

明 細 書

発明の名称：

類似データ検索装置、類似データ検索方法、及びコンピュータ読み取り可能な記録媒体

技術分野

[0001] 本発明は、類似データ検索装置、特には、文字列から変換された集合の集合間類似度に基づいて検索を行う、類似データ検索装置、及び類似データ検索方法に関し、更には、これらを実現するための類似データ検索性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に関する。

背景技術

[0002] 類似データ検索は、クラスタリング、重複データ照合、文字列ソフトマッチングなどに幅広く応用可能な、基本的で、且つ重要なデータ処理である。類似データ検索の具体的な方法としては、単純に、対象となる全てのデータの間の類似度を計算し、類似度に基づいて検索を行なう方法が挙げられるが、この方法では、データ量が多くなると処理時間も膨大になる。

[0003] 例えば、類似度が一定以上の全てのデータペアをデータベース内から検索する場合、与えられたデータが N 個とすると、 $(N(N-1)/2)$ 回の類似度計算が必要となる。これは、例えば、一回の類似度計算にかかる時間が、 0.001 ミリ秒とした場合、データの個数 N が $100,000$ 個であるならば、約 50 億回の類似度計算が必要となるため、計算には約 14 日を要することになる。

[0004] このため、非特許文献 1 は、類似度が一定以上の全データペアを高速に検索して処理時間を短縮するシステムを開示している。非特許文献 1 に開示されたシステムは、まず、文字列をその特徴の集合に変換し、集合のサイズを集合内の要素の個数と定義し、集合を同じサイズで分けて、同じサイズの集合毎に、転置インデックスを作成する。次に、非特許文献 1 に開示されたシステムは、検索時に、入力される集合のサイズと類似度閾値とから、検索すべき転置インデックスのサイズの最小値と最大値とを同定し、同定したサ

イズの範囲にある転置インデックスのみを検索する。

- [0005] 具体的には、非特許文献1は、その表1において、検索が要求された集合を X とし、ジャックカード係数 $(X \cap Y) / |X \cup Y|$ が閾値 α 以上の場合は、 $\alpha |X|$ 以上、 $|X| / \alpha$ 以下のサイズに該当する転置インデックスのみを検索すれば十分であることを開示している。従って、非特許文献1に開示されたシステムは、転置インデックスを集合のサイズ毎に作成し、検索条件（検索要求）から決まるサイズの上限と下限とを用いて、検索すべき転置インデックスを同定する。この結果、無駄な検索が省かれるので、検索処理の高速化が可能となる。

先行技術文献

非特許文献

- [0006] 非特許文献1：岡崎直観，辻井潤一「集合間類似度に対する簡潔かつ高速な類似文字列検索アルゴリズム」、自然言語処理、Vol. 18、No. 2、pp. 89-117、2011年。

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、非特許文献1に開示されたシステムには、検索対象となるデータ中に同じサイズの集合が少ない場合に、検索効率が悪化するという問題がある。
- [0008] その理由は、同じサイズの集合毎に転置インデックスが作成されているため、同じサイズのデータが少ない場合には、各転置インデックスから検索可能な集合が少なくなり、同じ結果を得るために必要な転置インデックスへの検索回数が増えることにある。
- [0009] 平均的には、同じサイズの集合の数と、同じ結果を取得するために必要な転置インデックスに対する検索回数とは反比例の関係となる。そのため、外部の記憶装置へのランダムアクセスが行われる場合など、転置インデックスに対する検索コストが高い場合には、特に検索効率が悪化する。

[0010] [発明の目的]

本発明の目的の一例は、上記問題を解消し、検索対象となるデータ中に同じサイズの集合が少ない場合であっても、転置インデックスにおける検索回数の増加による検索効率の低下を抑制し得る、類似データ検索装置、類似データ検索方法、及びコンピュータ読み取り可能な記録媒体を提供することにある。

課題を解決するための手段

[0011] 上記目的を達成するため、本発明の一側面における類似データ検索装置は、検索対象となるデータ及び検索条件となるデータとして集合を用いて検索を行うための類似データ検索装置であって、

検索に使用する転置インデックスの生成のため、生成予定の各転置インデックスに含まれる検索対象の集合の数が設定個数以上になるように、前記生成予定の転置インデックス毎に、前記検索対象の集合のサイズの範囲を決定し、決定した前記サイズの範囲毎に、前記検索対象の集合を分けて、前記転置インデックスを生成する、転置インデックス生成部と、

検索条件の集合のサイズと、前記検索条件の集合と前記検索対象の集合との類似度に対して設定された閾値と、に基づいて、前記類似度が閾値以上となるために必要な、前記検索対象の集合のサイズの条件を求め、前記転置インデックスのうち、それに含まれる集合のサイズの最小値が前記条件を満たす転置インデックス以外を、検索不要な転置インデックスとして同定する、不要転置インデックス同定部と、

同定された前記検索不要な転置インデックス以外の転置インデックスに対して、前記検索条件の集合を適用して、検索を実行する、データ検索部と、を備えることを特徴とする。

[0012] また、上記目的を達成するため、本発明の一側面における類似データ検索方法は、検索対象となるデータ及び検索条件となるデータとして集合を用いて検索を行うための方法であって、

(a) 検索に使用する転置インデックスの生成のため、生成予定の各転置イ

ンデックスに含まれる検索対象の集合の数が設定個数以上になるように、前記生成予定の転置インデックス毎に、前記検索対象の集合のサイズの範囲を決定し、決定した前記サイズの範囲毎に、前記検索対象の集合を分けて、前記転置インデックスを生成する、ステップと、

(b) 検索条件の集合のサイズと、前記検索条件の集合と前記検索対象の集合との類似度に対して設定された閾値と、に基づいて、前記類似度が閾値以上となるために必要な、前記検索対象の集合のサイズの条件を求め、前記転置インデックスのうち、それに含まれる集合のサイズの最小値が前記条件を満たす転置インデックス以外を、検索不要な転置インデックスとして同定する、ステップと、

(c) 前記(b)のステップで同定された前記検索不要な転置インデックス以外の転置インデックスに対して、前記検索条件の集合を適用して、検索を実行する、ステップと、

を有することを特徴とする。

[0013] 更に、上記目的を達成するため、本発明の一側面におけるコンピュータ読み取り可能な記録媒体は、コンピュータによって、検索対象となるデータ及び検索条件となるデータとして集合を用いて検索を行う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記録媒体であって、前記コンピュータに、

(a) 検索に使用する転置インデックスの生成のため、生成予定の各転置インデックスに含まれる検索対象の集合の数が設定個数以上になるように、前記生成予定の転置インデックス毎に、前記検索対象の集合のサイズの範囲を決定し、決定した前記サイズの範囲毎に、前記検索対象の集合を分けて、前記転置インデックスを生成する、ステップと、

(b) 検索条件の集合のサイズと、前記検索条件の集合と前記検索対象の集合との類似度に対して設定された閾値と、に基づいて、前記類似度が閾値以上となるために必要な、前記検索対象の集合のサイズの条件を求め、前記転置インデックスのうち、それに含まれる集合のサイズの最小値が前記条件を

