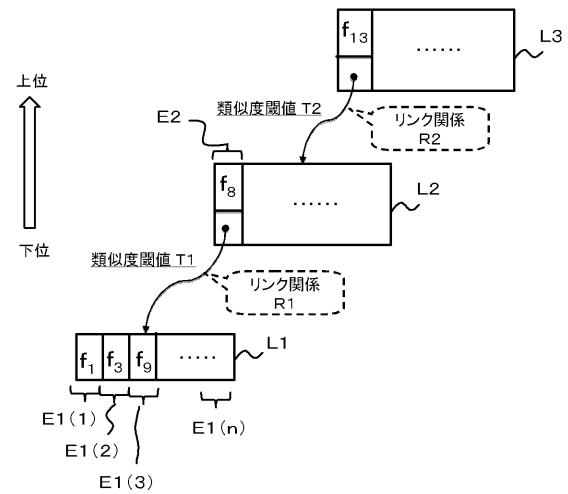
【0160】

この出願は、2013年1月11日に出願された日本出願特願2013-003970号を基礎とする優先権を主張し、その開示の全てをここに取り込む。

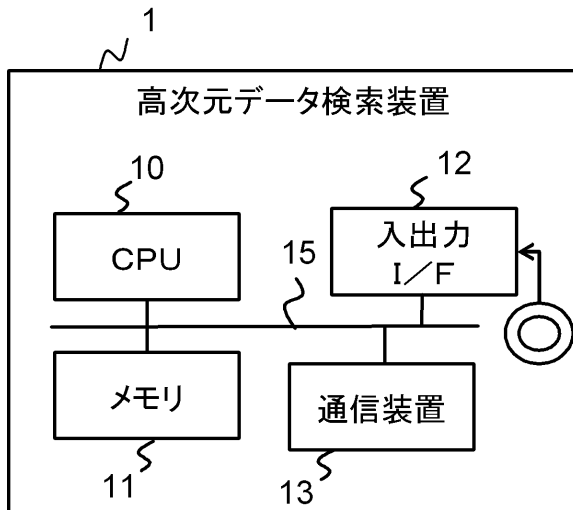
【図1】



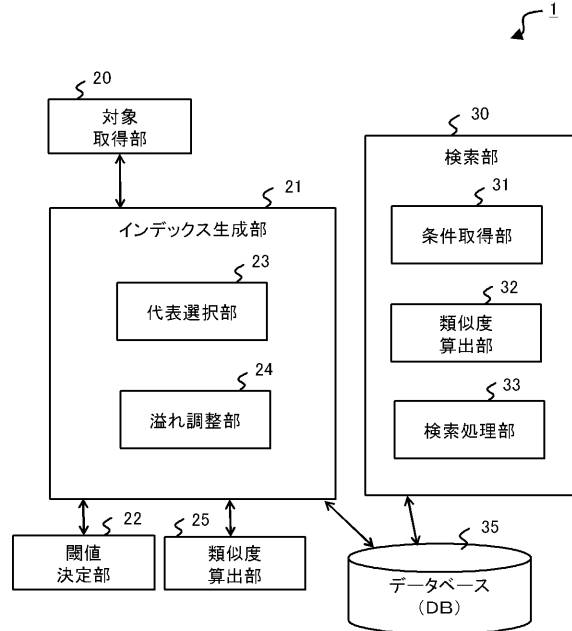
【図2】



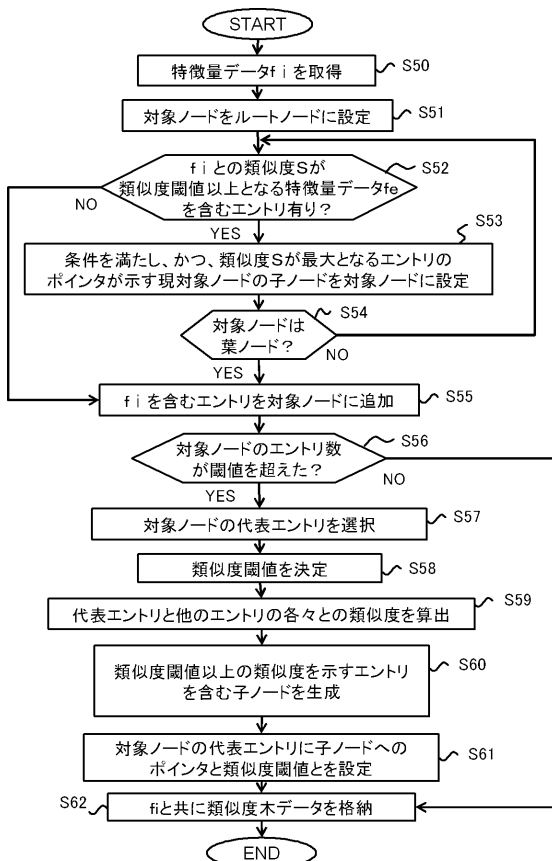
【図 3】



