(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7464193号
(P7464193)

(45)発行日　令和6年4月9日(2024.4.9)

(24)登録日　令和6年4月1日(2024.4.1)

(51)国際特許分類
Ｇ０６Ｆ　16/903(2019.01)

ＦＩ
Ｇ０６Ｆ　16/903

請求項の数　10　（全27頁）

(21)出願番号　特願2023-520672(P2023-520672)
(86)(22)出願日　令和3年5月13日(2021.5.13)
(86)国際出願番号　PCT/JP2021/018169
(87)国際公開番号　WO2022/239174
(87)国際公開日　令和4年11月17日(2022.11.17)
審査請求日　令和5年9月13日(2023.9.13)

(73)特許権者　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(74)代理人　100103090
弁理士　岩壁 冬樹
(74)代理人　100124501
弁理士　塩川 誠人
(72)発明者　大野 善之
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　齊藤 貴孝

最終頁に続く

(54)【発明の名称】　類似度導出システムおよび類似度導出方法

(57)【特許請求の範囲】

【請求項１】

複数の集合に含まれる各集合の個々の要素に対して複数のハッシュ関数を適用して得られる複数のハッシュ値を求める際に、前記複数のハッシュ関数のうちの所定のハッシュ関数によって得られるハッシュ値が一致し、かつ、要素自体が一致する複数の要素に関しては、前記所定のハッシュ関数以外の各ハッシュ関数の計算の重複を排除し、前記各集合の個々の要素に対して前記複数のハッシュ値を求めるハッシュ値計算手段と、

前記複数の集合に含まれるそれぞれの集合に関して、個々のハッシュ関数毎に、ハッシュ値の最小値である最小ハッシュ値を特定する最小ハッシュ値特定手段と、

前記複数の集合から得られる１組以上の集合のペアに関して、ペアをなす２つの集合の類似度を、個々のハッシュ関数に対応する最小ハッシュ値に基づいて導出する類似度導出手段とを備える

ことを特徴とする類似度導出システム。

【請求項２】

前記ハッシュ値計算手段は、

前記各集合の個々の要素に対して前記所定のハッシュ関数を適用することによって第１のハッシュ値を計算する第１のハッシュ値計算手段と、

前記各集合の個々の要素のうち、前記第１のハッシュ値が一致し、かつ、要素自体が一致する複数の要素からは１つの要素だけを取り出すとともに、前記複数の要素に該当しない各要素を取り出すことによって、重複なく全ての種類の要素を含む１つの集合である全

体集合を生成する全体集合生成手段と、
前記複数の集合に含まれるどの集合のどの要素が前記全体集合のどの要素に該当するかを示すインデックス情報を生成するインデックス情報生成手段と、
前記全体集合に属する各要素に対して、前記複数のハッシュ関数のうちの前記所定のハッシュ関数以外の各ハッシュ関数を適用することによって、前記各ハッシュ関数に対応するハッシュ値を計算する第２のハッシュ値計算手段と、
前記インデックス情報に基づいて、前記各集合の個々の要素の前記第１のハッシュ値と、当該第１のハッシュ値に対応する、前記第２のハッシュ値計算手段によって計算されたハッシュ値とを組み合わせることによって、前記各集合の個々の要素に対して、前記複数のハッシュ関数に対応する複数のハッシュ値を決定するハッシュ値決定手段とを含む
請求項１に記載の類似度導出システム。
【請求項３】
前記ハッシュ値計算手段は、
前記複数の集合から、順次、１つの集合を選択する集合選択手段と、
選択された集合から、順次、１つの要素を選択する要素選択手段と、
選択された要素に前記所定のハッシュ関数を適用することによって前記選択された要素の第１のハッシュ値を計算する第１のハッシュ値計算手段と、
前記選択された要素と、第１のハッシュ値が一致し、かつ、要素自体が一致する要素である一致要素が既に選択されているか否かを判定する判定手段と、
前記一致要素が既に選択されている場合に、前記選択された要素の前記所定のハッシュ関数以外の各ハッシュ関数に対応するハッシュ値を、前記一致要素の前記各ハッシュ関数に対応するハッシュ値と同一であると定め、
前記一致要素が選択されていない場合に、前記選択された要素の前記所定のハッシュ関数以外の各ハッシュ関数に対応するハッシュ値を計算する第２のハッシュ値計算手段とを含む
請求項１に記載の類似度導出システム。
【請求項４】
前記所定のハッシュ関数は、前記複数のハッシュ関数の中で最も値域が広いハッシュ関数である
請求項１から請求項３のうちのいずれか１項に記載の類似度導出システム。
【請求項５】
前記類似度導出手段は、
個々のハッシュ関数に対応する最小ハッシュ値同士が一致している数を、ペアをなす２つの集合の類似度として定める
請求項１から請求項４のうちのいずれか１項に記載の類似度導出システム。
【請求項６】
コンピュータが、
複数の集合に含まれる各集合の個々の要素に対して複数のハッシュ関数を適用して得られる複数のハッシュ値を求める際に、前記複数のハッシュ関数のうちの所定のハッシュ関数によって得られるハッシュ値が一致し、かつ、要素自体が一致する複数の要素に関しては、前記所定のハッシュ関数以外の各ハッシュ関数の計算の重複を排除し、前記各集合の個々の要素に対して前記複数のハッシュ値を求めるハッシュ値計算処理、
前記複数の集合に含まれるそれぞれの集合に関して、個々のハッシュ関数毎に、ハッシュ値の最小値である最小ハッシュ値を特定する最小ハッシュ値特定処理、および、
前記複数の集合から得られる１組以上の集合のペアに関して、ペアをなす２つの集合の類似度を、個々のハッシュ関数に対応する最小ハッシュ値に基づいて導出する類似度導出処理を実行する
ことを特徴とする類似度導出方法。
【請求項７】
前記コンピュータが、