満たす転置インデックス以外を、検索不要な転置インデックスとして同定する、ステップと、

(c) 前記(b)のステップで同定された前記検索不要な転置インデックス以外の転置インデックスに対して、前記検索条件の集合を適用して、検索を実行する、ステップと、

を実行させる命令を含む、プログラムを記録していることを特徴とする。

発明の効果

[0014] 以上のように本発明によれば、検索対象となるデータ中に同じサイズの集合が少ない場合であっても、転置インデックスにおける検索回数の増加による検索効率の低下を抑制することができる。

図面の簡単な説明

[0015] [図1]図1は、本発明の実施の形態1における類似データ検索装置の構成を示すブロック図である。

[図2]図2は、本発明の実施の形態1における類似データ検索装置の動作を示すフロー図である。

[図3]図3は、類似度としてジャカード係数が用いられる場合のステップA1の一例を説明する図である。

[図4]図4は、集合のサイズの範囲を求めるための数式の一例を示す図である。

[図5]図5は、類似度が閾値以上となる場合と等価であるとされる条件の一例を示す図である。

[図6]図6は、本発明の実施の形態2における類似データ検索装置の構成を示すブロック図である。

[図7]図7は、本発明の実施の形態2における類似データ検索装置の動作を示すフロー図である。

[図8]図8は、本発明の実施の形態1及び2における類似データ検索装置を実現するコンピュータの一例を示すブロック図である。

発明を実施するための形態

[0016] (実施の形態 1)

以下、本発明の実施の形態 1 における、類似データ検索装置、類似データ検索方法、及び類似データ検索用プログラムについて、図 1 ～図 5 を参照しながら説明する。

[0017] [装置構成]

最初に、本実施の形態 1 における類似データ検索装置の構成について図 1 を用いて説明する。図 1 は、本発明の実施の形態 1 における類似データ検索装置の構成を示すブロック図である。

[0018] 図 1 に示す本実施の形態 1 における類似データ検索装置 2 は、検索対象となるデータ及び検索条件となるデータとして集合を用いて検索を行う装置である。図 1 に示すように、類似データ検索装置 2 は、転置インデックス生成部 20 と、不要転置インデックス同定部 21 と、データ検索部 22 とを備えている。

[0019] 転置インデックス生成部 20 は、検索に使用する転置インデックスを生成する。このため、まず、転置インデックス生成部 20 は、生成予定の各転置インデックスに含まれる検索対象の集合の数が設定個数以上になるように、生成予定の転置インデックス毎に、検索対象の集合のサイズの範囲を決定する。そして、転置インデックス生成部 20 は、決定したサイズの範囲毎に、検索対象の集合を分けて、転置インデックスを生成する。

[0020] 不要転置インデックス同定部 21 は、まず、検索条件の集合のサイズと、検索条件の集合と検索対象の集合との類似度に対して設定された閾値と、に基づいて、類似度が閾値以上となるために必要な、検索対象の集合のサイズの条件を求める。次に、不要転置インデックス同定部 21 は、転置インデックスのうち、それに含まれる集合のサイズの最小値が条件を満たす転置インデックス以外を、検索不要な転置インデックスとして同定する。

[0021] データ検索部 22 は、不要転置インデックス同定部 21 によって同定された検索不要な転置インデックス以外の転置インデックスを特定し、この特定した転置インデックスに対して、検索条件の集合を適用して、検索を実行す

る。

[0022] このように、本実施の形態では、各転置インデックスにおいて、それに含まれる集合のサイズの範囲が決められ、このサイズの範囲に基づいて、検索条件の集合の検索先として不適切な転置インデックスが同定される。そして、同定された転置インデックスを除いて、検索が行なわれる。この結果、検索対象となるデータ中に同じサイズの集合が少ない場合であっても、転置インデックスにおける検索回数の増加による検索効率の低下を抑制することができる。

[0023] ここで、本実施の形態 1 における類似データ検索装置 2 の構成について更に具体的に説明する。図 1 に示すように、類似データ検索装置 2 は、データ記憶装置 1 と、入力装置 3 と、出力装置 4 とに接続され、これらと共に、類似データ検索システム 30 を構築している。

[0024] データ記憶装置 1 は、検索対象の集合で構成された検索対象データ 10 と、検索対象の集合に含まれる要素に予め付与された重要度を特定する要素重要度データ 11 と、を記憶している（後述の図 3 参照）。

[0025] また、入力装置 3 は、類似データ検索装置 2 に、検索条件の集合、及び類似度の閾値といったデータを入力する装置である。入力装置 3 としては、キーボード等の入力機器、類似データ検索装置 2 にネットワークを介して接続された端末装置、等が挙げられる。

[0026] 出力装置 4 は、検索結果の出力先となる装置である。出力装置 4 としては、表示装置、プリンタ等が挙げられるが、その他に、類似データ検索装置 2 にネットワークを介して接続された端末装置も挙げられる。なお、入力装置 3 と出力装置 4 とは、同一の端末装置であっても良い。

[0027] また、本実施の形態 1 において、「集合」は、1 つ以上の要素で構成されていれば良く、各要素には、上述したように、予め重要度が付与されていても良い（後述の図 3 参照）。なお、非特許文献 1 に記載されているように、集合は、要素となる文字 tri-gram から構成されていても良い。

[0028] また、本実施の形態 1 において、検索条件の集合と検索対象の集合との「

類似度」は、例えば、検索対象の集合をD、検索条件（検索要求）の集合をQ、重要度の重みを返す関数を $w(\cdot)$ とすると、これらを用いた数式によって計算される。例えば、集合間の類似度は、Dから見たQの重複度 $\text{overlap}(Q, D)$ 、Qから見たDの重複度 $\text{overlap}(D, Q)$ 、コサイン類似度 $\text{cosine}(Q, D)$ 、ダイス係数 $\text{dice}(Q, D)$ 、及びジャカード係数 $\text{jaccard}(Q, D)$ のうちいずれかによって計算される。

[0029] 具体的には、集合間の類似度は、下記の数1～数5のいずれかを用いることで計算することができる。但し、本実施の形態において、類似度は、下記の式で計算されるものに限られるわけではない。本実施の形態では、類似度は、集合の大きさによって条件を規定できるものであれば良く、特に限定なく適用することができる。

[0030] [数1]

$$\text{overlap}(Q, D) = \frac{\sum_{e \in Q \cap D} w(e)}{\sum_{e \in D} w(e)}$$

[0031] [数2]

$$\text{overlap}(D, Q) = \frac{\sum_{e \in Q \cap D} w(e)}{\sum_{e \in Q} w(e)}$$

[0032] [数3]

$$\text{cosine}(Q, D) = \frac{\sum_{e \in Q \cap D} w(e)}{\sqrt{\sum_{e \in Q} w(e)^2} \sqrt{\sum_{e \in D} w(e)^2}}$$

[0033] [数4]

$$\text{dice}(Q, D) = \frac{2 \sum_{e \in Q \cap D} w(e)}{\sum_{e \in Q} w(e) + \sum_{e \in D} w(e)}$$

[0034] [数5]

$$jaccard(Q, D) = \frac{\sum_{e \in Q \cap D} w(e)}{\sum_{e \in Q \cup D} w(e)}$$

[0035] [装置動作]