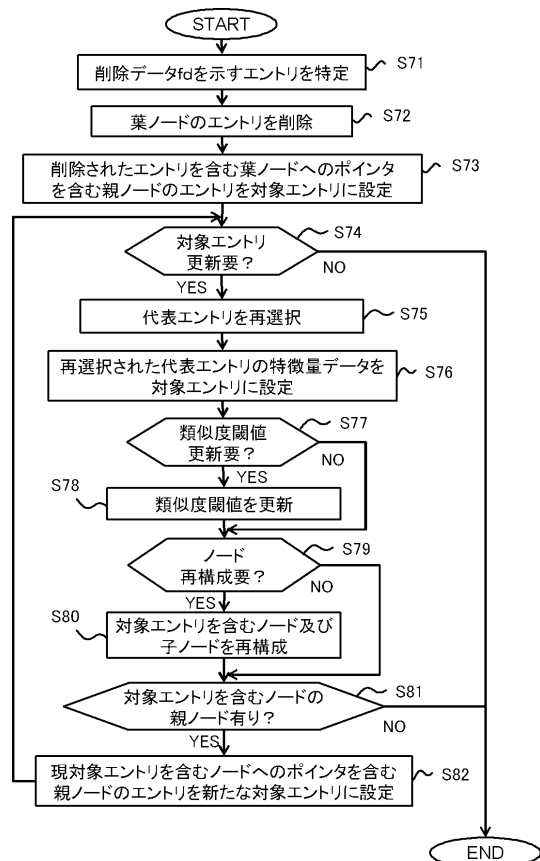
【図 4】



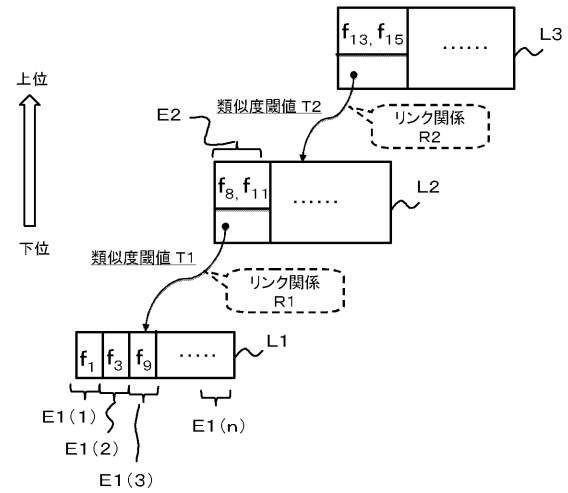
【図 5】



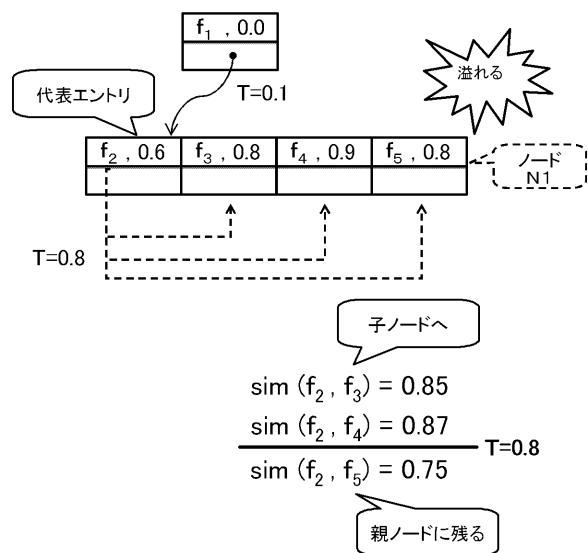
【図 6】



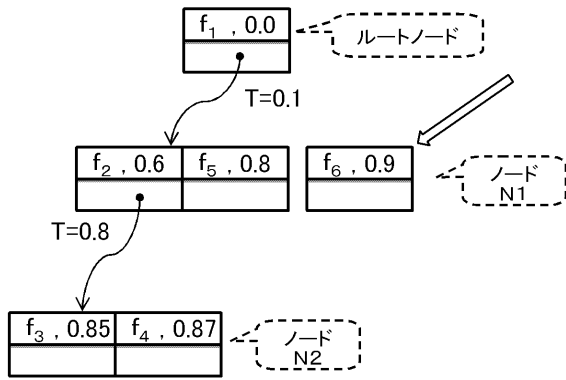
【图 8】



【図 10】

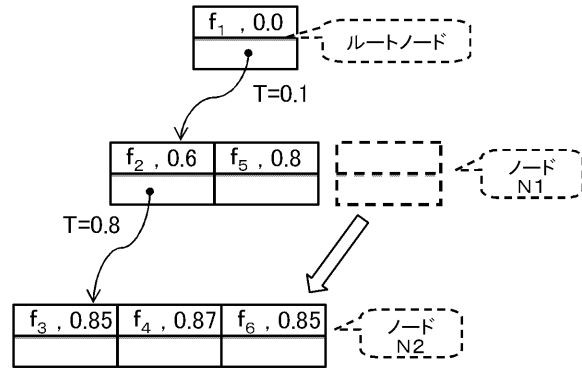


【図 1 1】



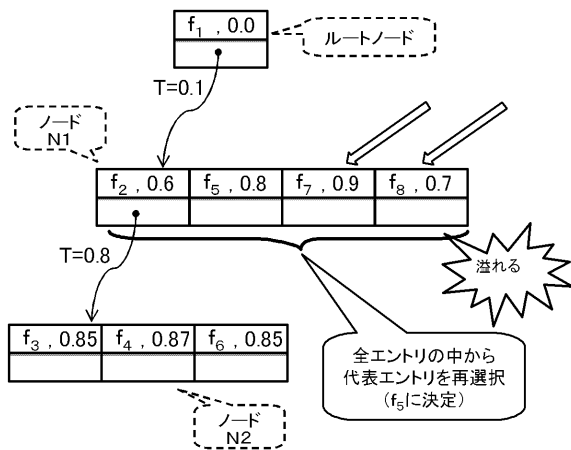