前記ハッシュ値計算処理で、
前記各集合の個々の要素に対して前記所定のハッシュ関数を適用することによって第１のハッシュ値を計算する第１のハッシュ値計算処理、
前記各集合の個々の要素のうち、前記第１のハッシュ値が一致し、かつ、要素自体が一致する複数の要素からは１つの要素だけを取り出すとともに、前記複数の要素に該当しない各要素を取り出すことによって、重複なく全ての種類の要素を含む１つの集合である全体集合を生成する全体集合生成処理、
前記複数の集合に含まれるどの集合のどの要素が前記全体集合のどの要素に該当するかを示すインデックス情報を生成するインデックス情報生成処理、
前記全体集合に属する各要素に対して、前記複数のハッシュ関数のうちの前記所定のハッシュ関数以外の各ハッシュ関数を適用することによって、前記各ハッシュ関数に対応するハッシュ値を計算する第２のハッシュ値計算処理、および、
前記インデックス情報に基づいて、前記各集合の個々の要素の前記第１のハッシュ値と、当該第１のハッシュ値に対応する、前記第２のハッシュ値計算処理で計算されたハッシュ値とを組み合わせることによって、前記各集合の個々の要素に対して、前記複数のハッシュ関数に対応する複数のハッシュ値を決定するハッシュ値決定処理を実行する
請求項６に記載の類似度導出方法。
【請求項８】
前記コンピュータが、
前記ハッシュ値計算処理で、
前記複数の集合から、順次、１つの集合を選択する集合選択処理、
選択された集合から、順次、１つの要素を選択する要素選択処理、
選択された要素に前記所定のハッシュ関数を適用することによって前記選択された要素の第１のハッシュ値を計算する第１のハッシュ値計算処理、
前記選択された要素と、第１のハッシュ値が一致し、かつ、要素自体が一致する要素である一致要素が既に選択されているか否かを判定する判定処理、および、
前記一致要素が既に選択されている場合に、前記選択された要素の前記所定のハッシュ関数以外の各ハッシュ関数に対応するハッシュ値を、前記一致要素の前記各ハッシュ関数に対応するハッシュ値と同一であると定め、
前記一致要素が選択されていない場合に、前記選択された要素の前記所定のハッシュ関数以外の各ハッシュ関数に対応するハッシュ値を計算する第２のハッシュ値計算処理を実行する
請求項６に記載の類似度導出方法。
【請求項９】
コンピュータに、
複数の集合に含まれる各集合の個々の要素に対して複数のハッシュ関数を適用して得られる複数のハッシュ値を求める際に、前記複数のハッシュ関数のうちの所定のハッシュ関数によって得られるハッシュ値が一致し、かつ、要素自体が一致する複数の要素に関しては、前記所定のハッシュ関数以外の各ハッシュ関数の計算の重複を排除し、前記各集合の個々の要素に対して前記複数のハッシュ値を求めるハッシュ値計算処理、
前記複数の集合に含まれるそれぞれの集合に関して、個々のハッシュ関数毎に、ハッシュ値の最小値である最小ハッシュ値を特定する最小ハッシュ値特定処理、および、
前記複数の集合から得られる１組以上の集合のペアに関して、ペアをなす２つの集合の類似度を、個々のハッシュ関数に対応する最小ハッシュ値に基づいて導出する類似度導出処理
を実行させるための類似度導出プログラム。
【請求項１０】
前記コンピュータに、
前記ハッシュ値計算処理で、
前記各集合の個々の要素に対して前記所定のハッシュ関数を適用することによって第１

のハッシュ値を計算する第１のハッシュ値計算処理、
　前記各集合の個々の要素のうち、前記第１のハッシュ値が一致し、かつ、要素自体が一致する複数の要素からは１つの要素だけを取り出すとともに、前記複数の要素に該当しない各要素を取り出すことによって、重複なく全ての種類の要素を含む１つの集合である全体集合を生成する全体集合生成処理、
　前記複数の集合に含まれるどの集合のどの要素が前記全体集合のどの要素に該当するかを示すインデックス情報を生成するインデックス情報生成処理、
　前記全体集合に属する各要素に対して、前記複数のハッシュ関数のうちの前記所定のハッシュ関数以外の各ハッシュ関数を適用することによって、前記各ハッシュ関数に対応するハッシュ値を計算する第２のハッシュ値計算処理、および、
　前記インデックス情報に基づいて、前記各集合の個々の要素の前記第１のハッシュ値と、当該第１のハッシュ値に対応する、前記第２のハッシュ値計算処理で計算されたハッシュ値とを組み合わせることによって、前記各集合の個々の要素に対して、前記複数のハッシュ関数に対応する複数のハッシュ値を決定するハッシュ値決定処理を実行させる
　請求項９に記載の類似度導出プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
　本発明は、集合同士の類似度を導出する類似度導出システム、類似度導出方法、および、類似度導出プログラムに関する。
【背景技術】
【０００２】
　集合の各要素に同一のハッシュ関数を適用することによって、要素毎にハッシュ値が得られる。そのハッシュ値の最小値をＭｉｎＨａｓｈと称する場合がある。本明細書では、同一のハッシュ関数に基づいて要素毎に得られるハッシュ値の最小値を最小ハッシュ値と記す。前述の集合の各要素に複数のハッシュ関数を適用すれば、ハッシュ関数と同数の最小ハッシュ値が得られる。
【０００３】
　また、２つの集合に、共通の複数のハッシュ関数を適用し、集合毎に複数の最小ハッシュ値を求め、その複数の最小ハッシュ値に基づいて、２つの集合の類似度を求めることが考えられる。
【０００４】
　また、特許文献１には、最小ハッシュ値を用いて類似テキストを探索する方法が記載されている。
【先行技術文献】
【特許文献】
【０００５】
【文献】特開２０２０－４１０７号公報
【発明の概要】
【発明が解決しようとする課題】
【０００６】
　前述の例のように集合の類似度を求めるためには、個々の集合の個々の要素毎に、複数のハッシュ関数を適用することによって複数のハッシュ値を求めなければならない。これは、集合毎に、複数の最小ハッシュ値を求める必要があるためである。ここで、集合の個数をｎ個とする。また、ここでは、説明を簡単にするために、各集合の要素数がｋ個で共通であるとする。また、ハッシュ関数の数をｍ個とする。この場合、ハッシュ値の計算を、ｎ・ｋ・ｍ回行う必要がある。
【０００７】
　しかし、ハッシュ値の計算量は少ない方が好ましい。
【０００８】

そこで、本発明は、集合の類似度を導出する際に、ハッシュ値の計算量を低減させることができる類似度導出システム、類似度導出方法、および、類似度導出プログラム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】

【０００９】

本発明による類似度導出システムは、複数の集合に含まれる各集合の個々の要素に対して複数のハッシュ関数を適用して得られる複数のハッシュ値を求める際に、その複数のハッシュ関数のうちの所定のハッシュ関数によって得られるハッシュ値が一致し、かつ、要素自体が一致する複数の要素に関しては、所定のハッシュ関数以外の各ハッシュ関数の計算の重複を排除し、各集合の個々の要素に対して複数のハッシュ値を求めるハッシュ値計算手段と、複数の集合に含まれるそれぞれの集合に関して、個々のハッシュ関数毎に、ハッシュ値の最小値である最小ハッシュ値を特定する最小ハッシュ値特定手段と、複数の集合から得られる１組以上の集合のペアに関して、ペアをなす２つの集合の類似度を、個々のハッシュ関数に対応する最小ハッシュ値に基づいて導出する類似度導出手段とを備えることを特徴とする。

【００１０】

本発明による類似度導出方法は、コンピュータが、複数の集合に含まれる各集合の個々の要素に対して複数のハッシュ関数を適用して得られる複数のハッシュ値を求める際に、その複数のハッシュ関数のうちの所定のハッシュ関数によって得られるハッシュ値が一致し、かつ、要素自体が一致する複数の要素に関しては、所定のハッシュ関数以外の各ハッシュ関数の計算の重複を排除し、各集合の個々の要素に対して複数のハッシュ値を求めるハッシュ値計算処理、複数の集合に含まれるそれぞれの集合に関して、個々のハッシュ関数毎に、ハッシュ値の最小値である最小ハッシュ値を特定する最小ハッシュ値特定処理、および、複数の集合から得られる１組以上の集合のペアに関して、ペアをなす２つの集合の類似度を、個々のハッシュ関数に対応する最小ハッシュ値に基づいて導出する類似度導出処理を実行することを特徴とする。

【００１１】

本発明による類似度導出プログラムは、コンピュータに、複数の集合に含まれる各集合の個々の要素に対して複数のハッシュ関数を適用して得られる複数のハッシュ値を求める際に、その複数のハッシュ関数のうちの所定のハッシュ関数によって得られるハッシュ値が一致し、かつ、要素自体が一致する複数の要素に関しては、所定のハッシュ関数以外の各ハッシュ関数の計算の重複を排除し、各集合の個々の要素に対して複数のハッシュ値を求めるハッシュ値計算処理、複数の集合に含まれるそれぞれの集合に関して、個々のハッシュ関数毎に、ハッシュ値の最小値である最小ハッシュ値を特定する最小ハッシュ値特定処理、および、複数の集合から得られる１組以上の集合のペアに関して、ペアをなす２つの集合の類似度を、個々のハッシュ関数に対応する最小ハッシュ値に基づいて導出する類似度導出処理を実行させる。