次に、本発明の実施の形態1における類似データ検索装置2の動作について図2～図5を用いて説明する。図2は、本発明の実施の形態1における類似データ検索装置の動作を示すフロー図である。以下の説明においては、適宜図1を参酌する。また、本実施の形態1では、類似データ検索装置2を動作させることによって、類似データ検索方法が実施される。よって、本実施の形態1における類似データ検索方法の説明は、以下の類似データ検索装置の動作説明に代える。

[0036] [ステップA1]

最初に、図2に示すように、転置インデックス作成部20が、データ記憶装置1から、検索対象の集合で構成された検索対象データ10と、集合の要素の重要度を示す要素重要度データ11とを、読み出す。そして、転置インデックス作成部20は、これらのデータを用いて検索対象の集合それぞれのサイズを計算する。

[0037] 続いて、転置インデックス作成部20は、生成予定の各転置インデックスに含まれる検索対象の集合の数が設定個数以上になるように、検索対象の集合のサイズの範囲を決定する。そして、転置インデックス作成部20は、決定したサイズの範囲毎に検索対象の集合を分けて、各転置インデックスを生成する（ステップA1）。

[0038] また、転置インデックス生成部20は、検索条件の集合が複数存在する場合には、検索条件の集合それぞれで検索を試行し、データ検索部22で要する検索時間の総和が小さくなるように、生成予定の各転置インデックスに含まれる検索対象の集合の最小数を決定することができる。

[0039] ここで、集合のサイズの計算の仕方について具体的に説明する。サイズの

計算の対象となる集合をXとすると、類似度として、上述の2種類の重複度、ダイス係数、又はジャカード係数が用いられる場合は、集合Xのサイズは、下記の数6によって定義される。また、類似度として、コサイン類似度が用いられる場合は、集合Xのサイズは、下記の数7によって定義される。

[0040] [数6]

$$\sum_{e \in X} w(e)$$

[0041] [数7]

$$\sqrt{\sum_{e \in X} w(e)^2}$$

[0042] また、ステップA1では、転置インデックス生成部20は、検索対象の集合に含まれる各要素に対して予め付与されている重要度を用いて、検索対象の集合それぞれのサイズを計算することができる。このとき、全ての要素の重要度は1としても良く、この場合、集合のサイズは集合の要素の個数に一致する。

[0043] 一方、要素の重要度が細かく設定されている程、同じサイズの集合が少なくなる。このため、上述の本実施の形態1で得られる効果は大きくなる。従って、本実施の形態1では、重要度は出来る限り、細かく設定されているのが好ましい。

[0044] また、ステップA1では、転置インデックス生成部20は、各転置インデックスが、必ず一定数以上の集合を持つように、サイズの範囲に閾値を定め、検索対象の集合の総数を、設定値で除算することによって設定個数を算出することができる。そして、転置インデックス生成部20は、算出した設定個数に基づいて、生成予定の転置インデックス毎に、検索対象の集合のサイズの範囲を決定することができる。つまり、転置インデックス生成部20は、検索対象の集合の総数を、設定個数Nで除算して均等に全体をN個に分割することで、サイズを決定しても良い。

- [0045] また、サイズの範囲の閾値及び設定個数 N は、検索条件の候補となる集合のサンプルを用いて実際に検索を行うことによって設定できる。この場合、最も計算時間が速くなるように設定するのが好ましい。
- [0046] 更に、各転置インデックスから検索可能な集合の個数の基準を決定できる場合は、検索対象の各集合のサイズを計算し、これらをサイズの昇順に並べ、サイズの小さい集合から各転置インデックスに所定の条件を満たすまで加えていくことで、転置インデックスの生成が可能である。
- [0047] ここで、図3を用いて、ステップA1の具体例について説明する。図3は、類似度としてジャカード係数が用いられる場合のステップA1の一例を説明する図である。また、図3の例では、設定個数 N は、50に設定されている。
- [0048] 図3に示すように、検索対象データとなる各集合には、識別子SIDが付与されている。また、各集合のサイズは、類似度としてジャカード係数が用いられるので、上記の数6を用いて計算されている。例えば、SID=1の集合のサイズは6.8となる。
- [0049] また、図3の例では、検索対象データが10000個存在するため、転置インデックスの数に合わせて、検索対象データを50分割する場合は、おおよそ200個の集合が属するように転置インデックスを作成すれば良いことが分かる。
- [0050] そのため、転置インデックス生成部20は、検索対象の集合それぞれをサイズの昇順に並べ、各転置インデックスには、200個を超えるまで集合が追加される。このとき、200個目の集合と同じサイズの集合が存在する場合は、転置インデックス生成部20は、この同じサイズの集合も、200個目の集合が追加された転置インデックスに追加する。そして、転置インデックス生成部20は、サイズが異なる集合がでてきて初めて、その集合を新しい転置インデックスに追加する。
- [0051] 続いて、転置インデックス生成部20は、各転置インデックスから、検索可能な集合のサイズの最小値 β を特定する。そして、転置インデックス生成部20は、特定した各 β に、その昇順に、転置インデックスのIDを付与する

。このとき、IDを*i*とし、各IDの転置インデックスに含まれる集合を D_i とすると、各転置インデックスのサイズの幅と集合のサイズとの関係は、下記の数8によって表される。

[0052] [数8]

$$\beta_i \leq |D_i| < \beta_{i+1}$$

[0053] 上記数8より、例えば、図3に示すID=3の転置インデックスが検索対象として用いられるのであれば、サイズが6.0以上、8.0未満の集合が検索可能となる。従って、ID=3の転置インデックスには、サイズが6.8であるSID=1の集合が含まれている。

[0054] [ステップA2]

次に、ステップA1の終了後、不要転置インデックス同定部21が、検索条件の集合のサイズと、類似度に設定された閾値とを用いて、類似度毎に定められた数式に従って、類似度が閾値以上となるために必要な、検索対象の集合のサイズの条件を求める。

[0055] 続いて、不要転置インデックス同定部21は、転置インデックスのうち、それに含まれる集合のサイズの最小値が条件を満たす転置インデックス以外、即ち、類似度が閾値以上になりえない転置インデックスを検索不要と同定する（ステップA2）。

[0056] ここで、図4を用いて、ステップA2について具体的に説明する。図4は、集合のサイズの範囲を求めるための数式の一例を示す図である。つまり、図4には、検索条件の集合Qと、閾値 α とが与えられた場合の、各類似度が α 以上になる、集合のサイズの範囲を求めるための数式が示されている。

[0057] 図4に示された各数式の証明については、重複度の場合を除き、非特許文献1に開示された、整数への切り上げ、切り下げを用いない場合と同様であるので、本実施の形態1での説明は省略する。重複度である $\text{Overlap}(Q, D)$ 、 $\text{Overlap}(D, Q)$ の証明は、以下の通りとなる。

[0058] まず、 $\text{Overlap}(Q, D)$ は、 α 以上になる場合、定義より、下記の数9の通り

となる。

[0059] [数9]

$$\frac{\sum_{e \in Q \cap D} w(e)}{\sum_{e \in D} w(e)} \geq \alpha$$

[0060] また、上記の数9を変形すると、下記の数10の通りとなる。下記の数10における $|D|$ の最大値は下記の数11の通りとなる。

[0061] [数10]

$$\frac{|Q|}{\alpha} \geq \frac{\sum_{e \in Q} w(e)}{\alpha} \geq \frac{\sum_{e \in Q \cap D} w(e)}{\alpha} \geq \sum_{e \in D} w(e) = |D|$$