$$\text{sim}(f_2, f_6) = 0.85$$

【図 1 2】



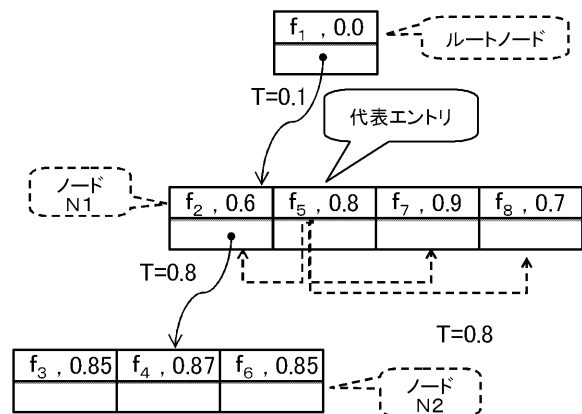
$$\text{sim}(f_2, f_6) = 0.85$$

【図 1 3】



$$\text{sim}(f_2, f_8) = 0.75$$

【図 1 4】



$$\text{sim}(f_5, f_2) = 0.75$$

$$\text{sim}(f_5, f_7) = 0.7$$

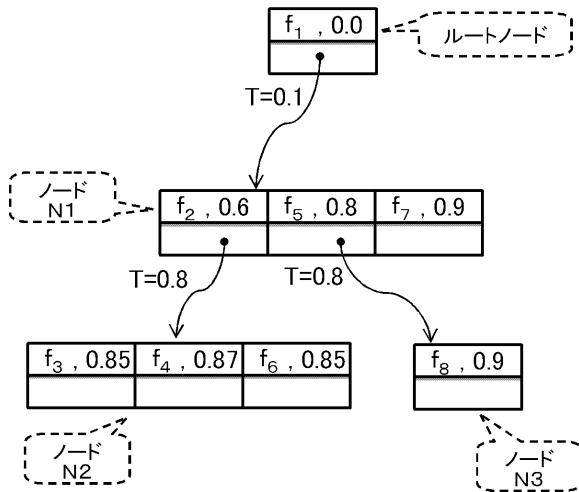