【発明の効果】

【００１２】

本発明によれば、集合の類似度を導出する際に、ハッシュ値の計算量を低減させることができる。

【図面の簡単な説明】

【００１３】

【図１】本発明の第１の実施形態の類似度導出システムの例を示すブロック図である。

【図２】第１の実施形態の処理経過の例を示すフローチャートである。

【図３】本発明の第２の実施形態の類似度導出システムの例を示すブロック図である。

【図４】各集合Ａ，Ｂの個々の要素毎に得られた第１のハッシュ値の例を示す模式図である。

【図５】図４に示す集合Ａ，Ｂから得られる全体集合を示す模式図である。

【図６】インデックス集合の例を示す模式図である。

【図７】図５に示す全体集合に属する各要素に対して、第２のハッシュ値計算部５４が計算した各ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値の例を示す模式図である。
【図８】各集合の個々の要素に対して決定された、複数のハッシュ関数ｈ１，ｈ２，・・・，ｈｍに対応する複数のハッシュ値の例を示す模式図である。
【図９】集合Ａ、集合Ｂに関してそれぞれ、ハッシュ関数毎に特定された最小ハッシュ値を示す模式図である。
【図１０】第２の実施形態の処理経過の例を示すフローチャートである。
【図１１】ステップＳ３の動作の具体例の一例を示すフローチャートである。
【図１２】本発明の第３の実施形態の類似度導出システムの例を示すブロック図である。
【図１３】第３の実施形態の処理経過の例を示すフローチャートである。
【図１４】第３の実施形態の処理経過の例を示すフローチャートである。
【図１５】複数の集合のうちの１つの集合の例を示す模式図である。
【図１６】最初にステップＳ７３を実行した後に得られたハッシュ値の例を示す模式図である。
【図１７】最初にステップＳ７６を実行した後に得られたハッシュ値の例を示す模式図である。
【図１８】ステップＳ７７で未選択の要素がないと判定されるまでに得られたハッシュ値の例を示す模式図である。
【図１９】複数の集合のうちの１つの集合の例を示す模式図である。
【図２０】集合Ｂの要素”mountain”が選択され、ステップＳ７３を実行した後に得られたハッシュ値の例を示す模式図である。
【図２１】集合Ｂの要素”mountain”を選択したときにおけるステップＳ７５の実行後に得られているハッシュ値の例を示す模式図である。
【図２２】各集合の各要素に対して求められた複数のハッシュ値の例を示す模式図である。
【図２３】本発明の実施形態の類似度導出システム１に係るコンピュータの構成例を示す概略ブロック図である。
【発明を実施するための形態】
【００１４】
以下、本発明の実施形態を図面を参照して説明する。以下に示す各実施形態では、複複数の集合が、予め、各実施形態の類似度導出システムに入力されているものとする。
【００１５】
実施形態１．
第１の実施形態は、本発明の概要を示す実施形態である。より具体的な事項については、後述の第２の実施形態、第３の実施形態で説明する。
【００１６】
図１は、本発明の第１の実施形態の類似度導出システムの例を示すブロック図である。類似度導出システム１は、ハッシュ値計算部２と、最小ハッシュ値特定部３と、類似度導出部４とを備える。
【００１７】
ハッシュ値計算部２は、複数の集合に含まれる各集合の個々の要素に対して、複数のハッシュ関数を適用して得られる複数のハッシュ値を求める。ただし、このとき、ハッシュ値計算部２は、その複数のハッシュ関数のうちの所定のハッシュ関数によって得られるハッシュ値が一致し、かつ、要素自体が一致する複数の要素に関しては、所定のハッシュ関数以外の各ハッシュ関数の計算の重複を排除し、各集合の個々の要素に対して複数のハッシュ値を求める。
【００１８】
複数のハッシュ関数の中で値域が最も広いハッシュ関数を所定のハッシュ関数として定めることが好ましい。すなわち、所定のハッシュ関数は、複数のハッシュ関数の中で最も値域が広いハッシュ関数であることが好ましい。
【００１９】

最小ハッシュ値特定部３は、複数の集合に含まれるそれぞれの集合に関して、個々のハッシュ関数毎に、ハッシュ値の最小値である最小ハッシュ値を特定する。例えば、１つの集合Ａに着目したとする。また、あるハッシュ関数に着目した場合、そのハッシュ関数に対応する、集合Ａの要素数分のハッシュ値が得られている。最小ハッシュ値特定部３は、そのハッシュ関数に対応する最小ハッシュ値として、それらのハッシュ値の最小値を特定する。最小ハッシュ値特定部３は、他の各ハッシュ関数に関してもそれぞれ、同様に、最小ハッシュ値を特定する。最小ハッシュ値特定部３は、この処理を、複数の集合に含まれるそれぞれの集合に対して行う。この結果、集合毎に、個々のハッシュ関数に対応する最小ハッシュ値が定まる。

【００２０】

類似度導出部４は、複数の集合から得られる１組以上の集合のペアに関して、ペアをなす２つの集合の類似度を、個々のハッシュ関数に対応する最小ハッシュ値に基づいて導出する。例えば、類似度導出部４は、複数の集合から得られる全てのペア（集合のペア）に関して、ペアをなす２つの集合の類似度を算出してもよい。

【００２１】

また、類似度導出部４は、例えば、個々のハッシュ値に対応する最小ハッシュ値同士が一致している数を、ペアをなす２つの集合の類似度として定めてもよい。

【００２２】

ハッシュ値計算部２、最小ハッシュ値特定部３、および、類似度導出部４は、例えば、類似度導出プログラムに従って動作するコンピュータのＣＰＵ（Central Processing Unit）によって実現される。例えば、ＣＰＵが、コンピュータのプログラム記憶装置等のプログラム記録媒体から類似度導出プログラムを読み込み、その類似度導出プログラムに従って、ハッシュ値計算部２、最小ハッシュ値特定部３、および、類似度導出部４として動作すればよい。

【００２３】

図２は、第１の実施形態の処理経過の例を示すフローチャートである。なお、既に説明した事項については、適宜、説明を省略する。

【００２４】

まず、ハッシュ値計算部２が、所定のハッシュ関数によって得られるハッシュ値が一致し、かつ、要素自体が一致する複数の要素に関しては、所定のハッシュ関数以外の各ハッシュ関数の計算の重複を排除し、各集合の個々の要素に対して複数のハッシュ値を求める（ステップＳ１）。

【００２５】

次に、最小ハッシュ値特定部３が、複数の集合に含まれるそれぞれの集合に関して、個々のハッシュ関数毎に最小ハッシュ値を特定する（ステップＳ２）。

【００２６】

次に、類似度導出部４が、複数の集合から得られる１組以上の集合のペアに関して、ペアをなす２つの集合の類似度を導出する（ステップＳ３）。

【００２７】

本実施形態によれば、ハッシュ値計算部２は、複数の集合に含まれる各集合の個々の要素に対して複数のハッシュ値を求める際に、所定のハッシュ関数によって得られるハッシュ値が一致し、かつ、要素自体が一致する複数の要素に関しては、所定のハッシュ関数以外の各ハッシュ関数の計算の重複を排除する。従って、その複数の要素に対して、同一のハッシュ関数の計算が繰り返し行われることはない。よって、集合の類似度を導出する際に、ハッシュ値の計算量を低減させることができる。