[0062] [数11]

$$\frac{|Q|}{\alpha} \geq |D|$$

[0063] $\text{Overlap}(D, Q)$ についてもほぼ同様で、定義より、下記の数12の通りとなる。

[0064] [数12]

$$\frac{\sum_{e \in Q \cap D} w(e)}{\sum_{e \in Q} w(e)} \geq \alpha$$

[0065] 上記の数12を変形すると下記の数13の通りとなる。下記の数13における $|D|$ の最小値は下記の数14の通りとなる。

[0066] [数13]

$$|D| = \sum_{e \in D} w(e) \geq \sum_{e \in Q \cap D} w(e) \geq \alpha \sum_{e \in Q} w(e) = \alpha |Q|$$

[0067] [数14]

$$|D| \geq \alpha |Q|$$

[0068] 例えば、検索条件の集合 Q 、そのサイズ $|Q|$ 、及び閾値 α が与えられた時、本例ではジャカード係数を対象としているため、閾値 α 以上となるためには、検索対象の集合 D のサイズは、 $\alpha |Q|$ 以上、且つ $|Q|/\alpha$ 以下である必要がある。具体的には、検索条件の集合 Q が、要素 e 、 f からなり、閾値 α が 0.6 の場合、 $|Q|$ は 2.2 となるため、サイズの最小値が $1.32(=2.2 \times 0.6)$ となり、最大値が $3.667(=2.2/0.6)$ となる。

[0069] ここで、図3に示した転置インデックスの下限 β を参照すると、転置インデックスID1と転置インデックスID2とに含まれる集合のサイズは、 $\beta_1(=0.5) \leq |D| < \beta_3(=6.0)$ となり、この時点で最小値と最大値を包含している。このため、ID3以降の転置インデックスは検索不要であることが分かる。このように、不要転置インデックス同定部21は、各転置インデックスのサイズの最小値を用いることで、検索不要な転置インデックスを同定する。

[0070] [ステップA3]

最後に、データ検索部22は、同定された検索不要な転置インデックス以外の転置インデックスについて、検索条件の対象となる集合と、その要素を含む各集合との類似度を計算し、閾値以上の集合を結果として、出力装置4に出力する（ステップA3）。

[0071] 例えば、上述した検索条件の集合 Q は、要素 e 、 f を含むため、データ検索部22は、図3に示したIDが1の転置インデックスからは、SID=3などを検索する。また、データ検索部22は、IDが2の転置インデックスからは、SID=10000などを検索する。そして、データ検索部22は、このようにして取得した集合と、検索条件の集合との類似度を計算して、類似度が実際に閾値 α 以上となるデータを検索結果として出力することができる。

[0072] また、ステップA3において、データ検索部22は、非特許文献1と同様に、 π オーバーラップ問題として検索を行なうことができる。即ち、データ検索部22は、同定された転置インデックス以外の転置インデックス（以下「非同定転置インデックス」と表記する。）に含まれる集合それぞれ毎に、検索条件の集合の要素と共通する要素を特定し、特定した要素の前記重要度の

和を計算する。そして、データ検索部 22 は、計算した和の値が、類似度が閾値 α 以上となる場合と等価になる条件を満たす場合に、計算の対象となった非同定転置インデックスの集合を検索結果とする

[0073] 具体的には、検索対象の集合のサイズ $|D|$ 、検索条件の集合のサイズ $|Q|$ 、閾値 α が与えられた時に、集合 Q と集合 D とで共通している要素の重要度の和が、図 5 に示す式によって計算される τ 以上になることが、集合 D と集合 Q との各類似度が α 以上と等価であるとされる。この場合、各検索対象の集合のサイズ $|D|$ の計算は、転置インデックスの生成時に一度行なわれれば良く、更に、検索条件の集合のサイズ $|Q|$ の計算も、一度行なわれれば良いため、毎回類似度を計算する必要がなくなり、計算効率の向上が図られる。図 5 は、類似度が閾値以上となる場合と等価であるとされる条件の一例を示す図である。

[0074] また、本実施の形態では、データ検索部 22 は、非同定転置インデックスに含まれる集合それぞれについて順に、検索条件の集合の要素を一つずつ照合することができる。そして、この場合、データ検索部 22 は、照合が未だ行われていない要素の重要度の和が、 τ 以上にならない（ τ 未満となる）場合は、転置インデックスに含まれる集合のうち、その時点までに照合が行われた集合のみを対象として、照合が未だ行われていない要素を用いた照合を実行する。そして、データ検索部 22 は、その時点までに照合が行われた集合についてのみ、共通する要素の重要度の和を計算する。

[0075] つまり、本実施の形態 1 では、非特許文献 1 に開示されているように、検索条件の集合 Q の未検索の要素の和が τ 未満になった時点で、それ以降で初めて検索される SID の集合と集合 Q との共通要素の重要度の和が τ 以上になれないという性質を利用することもできる。

[0076] 具体的には、データ検索部 22 は、各転置インデックスの集合のサイズの最小値 β を $|D|$ と見なし、その転置インデックス内の集合について最低限満たすべき τ を計算する。次に、データ検索部 22 は、検索条件の集合 Q の未検索の要素の和が τ 未満になったら、その時点までに検索された SID のみを候補として、残りの未検索の要素は、転置インデックスから取得したその要素のリ

ストに対して、各SIDの2分探索で存在確認を行う。これによって、各要素を含む集合の数を n とした場合に、線形探索では計算量が $O(n)$ となることに対し、2分探索による存在確認では $O(\log n)$ となるため、効率化が可能となる。

[0077] なお、この2分探索に切り替えた後の各集合に対する τ には、各SIDの集合のサイズが用いられることに注意する。また、効率よく検索条件の集合 Q の未検索の要素の和が τ 未満になることを確定させるため、データ検索部22は、要素を重要度の降順に検索（照合）するのが好ましい。

[0078] [実施の形態1における効果]

このように、本実施の形態1では、転置インデックス生成部20が検索対象の集合の数が少なくならないように各転置インデックスを作成する。また、不要転置インデックス同定部21が、検索条件と各転置インデックスの集合のサイズの最小値とから、類似度が閾値以上の集合を検索する際に、検索不要な転置インデックスを同定する。そして、データ検索部22が、検索不要な転置インデックス以外に対して検索を行う。このため、本実施の形態1によれば、検索条件の集合と同じサイズの集合が少ない場合でも、転置インデックスを参照する回数が総合的に少なくなるので、効率よく類似度が閾値以上となる全ての集合を検索できるようになる。

[0079] [プログラム]

本実施の形態におけるプログラムは、コンピュータに、図2に示すステップA1～A3を実行させるプログラムであれば良い。このプログラムをコンピュータにインストールし、実行することによって、本実施の形態における類似データ検索装置30と類似データ検索方法とを実現することができる。この場合、コンピュータのCPU（Central Processing Unit）は、転置インデックス生成部20、不要転置インデックス同定部21、及びデータ検索部22として機能し、処理を行なう。