$$\text{sim}(f_5, f_8) = 0.9$$

T=0.8

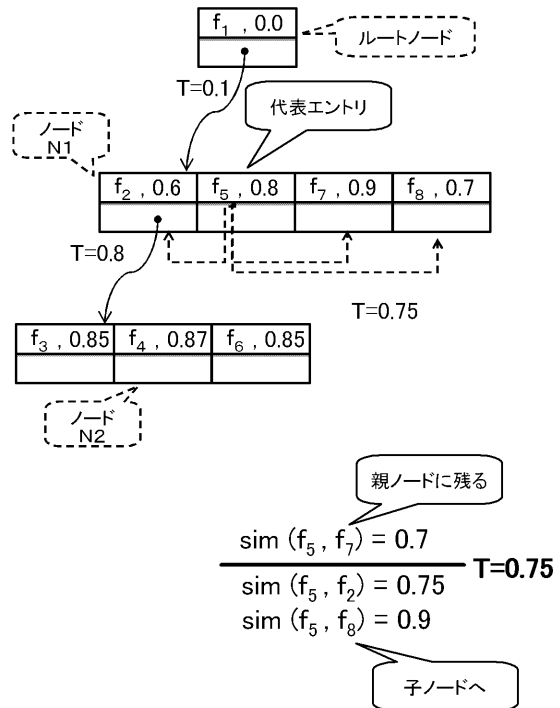
親ノードに残る

子ノードへ

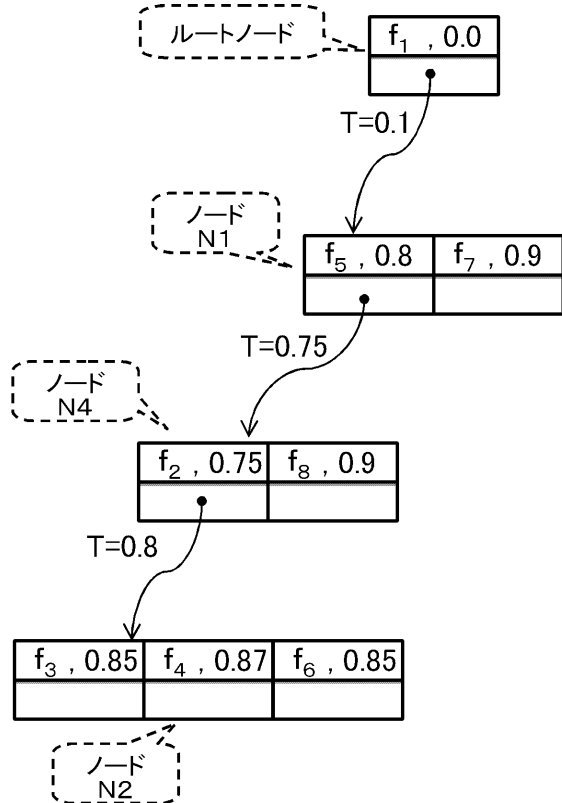
【図 15】



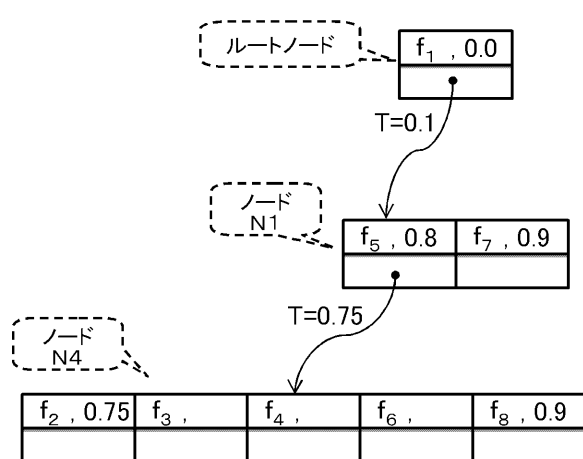
【図 16】



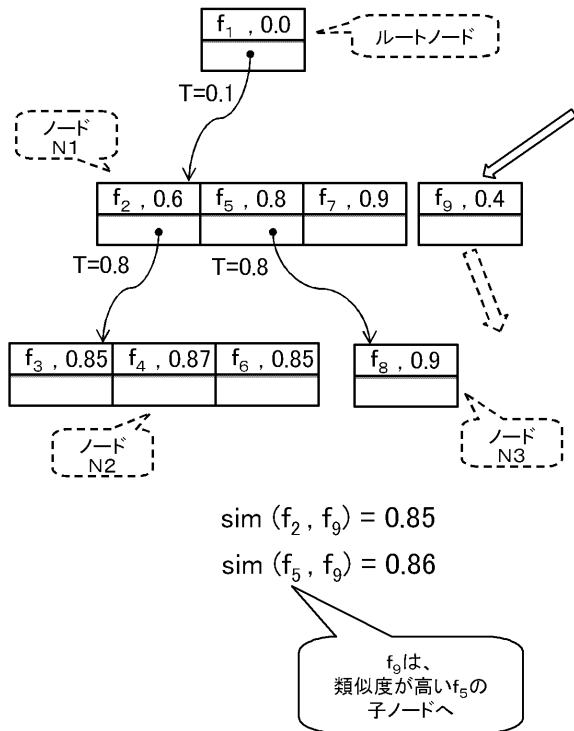
【図 17 A】



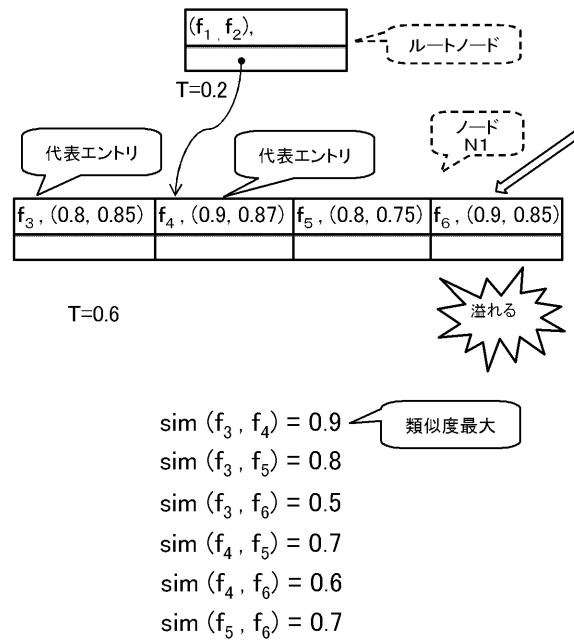