【００２８】

実施形態２．

第２の実施形態は、第１の実施形態をより具体的に示した実施形態である。図３は、本発明の第２の実施形態の類似度導出システムの例を示すブロック図である。本実施形態の類似度導出システム１も、ハッシュ値計算部２と、最小ハッシュ値特定部３と、類似度導

出部４とを備える。最小ハッシュ値特定部３および類似度導出部４は、第１の実施形態における最小ハッシュ値特定部３および類似度導出部４と同様である。

【００２９】

また、ハッシュ値計算部２は、第１のハッシュ値計算部５１と、全体集合生成部５２と、インデックス情報生成部５３と、第２のハッシュ値計算部５４と、ハッシュ値決定部５５とを備える。

【００３０】

第１のハッシュ値計算部５１は、複数の集合に含まれる各集合の個々の要素に対して、複数のハッシュ関数のうちの所定のハッシュ関数を適用することによって、ハッシュ値を計算する。この所定のハッシュ関数によって得られるハッシュ値を第１のハッシュ値と記す。したがって、各集合の個々の要素毎に第１のハッシュ値が計算される。

【００３１】

以下では、説明を簡単にするために２つの集合Ａ，Ｂに着目して説明するが、集合の数は３つ以上であってもよい。また、複数のハッシュ関数の数は、ｍ個であり、個々のハッシュ関数をｈ１，ｈ２，・・・，ｈｍと記すこととする。そして、ｈ１が、上記の所定のハッシュ関数であるものとする。

【００３２】

なお、複数のハッシュ関数の中から１つハッシュ関数を、予め所定の関数として定めておけばよい。このとき、複数のハッシュ関数の中で値域が最も広いハッシュ関数を所定のハッシュ関数として定めることが好ましい。すなわち、所定のハッシュ関数は、複数のハッシュ関数の中で最も値域が広いハッシュ関数であることが好ましい。本例では、ハッシュ関数ｈ１の値域が、複数のハッシュ関数ｈ１～ｈｍの値域の中で最も広いものとする。この点は、前述の第１の実施形態や後述の第３の実施形態においても同様である。

【００３３】

図４は、各集合Ａ，Ｂの個々の要素毎に得られた第１のハッシュ値の例を示す模式図である。図４に示すように、集合Ａは４つ文字列を要素として持ち、集合Ｂは、３つの文字列を要素として持っているものとする。第１のハッシュ値計算部５１は、これらの個々の要素に対して、所定のハッシュ関数ｈ１によって、第１のハッシュ値を計算する。図４では、個々の要素の右側に示した値が、第１のハッシュ値である。

【００３４】

全体集合生成部５２は、複数の集合に含まれる各集合の個々の要素のうち、第１のハッシュ値が一致し、かつ、要素自体が一致する複数の要素からは１つの要素だけを取り出すとともに、その複数の要素に該当しない各要素を取り出すことによって、重複なく全ての種類の要素を含む１つの集合を生成する。以下、この集合を全体集合と記す。

【００３５】

図５は、図４に示す集合Ａ，Ｂから得られる全体集合を示す模式図である。図４に示すように、各集合の個々の要素のうち、第１のハッシュ値が一致し、かつ、要素自体が一致する複数の要素は、集合Ａに属する”mountain”と、集合Ｂに属する”mountain”である。この２つの要素は、第１のハッシュ値が“６６６”で一致し、かつ、要素自体が”mountain”で一致している。全体集合生成部５２は、この２つの要素から１つの要素だけを取り出す。このとき、集合Ａに属する”mountain”を取りだしてもよく、あるいは、集合Ｂに属する”mountain”を取りだしてもよい。そして、全体集合生成部５２は、第１のハッシュ値が一致し、かつ、要素自体が一致する複数の要素（集合Ａに属する”mountain”、および、集合Ｂに属する”mountain”）に該当しない各要素（本例では、”the”，”highest”，”Fuji”，”room”，”view”）を取り出す。そして、全体集合生成部５２は、取り出した各要素を含む１つの集合を、全体集合として生成する。本例では、図５に示すように、全体集合は、要素として、６個の文字列（”the”，”highest”，”mountain”，”Fuji”，”room”，”view”）を含む。全体集合生成部５２は、前述のように集合Ａ，Ｂから要素を取り出しているので、全体集合に属する要素に重複はなく、また、全体集合は、集合Ａ，Ｂに属する全ての種類の要素を含んでいる。

【００３６】
インデックス情報生成部５３は、複数の集合に含まれるどの集合のどの要素が全体集合のどの要素に該当するかを示すインデックス情報を生成する。
【００３７】
図６は、インデックス集合の例を示す模式図である。図６に例示するインデックス情報を生成する例を説明する。インデックス情報生成部５３は、全体集合に属する各要素に対して識別情報を割り当てる。図６に示す例では、全体集合に属する各要素に対して“１”から“６”の識別情報が割り当てられている。そして、インデックス情報生成部５３は、複数の集合に含まれる各集合の各要素毎に、その要素に該当する全体集合内の要素の識別情報を定めた情報を、インデックス情報として生成する。図６に示す例では、例えば、集合Ａに属する要素”the”には、“１”というインデックス情報が定められている。このことは、集合Ａに属する要素”the”が、全体集合内の要素のうちの、識別要素“１”が割り当てられた要素に該当することを示している。また、図６に示す例では、集合Ａに属する要素”mountain”、および、集合Ｂに属する要素”mountain”は、いずれも、全体集合内の要素のうちの、識別要素“３”が割り当てられた要素に該当することを示している。すなわち、集合Ａに属する要素”mountain”、および、集合Ｂに属する要素”mountain”は、全体集合内の同一の要素に該当することを示している。
【００３８】
第２のハッシュ値計算部５４は、全体集合に属する各要素に対して、複数のハッシュ関数（本例では、ハッシュ関数ｈ１，ｈ２，・・・，ｈｍ）のうちの所定のハッシュ関数（本例では、ハッシュ関数ｈ１）以外の各ハッシュ関数（すなわち、各ハッシュ関数ｈ２，・・・，ｈｍ）を適用する。そして、第２のハッシュ値計算部５４は、全体集合に属する各要素に対して、その各ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値をそれぞれ計算する。
【００３９】
図７は、図５に示す全体集合に属する各要素に対して、第２のハッシュ値計算部５４が計算した各ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値の例を示す模式図である。
【００４０】
ハッシュ値決定部５５は、インデックス情報に基づいて、複数の集合に含まれる各集合（本例では、集合Ａ，Ｂ）の個々の要素の第１のハッシュ値（図４参照）と、その第１のハッシュ値に対応する、第２のハッシュ値計算部５４によって計算されたハッシュ値（図７参照）とを組み合わせることことによって、各集合の個々の要素に対して、複数のハッシュ関数ｈ１，ｈ２，・・・，ｈｍに対応する複数のハッシュ値を決定する。
【００４１】
本例では、各集合の各要素毎に、その要素に該当する全体集合内の要素の識別情報を定めた情報を、インデックス情報としている（図６参照）。また、第１のハッシュ値計算部５１によって、各集合の各要素に対応付けて第１のハッシュ値が計算されている（図４参照）。さらに、第２のハッシュ値計算部５４によって、全体集合に属する各要素に対して、所定のハッシュ関数ｈ１以外の各ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値が計算されている。従って、ハッシュ値決定部５５は、インデックス情報に基づいて、各集合の個々の要素の第１のハッシュ値と、第２のハッシュ値計算部５４によって計算された各ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値を組み合わせることができる。そして、その結果、ハッシュ値決定部５５は、各集合の個々の要素に対して、複数のハッシュ関数ｈ１，ｈ２，・・・，ｈｍに対応する複数のハッシュ値を決定することができる。
【００４２】
図８は、各集合（本例では、集合Ａ，Ｂ）の個々の要素に対して決定された、複数のハッシュ関数ｈ１，ｈ２，・・・，ｈｍに対応する複数のハッシュ値の例を示す模式図である。