[0080] （実施の形態2）

次に、本発明の実施の形態2における、類似データ検索装置、類似データ検索方法、及び類似データ検索用プログラムについて、図6～図7を参照し

ながら説明する。

[0081] [装置構成]

最初に、本実施の形態 2 における類似データ検索装置の構成について図 6 を用いて説明する。図 6 は、本発明の実施の形態 2 における類似データ検索装置の構成を示すブロック図である。

[0082] 図 6 に示すように、本実施の形態 2 においては、類似データ検索装置 5 は、以下の点で、実施の形態 1 において図 1 に示した類似データ検索装置 2 と異なっている。まず、本実施の形態 2 における類似データ検索装置 5 は、転置インデックス生成部 20、不要転置インデックス同定部 21、データ検索部 22 に加えて、同義要素変換部 23 を更に備えている。同義要素変換部 23 は、検索対象の集合及び検索条件の集合に含まれる要素のうち、定められた同義要素の集合に属する要素を、同義要素の代表要素に変換する。

[0083] また、本実施の形態 2 では、類似データ検索装置 5 は、検索対象データ 10、要素重要度データ 11 に加えて、同義要素データ 12 も利用する。同義要素データ 12 は、同義と考える要素を定義するデータであり、検索対象データ 10 及び要素重要度データ 11 と同様に、データ記憶装置 1 に格納されている。

[0084] 具体的には、同義要素変換部 23 は、検索対象データ 10、要素重要度データ 11、及び同義要素データ 12 を読み込み、同義要素の集合を生成する。そして、同義要素変換部 23 は、検索対象の集合と検索条件の集合とのそれぞれにおいて、各同義要素集合内に属する要素を、各同義要素集合の代表要素に置換し、置換後の各集合を、転置インデックス作成部 20 に出力する。

[0085] [装置動作]

次に、本発明の実施の形態 2 における類似データ検索装置 5 の動作について図 7 を用いて説明する。図 7 は、本発明の実施の形態 2 における類似データ検索装置の動作を示すフロー図である。以下の説明においては、適宜図 6 を参酌する。また、本実施の形態 2 では、類似データ検索装置 5 を動作させ

ることによって、類似データ検索方法が実施される。よって、本実施の形態 2 における類似データ検索方法の説明は、以下の類似データ検索装置の動作説明に代える。

[0086] [ステップ B 1]

最初に、図 7 に示すように、同義要素変換部 23 は、同義要素データ、要素重要度データを読み出し、検索対象の集合及び検索条件の集合それぞれについて、同義要素の集合を作成する。

[0087] 続いて、同義要素変換部 23 は、各同義要素集合の代表要素を選出し、検索対象の集合及び検索条件の集合それぞれにおける、各同義要素集合に属する要素を、代表要素に置換する（ステップ B 1）。

[0088] ステップ B 1 において、同義要素集合の作成は、要素をノードとし、同義と見なされる要素のペアに無向辺をひき、この場合に、各要素から連結成分で辿れるノードを同義要素とすることによって行なわれる。

[0089] また、代表要素としては、重要度が最大の要素、重要度が最小の要素、重要度が中央値の要素、要素が全順序である場合の最初の要素等のいずれかを用いることができる。なお、代表要素の選択方法は、特に限定されない。

[0090] [ステップ B 2～B 4]

次に、代表要素に変換済の検索対象の集合と、同じく代表要素に変換済の検索条件の集合とが用いられて、ステップ B 2～B 4 が実行される。なお、ステップ B 2～B 4 は、それぞれ、図 2 に示したステップ A 1～A 3 と同様のステップであり、これらと同様に実行され、最終的に検索結果が出力される。

[0091] [実施の形態 2 における効果]

以上のように、本実施の形態 2 では、同義要素変換部 23 が、同義要素を代表要素に置換した上で、検索処理が実行される。このため、互いに異なる要素同士であっても、同義である場合は、同一視の要素とみなされて、類似データ検索が行なわれるので、検索精度が高められることになる。

[0092] (コンピュータ)

ここで、実施の形態１及び２におけるプログラムを実行することによって、類似データ検索装置を実現するコンピュータについて図８を用いて説明する。図８は、本発明の実施の形態１及び２における類似データ検索装置を実現するコンピュータの一例を示すブロック図である。

[0093] 図８に示すように、コンピュータ１１０は、ＣＰＵ１１１と、メインメモリ１１２と、記憶装置１１３と、入力インターフェイス１１４と、表示コントローラ１１５と、データリーダ／ライタ１１６と、通信インターフェイス１１７とを備える。これらの各部は、バス１２１を介して、互いにデータ通信可能に接続される。

[0094] ＣＰＵ１１１は、記憶装置１１３に格納された、本実施の形態におけるプログラム（コード）をメインメモリ１１２に展開し、これらを所定順序で実行することにより、各種の演算を実施する。メインメモリ１１２は、典型的には、ＤＲＡＭ（Dynamic Random Access Memory）等の揮発性の記憶装置である。また、本実施の形態におけるプログラムは、コンピュータ読み取り可能な記録媒体１２０に格納された状態で提供される。なお、本実施の形態におけるプログラムは、通信インターフェイス１１７を介して接続されたインターネット上で流通するものであっても良い。

[0095] また、記憶装置１１３の具体例としては、ハードディスクドライブの他、フラッシュメモリ等の半導体記憶装置が挙げられる。入力インターフェイス１１４は、ＣＰＵ１１１と、キーボード及びマウスといった入力機器１１８との間のデータ伝送を仲介する。表示コントローラ１１５は、ディスプレイ装置１１９と接続され、ディスプレイ装置１１９での表示を制御する。

[0096] データリーダ／ライタ１１６は、ＣＰＵ１１１と記録媒体１２０との間のデータ伝送を仲介し、記録媒体１２０からのプログラムの読み出し、及びコンピュータ１１０における処理結果の記録媒体１２０への書き込みを実行する。通信インターフェイス１１７は、ＣＰＵ１１１と、他のコンピュータとの間のデータ伝送を仲介する。

[0097] また、記録媒体１２０の具体例としては、ＣＦ（Compact Flash（登録商標

））及びSD（Secure Digital）等の汎用的な半導体記憶デバイス、フレキシブルディスク（Flexible Disk）等の磁気記憶媒体、又はCD-ROM（Compact Disk Read Only Memory）などの光学記憶媒体が挙げられる。

[0098] 上述した実施の形態の一部又は全部は、以下に記載する（付記1）～（付記30）によって表現することができるが、以下の記載に限定されるものではない。

[0099] （付記1）

検索対象となるデータ及び検索条件となるデータとして集合を用いて検索を行うための類似データ検索装置であって、

検索に使用する転置インデックスの生成のため、生成予定の各転置インデックスに含まれる検索対象の集合の数が設定個数以上になるように、前記生成予定の転置インデックス毎に、前記検索対象の集合のサイズの範囲を決定し、決定した前記サイズの範囲毎に、前記検索対象の集合を分けて、前記転置インデックスを生成する、転置インデックス生成部と、

検索条件の集合のサイズと、前記検索条件の集合と前記検索対象の集合との類似度に対して設定された閾値と、に基づいて、前記類似度が閾値以上となるために必要な、前記検索対象の集合のサイズの条件を求め、前記転置インデックスのうち、それに含まれる集合のサイズの最小値が前記条件を満たす転置インデックス以外を、検索不要な転置インデックスとして同定する、不要転置インデックス同定部と、