【図 17 B】



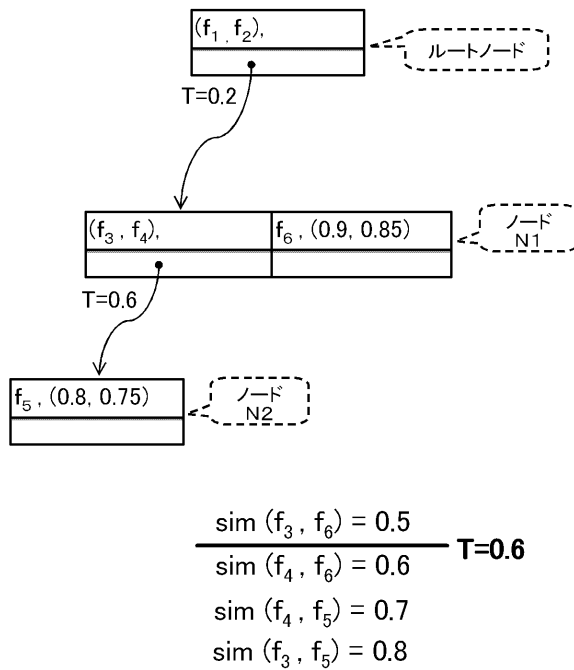
【図 18】



【図 19】



【図 20】



フロントページの続き

(56)参考文献 特開2012-234446 (JP, A)
特開2008-217246 (JP, A)
特開2008-224343 (JP, A)
特開2002-163272 (JP, A)
特開2000-112973 (JP, A)
米国特許第8352465 (US, B1)

(58)調査した分野(Int.Cl., DB名)

I P C G 0 6 F 1 7 / 3 0