【００４３】
このように、第１のハッシュ値計算部５１、全体集合生成部５２、インデックス情報生成部５３、第２のハッシュ値計算部５４、および、ハッシュ値決定部５５の動作によって、各集合の個々の要素に対して、複数のハッシュ関数を適用して得られる複数のハッシュ値を求める。また、このとき、第１のハッシュ値が一致し、かつ、要素自体が一致する複数の要素（本例では、集合Ａに属する”mountain”、および、集合Ｂに属する”mountain”）については、所定のハッシュ関数ｈ１以外のハッシュ関数ｈ２，・・・，ｈｍの計算を重複して行っていない。具体的には、”mountain”に対する、ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値の計算は、第２のハッシュ値計算部５４がそれぞれ１回行っているのみである。

【００４４】
最小ハッシュ値特定部３は、複数の集合に含まれるそれぞれの集合（本例では、集合Ａ、集合Ｂ）に関して、個々のハッシュ関数毎に最小ハッシュ値を特定する。図９は、集合Ａ、集合Ｂに関してそれぞれ、ハッシュ関数毎に特定された最小ハッシュ値を示す模式図である。例えば、図９に示す例では、集合Ａにおけるハッシュ関数ｈ１の最小ハッシュ値は１２であり、ハッシュ関数ｈ２の最小ハッシュ値は５６である。このように、最小ハッシュ値特定部３は、集合Ａに関して、ハッシュ関数毎に最小ハッシュ値を特定する。さらに、最小ハッシュ値特定部３は、他の各集合に関しても、同様に、ハッシュ関数毎に最小ハッシュ値を特定する。

【００４５】
図４から図９では、複数の集合が２つの集合Ａ，Ｂである場合を例にして説明した。類似度導出システム１に与えられる集合が３つ以上の場合であっても、第１のハッシュ値計算部５１、全体集合生成部５２、インデックス情報生成部５３、第２のハッシュ値計算部５４、ハッシュ値決定部５５、および、最小ハッシュ値特定部３の動作は、上記の動作と同様である。

【００４６】
第１のハッシュ値計算部５１、全体集合生成部５２、インデックス情報生成部５３、第２のハッシュ値計算部５４、ハッシュ値決定部５５を含むハッシュ値計算部２は、例えば、類似度導出プログラムに従って動作するコンピュータのＣＰＵによって実現される。例えば、ＣＰＵが、コンピュータのプログラム記憶装置等のプログラム記録媒体から類似度導出プログラムを読み込み、その類似度導出プログラムに従って、第１のハッシュ値計算部５１、全体集合生成部５２、インデックス情報生成部５３、第２のハッシュ値計算部５４、ハッシュ値決定部５５を含むハッシュ値計算部２として動作すればよい。

【００４７】
次に、第２の実施形態の処理経過の例を示す。なお、既に説明した事項については、適宜、説明を省略する。図１０は、第２の実施形態の処理経過の例を示すフローチャートである。また、上記の例と同様に、複数のハッシュ関数をｈ１，ｈ２，・・・，ｈｍとし、ｈ１が所定のハッシュ関数であるものとする。

【００４８】
まず、第１のハッシュ値計算部５１が、複数の集合に含まれる各集合の個々の要素に対して所定のハッシュ関数ｈ１を適用することによって、第１のハッシュ値を計算する（ステップＳ５１）。この結果、各集合の個々の要素毎に、第１のハッシュ値が得られる。

【００４９】
次に、全体集合生成部５２が、全体集合を生成する（ステップＳ５２）。ステップＳ５２において、全体集合生成部５２は、各集合の個々の要素のうち、第１のハッシュ値が一致し、かつ、要素自体が一致する複数の要素からは１つの要素だけを取り出すとともに、その複数の要素に該当しない各要素を取り出し、取り出した各要素を含む１つの集合を全体集合とすればよい。

【００５０】
次に、インデックス情報生成部５３が、インデックス情報を生成する（ステップＳ５３

）。例えば、インデックス情報生成部５３は、全体集合に属する各要素に対して識別情報を割り当てる。そして、インデックス情報生成部５３は、複数の集合に含まれる各集合の各要素毎に、その要素に該当する全体集合内の要素の識別情報を定めた情報を、インデックス情報として生成する。

【００５１】

次に、第２のハッシュ値計算部５４が、全体集合に属する各要素に対して、所定のハッシュ関数ｈ１以外の各ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値をそれぞれ計算する（ステップＳ５４）。

【００５２】

次に、ハッシュ値決定部５５が、インデックス情報に基づいて、各集合の個々の要素の第１のハッシュ値と、ステップＳ５４で計算された各ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値とを組み合わせることによって、各集合の個々の要素に対して、複数のハッシュ関数ｈ１，ｈ２，・・・，ｈｍに対応する複数のハッシュ値を決定する（ステップＳ５５）。ステップＳ５の結果、例えば、図８に例示するように、各集合の個々の要素に対して、複数のハッシュ関数ｈ１，ｈ２，・・・，ｈｍに対応する複数のハッシュ値が得られる。

【００５３】

ステップＳ５１～ステップＳ５５は、第１の実施形態のステップＳ１をより具体化した処理の一例である。

【００５４】

ステップＳ５５の次に、最小ハッシュ値特定部３が、複数の集合に含まれるそれぞれの集合に関して、個々のハッシュ関数毎に最小ハッシュ値を特定する（ステップＳ２）。ステップＳ２の結果、例えば、図９に例示するように、各集合に関して、ハッシュ関数毎に最小ハッシュ値が特定される。

【００５５】

次に、類似度導出部４が、複数の集合から得られる１組以上の集合のペアに関して、ペアをなす２つの集合の類似度を導出する（ステップＳ３）。

【００５６】

ステップＳ２，Ｓ３は、第１の実施形態におけるステップＳ２，Ｓ３と同様である。

【００５７】

以下、ステップＳ３についてより具体的に説明する。図１１は、ステップＳ３の動作の具体例の一例を示すフローチャートである。なお、類似度導出システム１に与えられる集合の数は、２つであっても、３つ以上であってもよい。以下の説明では、類似度導出システム１が、与えられた集合から得られる全てのペア（集合のペア）に関して、ペアをなす２つの集合の類似度を算出する場合を示す。ただし、類似度導出システム１に与えられた集合から得られる全てのペアのうち、例えば、類似度導出システム１の操作者によって指定されたペアに関してのみ、類似度を算出してもよい。

【００５８】

類似度導出部４は、複数の集合から、集合のペアを１つ取り出す（ステップＳ６１）。１つのペアは、２つの集合からなる。また、ステップＳ６１では、類似度導出部４は、まだ選択されていないペアを１つ取り出す。