同定された前記検索不要な転置インデックス以外の転置インデックスに対して、前記検索条件の集合を適用して、検索を実行する、データ検索部と、を備えることを特徴とする類似データ検索装置。

[0100] （付記2）

前記転置インデックス生成部が、前記検索対象の集合の総数を、設定された値で除算することによって、前記設定個数を算出し、算出した前記設定個数に基づいて、前記生成予定の転置インデックス毎に、前記検索対象の集合のサイズの範囲を決定する、付記1に記載の類似データ検索装置。

[0101] (付記3)

前記転置インデックス生成部が、前記検索条件の集合が複数存在する場合に、前記データ検索部による検索にかかる時間の総和が小さくなるように、前記生成予定の各転置インデックスに含まれる検索対象の集合の最小数を決定する、付記1または2に記載の類似データ検索装置。

[0102] (付記4)

前記不要転置インデックス同定部は、
前記検索条件の集合から前記検索対象の集合に対する重複度、前記検索対象の集合から前記検索条件の集合に対する重複度、ジャカード係数、ダイス係数、コサイン類似度のうち、いずれかによって規定される数式と、前記閾値とを用いて、前記条件を計算する、
付記1から3のいずれかに記載の類似データ検索装置。

[0103] (付記5)

前記データ検索部が、同定された前記検索不要な転置インデックス以外の転置インデックスに含まれる集合の中から、前記検索条件の集合の要素を含む集合を特定し、特定した集合と前記検索条件の集合との前記類似度が前記閾値以上となる場合に、特定した集合を検索結果とする、
付記1～4のいずれかに記載の類似データ検索装置。

[0104] (付記6)

前記転置インデックス生成部が、更に、前記検索対象の集合に含まれる各要素に対して予め付与されている重要度を用いて、前記検索対象の集合それぞれのサイズを計算する、
付記1～5のいずれかに記載の類似データ検索装置。

[0105] (付記7)

前記類似度が、前記検索条件の集合から前記検索対象の集合に対する重複度、前記検索対象の集合から前記検索条件の集合に対する重複度、ジャカード係数、ダイス係数、コサイン類似度のうち、いずれかによって規定される数式を用いて計算されており、

前記データ検索部が、同定された前記検索不要な転置インデックス以外の転置インデックスに含まれる集合それぞれ毎に、前記検索条件の集合の要素と共通する要素を特定し、特定した要素の前記重要度の和を計算し、計算した和の値が、前記類似度が前記閾値以上となる場合と等価になる条件を満たす場合に、計算の対象となった前記集合を検索結果とする、
付記 6 に記載の類似データ検索装置。

[0106] (付記 8)

前記データ検索部が、
同定された前記検索不要な転置インデックス以外の転置インデックスに含まれる集合それぞれについて順に、前記検索条件の集合の要素を一つずつ照合し、
照合が未だ行われていない要素の前記重要度の和が、前記条件を満たさなくなった場合は、前記転置インデックスに含まれる集合のうち、その時点までに照合が行われた集合のみを対象として、前記照合が未だ行われていない要素を用いた照合を行い、
前記その時点までに照合が行われた集合についてのみ、前記共通する要素の前記重要度の和を計算する、
付記 7 に記載の類似データ検索装置。

[0107] (付記 9)

前記データ検索部が、重要度の降順に、前記検索条件の集合の要素を照合する、付記 8 に記載の類似データ検索装置。

[0108] (付記 10)

前記検索対象の集合及び前記検索条件の集合に含まれる要素のうち、定められた同義要素の集合に属する要素を、前記同義要素の代表要素に変換する、同義要素変換部を、更に備えている、付記 1 から 9 のいずれかに記載の類似データ検索装置。

[0109] (付記 11)

検索対象となるデータ及び検索条件となるデータとして集合を用いて検索

を行うための方法であって、

(a) 検索に使用する転置インデックスの生成のため、生成予定の各転置インデックスに含まれる検索対象の集合の数が設定個数以上になるように、前記生成予定の転置インデックス毎に、前記検索対象の集合のサイズの範囲を決定し、決定した前記サイズの範囲毎に、前記検索対象の集合を分けて、前記転置インデックスを生成する、ステップと、

(b) 検索条件の集合のサイズと、前記検索条件の集合と前記検索対象の集合との類似度に対して設定された閾値と、に基づいて、前記類似度が閾値以上となるために必要な、前記検索対象の集合のサイズの条件を求め、前記転置インデックスのうち、それに含まれる集合のサイズの最小値が前記条件を満たす転置インデックス以外を、検索不要な転置インデックスとして同定する、ステップと、

(c) 前記(b)のステップで同定された前記検索不要な転置インデックス以外の転置インデックスに対して、前記検索条件の集合を適用して、検索を実行する、ステップと、

を有することを特徴とする類似データ検索方法。

[0110] (付記 1 2)

前記(a)のステップで、前記検索対象の集合の総数を、設定された値で除算することによって、前記設定個数を算出し、算出した前記設定個数に基づいて、前記生成予定の転置インデックス毎に、前記検索対象の集合のサイズの範囲を決定する、付記 1 1 に記載の類似データ検索方法。

[0111] (付記 1 3)

前記(a)のステップで、前記検索条件の集合が複数存在する場合に、前記(c)のステップによる検索にかかる時間の総和が小さくなるように、前記生成予定の各転置インデックスに含まれる検索対象の集合の最小数を決定する、付記 1 1 または 1 2 に記載の類似データ検索方法。

[0112] (付記 1 4)

前記(b)のステップで、前記検索条件の集合から前記検索対象の集合に

対する重複度、前記検索対象の集合から前記検索条件の集合に対する重複度、ジャカード係数、ダイス係数、コサイン類似度のうち、いずれかによって規定される数式と、前記閾値とを用いて、前記条件を計算する、
付記 1 1 から 1 3 のいずれかに記載の類似データ検索方法。

[0113] (付記 1 5)

前記 (c) のステップで、前記 (b) のステップで同定された前記検索不要な転置インデックス以外の転置インデックスに含まれる集合の中から、前記検索条件の集合の要素を含む集合を特定し、特定した集合と前記検索条件の集合との前記類似度が前記閾値以上となる場合に、特定した集合を検索結果とする、

付記 1 1 ～ 1 4 のいずれかに記載の類似データ検索方法。

[0114] (付記 1 6)

前記 (a) のステップで、更に、前記検索対象の集合に含まれる各要素に対して予め付与されている重要度を用いて、前記検索対象の集合それぞれのサイズを計算する、

付記 1 1 ～ 1 5 のいずれかに記載の類似データ検索方法。

[0115] (付記 1 7)

前記類似度が、前記検索条件の集合から前記検索対象の集合に対する重複度、前記検索対象の集合から前記検索条件の集合に対する重複度、ジャカード係数、ダイス係数、コサイン類似度のうち、いずれかによって規定される数式を用いて計算されており、

前記 (c) のステップで、前記 (b) のステップで同定された前記検索不要な転置インデックス以外の転置インデックスに含まれる集合それぞれ毎に、前記検索条件の集合の要素と共通する要素を特定し、特定した要素の前記重要度の和を計算し、計算した和の値が、前記類似度が前記閾値以上となる場合と等価になる条件を満たす場合に、計算の対象となった前記集合を検索結果とする、

付記 1 6 に記載の類似データ検索方法。

[0116] (付記 18)

前記(c)のステップで、
同定された前記検索不要な転置インデックス以外の転置インデックスに含まれる集合それぞれについて順に、前記検索条件の集合の要素を一つずつ照合し、
照合が未だ行われていない要素の前記重要度の和が、前記条件を満たさなくなった場合は、前記転置インデックスに含まれる集合のうち、その時点までに照合が行われた集合のみを対象として、前記照合が未だ行われていない要素を用いた照合を行い、
前記その時点までに照合が行われた集合についてのみ、前記共通する要素の前記重要度の和を計算する、
付記 17 に記載の類似データ検索方法。

[0117] (付記 19)

前記(c)のステップで、重要度の降順に、前記検索条件の集合の要素を照合する、付記 18 に記載の類似データ検索方法。

[0118] (付記 20)

(d) 前記検索対象の集合及び前記検索条件の集合に含まれる要素のうち、定められた同義要素の集合に属する要素を、前記同義要素の代表要素に変換する、ステップを更に有する、付記 11 から 19 のいずれかに記載の類似データ検索方法。

[0119] (付記 21)

コンピュータによって、検索対象となるデータ及び検索条件となるデータとして集合を用いて検索を行う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記録媒体であって、
前記コンピュータに、
(a) 検索に使用する転置インデックスの生成のため、生成予定の各転置インデックスに含まれる検索対象の集合の数が設定個数以上になるように、前記生成予定の転置インデックス毎に、前記検索対象の集合のサイズの範囲を

決定し、決定した前記サイズの範囲毎に、前記検索対象の集合を分けて、前記転置インデックスを生成する、ステップと、

(b) 検索条件の集合のサイズと、前記検索条件の集合と前記検索対象の集合との類似度に対して設定された閾値と、に基づいて、前記類似度が閾値以上となるために必要な、前記検索対象の集合のサイズの条件を求め、前記転置インデックスのうち、それに含まれる集合のサイズの最小値が前記条件を満たす転置インデックス以外を、検索不要な転置インデックスとして同定する、ステップと、

(c) 前記(b)のステップで同定された前記検索不要な転置インデックス以外の転置インデックスに対して、前記検索条件の集合を適用して、検索を実行する、ステップと、

を実行させる命令を含む、プログラムを記録しているコンピュータ読み取り可能な記録媒体。

[0120] (付記 2 2)

前記(a)のステップで、前記検索対象の集合の総数を、設定された値で除算することによって、前記設定個数を算出し、算出した前記設定個数に基づいて、前記生成予定の転置インデックス毎に、前記検索対象の集合のサイズの範囲を決定する、付記 2 1 に記載のコンピュータ読み取り可能な記録媒体。

[0121] (付記 2 3)

前記(a)のステップで、前記検索条件の集合が複数存在する場合に、前記(c)のステップによる検索にかかる時間の総和が小さくなるように、前記生成予定の各転置インデックスに含まれる検索対象の集合の最小数を決定する、付記 2 1 または 2 2 に記載のコンピュータ読み取り可能な記録媒体。

[0122] (付記 2 4)

前記(b)のステップで、前記検索条件の集合から前記検索対象の集合に対する重複度、前記検索対象の集合から前記検索条件の集合に対する重複度、ジャカード係数、ダイス係数、コサイン類似度のうち、いずれかによって

規定される数式と、前記閾値とを用いて、前記条件を計算する、
付記 21 から 23 のいずれかに記載のコンピュータ読み取り可能な記録媒体
。

[0123] (付記 25)

前記 (c) のステップで、前記 (b) のステップで同定された前記検索不要な転置インデックス以外の転置インデックスに含まれる集合の中から、前記検索条件の集合の要素を含む集合を特定し、特定した集合と前記検索条件の集合との前記類似度が前記閾値以上となる場合に、特定した集合を検索結果とする、

付記 21 ～ 24 のいずれかに記載のコンピュータ読み取り可能な記録媒体。

[0124] (付記 26)

前記 (a) のステップで、更に、前記検索対象の集合に含まれる各要素に対して予め付与されている重要度を用いて、前記検索対象の集合それぞれのサイズを計算する、

付記 21 ～ 25 のいずれかに記載のコンピュータ読み取り可能な記録媒体。

[0125] (付記 27)

前記類似度が、前記検索条件の集合から前記検索対象の集合に対する重複度、前記検索対象の集合から前記検索条件の集合に対する重複度、ジャカード係数、ダイス係数、コサイン類似度のうち、いずれかによって規定される数式を用いて計算されており、

前記 (c) のステップで、前記 (b) のステップで同定された前記検索不要な転置インデックス以外の転置インデックスに含まれる集合それぞれ毎に、前記検索条件の集合の要素と共通する要素を特定し、特定した要素の前記重要度の和を計算し、計算した和の値が、前記類似度が前記閾値以上となる場合と等価になる条件を満たす場合に、計算の対象となった前記集合を検索結果とする、

付記 26 に記載のコンピュータ読み取り可能な記録媒体。

[0126] (付記 28)

前記（c）のステップで、
同定された前記検索不要な転置インデックス以外の転置インデックスに含まれる集合それぞれについて順に、前記検索条件の集合の要素を一つずつ照合し、
照合が未だ行われていない要素の前記重要度の和が、前記条件を満たさなくなった場合は、前記転置インデックスに含まれる集合のうち、その時点までに照合が行われた集合のみを対象として、前記照合が未だ行われていない要素を用いた照合を行い、
前記その時点までに照合が行われた集合についてのみ、前記共通する要素の前記重要度の和を計算する、
付記 27 に記載のコンピュータ読み取り可能な記録媒体。

[0127] （付記 29）

前記（c）のステップで、重要度の降順に、前記検索条件の集合の要素を照合する、付記 28 に記載のコンピュータ読み取り可能な記録媒体。

[0128] （付記 30）

前記プログラムが、更に、前記コンピュータに、
（d）前記検索対象の集合及び前記検索条件の集合に含まれる要素のうち、定められた同義要素の集合に属する要素を、前記同義要素の代表要素に変換する、ステップを実行させる命令を含む、付記 21 から 29 のいずれかに記載のコンピュータ読み取り可能な記録媒体。

[0129] 以上、実施の形態を参照して本願発明を説明したが、本願発明は上記実施の形態に限定されるものではない。本願発明の構成や詳細には、本願発明の Scope 内で当業者が理解し得る様々な変更をすることができる。

[0130] この出願は、2013 年 3 月 7 日に提出された日本出願特願 2013-045566 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。

産業上の利用可能性

[0131] 以上のように本発明によれば、検索対象となるデータ中に同じサイズの集

合が少ない場合であっても、転置インデックスにおける検索回数の増加による検索効率の低下を抑制することができる。本発明は、重複データを削除するための重複データの照合処理及び類似データをまとめるための処理を行なうデータクラスタリングシステム、辞書エントリとのソフトマッチングによって辞書ソフト照合を行なうシステム、等に有用である。