【００５９】

類似度導出部４は、ステップＳ６１で選択されたペアをなす２つの集合の類似度を、その２つの集合における個々のハッシュ関数に対応する最小ハッシュ値に基づいて導出する（ステップＳ６２）。例えば、類似度導出部４は、ペアをなす２つの集合に関して、個々のハッシュ値に対応する最小ハッシュ値同士が一致している数（以下、一致数と記す。）を、ペアをなす２つの集合の類似度として定めてもよい。また、例えば、類似度導出部４は、ハッシュ関数の数に対する一致数の割合を、類似度として定めてもよい。

【００６０】

ステップＳ６２の次に、類似度導出部４は、未選択の集合のペアが存在するか否かを判

定する（ステップＳ６３）。
【００６１】
　未選択の集合のペアが存在するならば（ステップＳ６３のＹｅｓ）、類似度導出部４は、ステップＳ６１以降の動作を繰り返す。
【００６２】
　未選択の集合のペアが存在しないならば（ステップＳ６３のＮｏ）、全てのペアに関して、ペアをなす２つの集合の類似度が導出されていることになるので、その時点で処理を終了する。
【００６３】
　本実施形態によれば、全体集合生成部５２が、各集合の個々の要素のうち、第１のハッシュ値が一致し、かつ、要素自体が一致する複数の要素からは１つの要素だけを取り出すとともに、その複数の要素に該当しない各要素を取り出し、取り出した各要素を含む１つの全体集合を生成する。そして、第２のハッシュ値計算部５４が、全体集合に属する各要素に対して、所定のハッシュ関数ｈ１以外の各ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値をそれぞれ計算する。従って、第１のハッシュ値が一致し、かつ、要素自体が一致する複数の要素については、第２のハッシュ値計算部５４は、ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値の計算を１回のみ行う。そのような複数の要素それぞれに対して、ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値の計算を行っても、同一の計算を行うことになり、同一のハッシュ値が得られるが、上記のように、第２のハッシュ値計算部５４は、そのような複数の要素については、ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値の計算を１回のみ行う。従って、同一の計算を重複して行うことがなく、集合の類似度を導出する際に、ハッシュ値の計算量を低減させることができる。
【００６４】
実施形態３．
　第３の実施形態は、第１の実施形態をより具体的に示した実施形態である。図１２は、本発明の第３の実施形態の類似度導出システムの例を示すブロック図である。本実施形態の類似度導出システム１も、ハッシュ値計算部２と、最小ハッシュ値特定部３と、類似度導出部４とを備える。最小ハッシュ値特定部３および類似度導出部４は、第１の実施形態および第２の実施形態における最小ハッシュ値特定部３および類似度導出部４と同様である。
【００６５】
　また、本実施の形態においても、第２の実施形態と同様に、複数のハッシュ関数の数は、ｍ個であり、個々のハッシュ関数をｈ１，ｈ２，・・・，ｈｍと記すこととする。そして、ｈ１が、所定のハッシュ関数であるものとする。
【００６６】
　既に説明したように、複数のハッシュ関数の中で値域が最も広いハッシュ関数を所定のハッシュ関数として定めることが好ましい。すなわち、所定のハッシュ関数は、複数のハッシュ関数の中で最も値域が広いハッシュ関数であることが好ましい。本例では、ハッシュ関数ｈ１の値域が、複数のハッシュ関数ｈ１～ｈｍの値域の中で最も広いものとする。
【００６７】
　また、所定のハッシュ関数ｈ１で計算されたハッシュ値を、第１のハッシュ値と記す。
【００６８】
　第３の実施形態のハッシュ値計算部２は、集合選択部６１と、要素選択部６２と、第１のハッシュ値計算部６３と、判定部６４と、第２のハッシュ値計算部６５とを備える。なお、第３の実施形態における第１のハッシュ値計算部６３および第２のハッシュ値計算部６５の動作は、第２の実施形態における第１のハッシュ値計算部５１および第２のハッシュ値計算部５４の動作とは異なる。
【００６９】
　集合選択部６１は、複数の集合から、順次、１つの集合を選択する。
【００７０】

要素選択部６２は、集合選択部６１によって選択された集合から、順次、１つの要素を選択する。

【００７１】

第１のハッシュ値計算部６３は、要素選択部６２によって選択された要素に所定のハッシュ関数を適用することによって、その選択された要素の第１のハッシュ値を計算する。

【００７２】

判定部６４は、要素選択部６２によって選択された要素と、第１のハッシュ値が一致し、かつ、要素自体が一致する要素が既に選択されているか否かを判定する。

【００７３】

以下、要素選択部６２によって選択された要素と、第１のハッシュ値が一致し、かつ、要素自体が一致する要素を、一致要素と記す。

【００７４】

第２のハッシュ値計算部６５は、一致要素が既に選択されている場合に、要素選択部６２によって選択された要素の所定のハッシュ関数ｈ１以外の各ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値を、その一致要素の各ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値と同一であると定める。

【００７５】

また、第２のハッシュ値計算部６５は、一致要素が 選択されていない場合に、要素選択部６２によって選択された要素の所定のハッシュ関数ｈ１以外の各ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値を計算する。

【００７６】

集合選択部６１、要素選択部６２、第１のハッシュ値計算部６３、判定部６４、第２のハッシュ値計算部６５を含むハッシュ値計算部２は、例えば、類似度導出プログラムに従って動作するコンピュータのＣＰＵによって実現される。例えば、ＣＰＵが、コンピュータのプログラム記憶装置等のプログラム記録媒体から類似度導出プログラムを読み込み、その類似度導出プログラムに従って、集合選択部６１、要素選択部６２、第１のハッシュ値計算部６３、判定部６４、第２のハッシュ値計算部６５を含むハッシュ値計算部２として動作すればよい。

【００７７】

次に、第３の実施形態の処理経過の例を示す。なお、既に説明した事項については、適宜、説明を省略する。図１３および図１４は、第３の実施形態の処理経過の例を示すフローチャートである。

【００７８】

まず、集合選択部６１が、複数の集合から１つの集合を選択する（ステップＳ７１）。ステップＳ７１で、集合選択部６１は、まだ選択されていない集合を１つ選択する。

【００７９】

本例では、図１５に示す集合が選択されたものとする。以下、図１５に示す集合を集合Ａと記す。なお、図１５は、複数の集合のうちの１つの集合の例を示す模式図である。

【００８０】

次に、要素選択部６２が、ステップＳ７１で選択された集合から１つの要素を選択する（ステップＳ７２）。ステップＳ７２で、要素選択部６２は、まだ選択されていない要素を１つ選択する。ここでは、集合Ａから要素”the”を選択したものとする。

【００８１】

次に、第１のハッシュ値計算部６３が、ステップＳ７２で選択された要素に所定のハッシュ関数ｈ１を適用することによって、その要素の第１のハッシュ値を計算する（ステップＳ７３）。図１６は、最初にステップＳ７３を実行した後に得られたハッシュ値の例を示す模式図である。図１６に示すように、この時点では、要素”the”の第１のハッシュ値が得られているが、他のハッシュ値は得られていない。