符号の説明

- [0132]
- 1 データ記憶装置
 - 2 類似データ検索装置（実施の形態１）
 - 3 入力装置
 - 4 出力装置
 - 5 類似データ検索装置（実施の形態２）
 - 10 検索対象データ
 - 11 要素重要度データ
 - 12 同義要素データ
 - 20 転置インデックス生成部
 - 21 検索不要転置インデックス同定部
 - 22 データ検索部
 - 23 同義要素変換部
 - 30 類似データ検索システム
 - 110 コンピュータ
 - 111 CPU
 - 112 メインメモリ
 - 113 記憶装置
 - 114 入力インターフェイス
 - 115 表示コントローラ
 - 116 データリーダ／ライタ
 - 117 通信インターフェイス
 - 118 入力機器

1 1 9 ディスプレイ装置

1 2 0 記録媒体

1 2 1 バス

請求の範囲

[請求項1] 検索対象となるデータ及び検索条件となるデータとして集合を用いて検索を行うための類似データ検索装置であって、

 検索に使用する転置インデックスの生成のため、生成予定の各転置インデックスに含まれる検索対象の集合の数が設定個数以上になるように、前記生成予定の転置インデックス毎に、前記検索対象の集合のサイズの範囲を決定し、決定した前記サイズの範囲毎に、前記検索対象の集合を分けて、前記転置インデックスを生成する、転置インデックス生成部と、

 検索条件の集合のサイズと、前記検索条件の集合と前記検索対象の集合との類似度に対して設定された閾値と、に基づいて、前記類似度が閾値以上となるために必要な、前記検索対象の集合のサイズの条件を求め、前記転置インデックスのうち、それに含まれる集合のサイズの最小値が前記条件を満たす転置インデックス以外を、検索不要な転置インデックスとして同定する、不要転置インデックス同定部と、

 同定された前記検索不要な転置インデックス以外の転置インデックスに対して、前記検索条件の集合を適用して、検索を実行する、データ検索部と、

 を備えることを特徴とする類似データ検索装置。

[請求項2] 前記転置インデックス生成部が、前記検索対象の集合の総数を、設定された値で除算することによって、前記設定個数を算出し、算出した前記設定個数に基づいて、前記生成予定の転置インデックス毎に、前記検索対象の集合のサイズの範囲を決定する、請求項1に記載の類似データ検索装置。

[請求項3] 前記転置インデックス生成部が、前記検索条件の集合が複数存在する場合に、前記データ検索部による検索にかかる時間の総和が小さくなるように、前記生成予定の各転置インデックスに含まれる検索対象の集合の最小数を決定する、請求項1または2に記載の類似データ検

索装置。

[請求項4] 前記不要転置インデックス同定部は、
前記検索条件の集合から前記検索対象の集合に対する重複度、前記検索対象の集合から前記検索条件の集合に対する重複度、ジャカード係数、ダイス係数、コサイン類似度のうち、いずれかによって規定される数式と、前記閾値とを用いて、前記条件を計算する、
請求項1から3のいずれかに記載の類似データ検索装置。

[請求項5] 前記データ検索部が、同定された前記検索不要な転置インデックス以外の転置インデックスに含まれる集合の中から、前記検索条件の集合の要素を含む集合を特定し、特定した集合と前記検索条件の集合との前記類似度が前記閾値以上となる場合に、特定した集合を検索結果とする、
請求項1～4のいずれかに記載の類似データ検索装置。

[請求項6] 前記転置インデックス生成部が、更に、前記検索対象の集合に含まれる各要素に対して予め付与されている重要度を用いて、前記検索対象の集合それぞれのサイズを計算する、
請求項1～5のいずれかに記載の類似データ検索装置。

[請求項7] 前記類似度が、前記検索条件の集合から前記検索対象の集合に対する重複度、前記検索対象の集合から前記検索条件の集合に対する重複度、ジャカード係数、ダイス係数、コサイン類似度のうち、いずれかによって規定される数式を用いて計算されており、
前記データ検索部が、
同定された前記検索不要な転置インデックス以外の転置インデックスに含まれる集合それぞれについて順に、前記検索条件の集合の要素を、重要度の降順に、一つずつ照合し、
照合が未だ行われていない要素の前記重要度の和が、前記条件を満たさなくなった場合は、前記転置インデックスに含まれる集合のうち、その時点までに照合が行われた集合のみを対象として、前記照合が未

だ行われていない要素を用いた照合を行い、
前記その時点までに照合が行われた集合についてののみ、前記検索条件の集合の要素と共通する要素の前記重要度の和を計算し、
計算した和の値が、前記類似度が前記閾値以上となる場合と等価になる条件を満たす場合に、計算の対象となった前記集合を検索結果とする、
請求項 6 に記載の類似データ検索装置。

[請求項8] 前記検索対象の集合及び前記検索条件の集合に含まれる要素のうち、定められた同義要素の集合に属する要素を、前記同義要素の代表要素に変換する、同義要素変換部を、更に備えている、請求項 1 から 7 のいずれかに記載の類似データ検索装置。

[請求項9] 検索対象となるデータ及び検索条件となるデータとして集合を用いて検索を行うための方法であって、

(a) 検索に使用する転置インデックスの生成のため、生成予定の各転置インデックスに含まれる検索対象の集合の数が設定個数以上になるように、前記生成予定の転置インデックス毎に、前記検索対象の集合のサイズの範囲を決定し、決定した前記サイズの範囲毎に、前記検索対象の集合を分けて、前記転置インデックスを生成する、ステップと、

(b) 検索条件の集合のサイズと、前記検索条件の集合と前記検索対象の集合との類似度に対して設定された閾値と、に基づいて、前記類似度が閾値以上となるために必要な、前記検索対象の集合のサイズの条件を求め、前記転置インデックスのうち、それに含まれる集合のサイズの最小値が前記条件を満たす転置インデックス以外を、検索不要な転置インデックスとして同定する、ステップと、

(c) 前記 (b) のステップで同定された前記検索不要な転置インデックス以外の転置インデックスに対して、前記検索条件の集合を適用して、検索を実行する、ステップと、

を有することを特徴とする類似データ検索方法。

[請求項10]

コンピュータによって、検索対象となるデータ及び検索条件となるデータとして集合を用いて検索を行う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記録媒体であって、
前記コンピュータに、

(a) 検索に使用する転置インデックスの生成のため、生成予定の各転置インデックスに含まれる検索対象の集合の数が設定個数以上になるように、前記生成予定の転置インデックス毎に、前記検索対象の集合のサイズの範囲を決定し、決定した前記サイズの範囲毎に、前記検索対象の集合を分けて、前記転置インデックスを生成する、ステップと、

(b) 検索条件の集合のサイズと、前記検索条件の集合と前記検索対象の集合との類似度に対して設定された閾値と、に基づいて、前記類似度が閾値以上となるために必要な、前記検索対象の集合のサイズの条件を求め、前記転置インデックスのうち、それに含まれる集合のサイズの最小値が前記条件を満たす転置インデックス以外を、検索不要な転置インデックスとして同定する、ステップと、

(c) 前記(b)のステップで同定された前記検索不要な転置インデックス以外の転置インデックスに対して、前記検索条件の集合を適用して、検索を実行する、ステップと、

を実行させる命令を含む、プログラムを記録しているコンピュータ読み取り可能な記録媒体。