【００８２】

ステップＳ７３の後に、判定部６４は、ステップＳ７２で選択された要素と、第１のハ

ッシュ値が一致し、かつ、要素自体が一致する要素（すなわち、一致要素）が既に選択されているか否かを判定する（ステップＳ７４）。

【００８３】

一致要素が既に選択されている場合（ステップＳ７４のＹｅｓ）、ステップＳ７５に移行し、一致要素が選択されていない場合（ステップＳ７４のＮｏ）、ステップＳ７６に移行する。本例では、集合Ａの要素”the”は、１番目に選択された要素であり、一致要素は選択されていない。したがって、ステップＳ７６に移行する。

【００８４】

ステップＳ７６では、第２のハッシュ値計算部６５が、ステップＳ７２で選択された要素の所定のハッシュ関数ｈ１以外の各ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値を計算する。図１７は、最初にステップＳ７６を実行した後に得られたハッシュ値の例を示す模式図である。図１７に示すように、この時点では、要素”the”の複数のハッシュ関数ｈ１，ｈ２・・・，ｈｍに対応するハッシュ値が得られている。

【００８５】

ステップＳ７５またはステップＳ７６の実行後には、要素選択部６２が、選択されている集合の中に未選択の要素があるか否かを判定する（ステップＳ７７）。本例では、選択されている集合Ａの中に未選択の要素があると判定する（ステップＳ７７のＹｅｓ）。この場合、ステップＳ７２以降の処理を繰り返す。

【００８６】

集合Ａの要素は全て異なる要素であるので（図１５参照）、集合Ａを選択しているときにステップＳ７４では、一致要素は選択されていないと判定され、ステップＳ７６に移行する。図１８は、ステップＳ７７で未選択の要素がないと判定されるまでに得られたハッシュ値の例を示す模式図である。

【００８７】

ステップＳ７７で、選択している集合の中に未選択の要素がないと判定した場合（ステップＳ７７のＮｏ）、ステップＳ７８に移行する。ステップＳ７８では、集合選択部６１が、未選択の集合があるか否かを判定する。

【００８８】

未選択の集合があると判定した場合（ステップＳ７８のＹｅｓ）、ステップＳ７１以降の処理を繰り返す。ここでは、まだ、図１９に示す集合Ｂが選択されていないものとする。なお、図１９は、複数の集合のうちの１つの集合の例を示す模式図である。

【００８９】

ステップＳ７１で集合選択部６１が集合Ｂを選択し、ステップＳ７２で要素選択部６２が集合Ｂの要素”room”を選択したとする。その後、ステップＳ７３が実行される。この場合、要素”room”に応じた一致要素は選択されていないので（ステップＳ７４のＮｏ）、ステップＳ７６に移行する。この結果、集合Ｂの要素”room” の複数のハッシュ関数ｈ１，ｈ２・・・，ｈｍに対応するハッシュ値が得られる。

【００９０】

次に、ステップＳ７２で、要素選択部６２が、集合Ｂの要素”mountain”を選択したとする。すると、第１のハッシュ値計算部６３は、集合Ｂの要素”mountain”に所定のハッシュ関数を適用することによって、その要素の第１のハッシュ値を計算する（ステップＳ７３）。図２０は、集合Ｂの要素”mountain”が選択され、ステップＳ７３を実行した後に得られたハッシュ値の例を示す模式図である。

【００９１】

次に、判定部６４は、ステップＳ７２で選択された要素（ここでは、集合Ｂの要素”mountain”）と、第１のハッシュ値が一致し、かつ、要素自体が一致する要素（すなわち、一致要素）が既に選択されているか否かを判定する（ステップＳ７４）。

【００９２】

この時点では、集合Ｂの要素”mountain”と、第１のハッシュ値が一致し、かつ、要素自体が一致する要素である「集合Ａの要素”mountain”」が選択済みである（ステップＳ

７４のＹｅｓ）。従って、ステップＳ７５に移行する。なお、この場合、「集合Ａの要素”mountain”」が一致要素に該当する。

【００９３】

ステップＳ７５では、第２のハッシュ値計算部６５は、ステップＳ７２で選択された「集合Ｂの要素”mountain”」の所定のハッシュ関数ｈ１以外の各ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値を、一致要素（集合Ａの要素”mountain”）の各ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値と同一であると定める。図２１は、集合Ｂの要素”mountain”を選択したときにおけるステップＳ７５の実行後に得られているハッシュ値の例を示す模式図である。

【００９４】

未選択の要素”view”が存在するので（ステップＳ７７のＹｅｓ）、ステップＳ７２で、要素選択部６２が、集合Ｂの要素”view”を選択し、ステップＳ７３以降の処理を繰り返す。この場合、要素”view”に応じた一致要素は選択されていないので（ステップＳ７４のＮｏ）、ステップＳ７６に移行する。この結果、集合Ｂの要素”view” の複数のハッシュ関数ｈ１，ｈ２・・・，ｈｍに対応するハッシュ値が得られる。

【００９５】

ここで、集合Ｂの要素は全て選択された状態になっているので（ステップＳ７７のＮｏ）、ステップＳ７８に移行する。図２２は、この時点で得られているハッシュ値の例を示す模式図である。

【００９６】

ステップＳ７８で、判定部６４が、未選択の集合がないと判定したとする（ステップＳ７８のＮｏ）。この場合、ステップＳ２（図１４参照）に移行する。

【００９７】

ステップＳ２に移行するまでの処理（ステップＳ７８で未選択の集合がないと判定するまでの処理）は、第１の実施形態のステップＳ１をより具体化した処理の一例である。

【００９８】

ステップＳ２では、最小ハッシュ値特定部３が、複数の集合に含まれるそれぞれの集合に関して、個々のハッシュ関数毎に最小ハッシュ値を特定する。本例では、複数の集合が、２つの集合Ａ，Ｂであるとする。この場合、ステップＳ２の結果、第２の実施形態で示した図９に例示するように、各集合に関して、ハッシュ関数毎に最小ハッシュ値が特定される。

【００９９】

ステップＳ２の次に、類似度導出部４が、複数の集合から得られる１組以上の集合のペアに関して、ペアをなす２つの集合の類似度を導出する（ステップＳ３）。

【０１００】

ステップＳ２，Ｓ３は、第１の実施形態および第２の実施形態におけるステップＳ２，Ｓ３と同様である。

【０１０１】

ステップＳ３のより具体的な動作の例については、第２の実施形態で図１１を参照して説明したので、ここでは説明を省略する。

【０１０２】

本実施形態によれば、判定部６４が、ステップＳ７２で選択された要素に応じた一致要素が既に選択済みであると判定した場合（ステップＳ７４のＹｅｓ）、ステップＳ７５に移行する。一致要素の各ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値は、既に計算されている。そして、ステップＳ７５において、第２のハッシュ値計算部６５は、ステップＳ７２で選択された要素の各ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値を、一致要素の各ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値と同一であると定める。ステップＳ７２で選択された要素に応じた一致要素は、ステップＳ７２で選択された要素と一致する。従って、ステップＳ７５において、第２のハッシュ値計算部６５は、ステップＳ７２で選択された要素の各ハッシュ関数ｈ２，・・・，ｈｍに対応するハ

ッシュ値を、一致要素の各ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値と同一であると定めることができる。そして、ステップＳ７２で選択された要素に応じた一致要素が既に選択済みである場合に、ステップＳ７５に移行する。その結果、既に計算された各ハッシュ関数ｈ２，・・・，ｈｍに対応するハッシュ値を再度計算する必要がない。よって、集合の類似度を導出する際に、ハッシュ値の計算量を低減させることができる。

【０１０３】

次に、各実施形態の変形例について説明する。この変形例では、複数の集合に含まれる各集合の各要素を、複数のハッシュ関数ｈ１，ｈ２，・・・，ｈｍとは別のハッシュ関数（ｈ０と記す。）によって、ハッシュ値（数値）に変換する。このハッシュ関数ｈ０を数値化ハッシュ関数と記す。例えば、各集合の各要素が文字列である場合、それぞれの文字列に数値化ハッシュ関数ｈ０を適用して、各集合の各要素を文字列から数値に変換する。

【０１０４】

そして、変換後の各要素を、各集合の各要素として、上記の第１の実施形態、第２の実施形態、第３の実施形態を適用してもよい。文字列に対して複数のハッシュ関数のハッシュ値を計算する計算量よりも、数値に対して複数のハッシュ関数のハッシュ値を計算する計算量の方が少ない。従って、上記のように、各集合の各要素を文字列から数値に変換することで計算量を少なくすることができる。

【０１０５】

ただし、数値化ハッシュ関数ｈ０による変換後の各要素は、数値である。この場合、第１の実施形態のステップＳ１では、ハッシュ値計算部２は、数値化ハッシュ関数ｈ０による変換後のハッシュ値が一致する複数の要素に関しては、複数のハッシュ関数の計算の重複を排除し、各集合の個々の要素に対して複数のハッシュ値を求めればよい。

【０１０６】

また、本変形例を第２の実施形態に適用する場合には、第１のハッシュ値計算部５１が、数値化ハッシュ関数ｈ０を用いて、各集合の各要素を数値に変換する。そして、全体集合生成部５２は、各集合の個々の要素（変換後の要素）のうち、数値化ハッシュ関数ｈ０による変換後のハッシュ値が一致する複数の要素からは１つの要素だけを取り出すとともに、その複数の要素に該当しない各要素を取り出すことによって、重複なく全ての種類の要素を１つ含む１つの全体集合を生成すればよい。また、この場合、第２のハッシュ値計算部５４は、全体集合に属する各要素に対して、複数のハッシュ関数ｈ１，ｈ２，・・・，ｈｍに対応するハッシュ値をそれぞれ計算する。ハッシュ値決定部５５は、インデックス情報に基づいて、各集合の個々の要素に対して、複数のハッシュ関数ｈ１，ｈ２，・・・，ｈｍに対応するハッシュ値を特定する。

【０１０７】

また、本変形例を第３の実施形態に適用する場合には、ステップＳ７３において、第１のハッシュ値計算部６３が、ステップＳ７２で選択された要素に数値化ハッシュ関数ｈ０を適用して、その要素を数値の要素に変換する。ステップＳ７４では、判定部６４は、数値化ハッシュ関数ｈ０で変換された要素と一致する要素を一致要素とし、一致要素が既に得られているか否かを判定すればよい。そして、一致要素が得られていない場合、ステップＳ７６で、第２のハッシュ値計算部６５は、変換後の要素に対して、複数のハッシュ関数ｈ１，ｈ２，・・・，ｈｍを適用することによって、複数のハッシュ関数ｈ１，ｈ２，・・・，ｈｍに対応する各ハッシュ値を計算する。また、一致要素が得られている場合、ステップＳ７５で、第２のハッシュ値計算部６５は、変換後の要素の複数のハッシュ関数ｈ１，ｈ２，・・・，ｈｍに対応する各ハッシュ値を、一致要素の複数のハッシュ関数ｈ１，ｈ２，・・・，ｈｍに対応する各ハッシュ値と同一であると定める。

【０１０８】

図２３は、本発明の実施形態の類似度導出システム１に係るコンピュータの構成例を示す概略ブロック図である。コンピュータ１０００は、ＣＰＵ１００１と、主記憶装置１００２と、補助記憶装置１００３と、インタフェース１００４とを備える。

【０１０９】

本発明の各実施形態の類似度導出システム１は、例えば、コンピュータ１０００によって実現される。類似度導出システム１の動作は、類似度導出プログラムの形式で補助記憶装置１００３に記憶されている。ＣＰＵ１００１は、その類似度導出プログラムを読み出し、類似度導出プログラムを主記憶装置１００２に展開し、その類似度導出プログラムに従って、上記の各実施形態で説明した処理を実行する。

【０１１０】

補助記憶装置１００３は、一時的でない有形の媒体の例である。一時的でない有形の媒体の他の例として、インタフェース１００４を介して接続される磁気ディスク、光磁気ディスク、ＣＤ－ＲＯＭ（Compact Disk Read Only Memory ）、ＤＶＤ－ＲＯＭ（Digital Versatile Disk Read Only Memory ）、半導体メモリ等が挙げられる。また、プログラムが通信回線によってコンピュータ１０００に配信される場合、配信を受けたコンピュータ１０００がそのプログラムを主記憶装置１００２に展開し、そのプログラムに従って上記の実施形態で説明した処理を実行してもよい。

【０１１１】

また、各構成要素の一部または全部は、汎用または専用の回路（circuitry ）、プロセッサ等やこれらの組合せによって実現されてもよい。これらは、単一のチップによって構成されてもよいし、バスを介して接続される複数のチップによって構成されてもよい。各構成要素の一部または全部は、上述した回路等とプログラムとの組合せによって実現されてもよい。

【０１１２】

各構成要素の一部または全部が複数の情報処理装置や回路等により実現される場合には、複数の情報処理装置や回路等は集中配置されてもよいし、分散配置されてもよい。例えば、情報処理装置や回路等は、クライアントアンドサーバシステム、クラウドコンピューティングシステム等、各々が通信ネットワークを介して接続される形態として実現されてもよい。

【０１１３】

上記の本発明の実施形態は、以下の付記のようにも記載され得るが、以下に限定されるわけではない。

【０１１４】

（付記１）

複数の集合に含まれる各集合の個々の要素に対して複数のハッシュ関数を適用して得られる複数のハッシュ値を求める際に、前記複数のハッシュ関数のうちの所定のハッシュ関数によって得られるハッシュ値が一致し、かつ、要素自体が一致する複数の要素に関しては、前記所定のハッシュ関数以外の各ハッシュ関数の計算の重複を排除し、前記各集合の個々の要素に対して前記複数のハッシュ値を求めるハッシュ値計算手段と、

前記複数の集合に含まれるそれぞれの集合に関して、個々のハッシュ関数毎に、ハッシュ値の最小値である最小ハッシュ値を特定する最小ハッシュ値特定手段と、

前記複数の集合から得られる１組以上の集合のペアに関して、ペアをなす２つの集合の類似度を、個々のハッシュ関数に対応する最小ハッシュ値に基づいて導出する類似度導出手段とを備える

ことを特徴とする類似度導出システム。

【０１１５】

（付記２）

前記ハッシュ値計算手段は、

前記各集合の個々の要素に対して前記所定のハッシュ関数を適用することによって第１のハッシュ値を計算する第１のハッシュ値計算手段と、

前記各集合の個々の要素のうち、前記第１のハッシュ値が一致し、かつ、要素自体が一致する複数の要素からは１つの要素だけを取り出すとともに、前記複数の要素に該当しない各要素を取り出すことによって、重複なく全ての種類の要素を含む１つの集合である全体集合を生成する全体集合生成手段と、

前記複数の集合に含まれるどの集合のどの要素が前記全体集合のどの要素に該当するかを示すインデックス情報を生成するインデックス情報生成手段と、
前記全体集合に属する各要素に対して、前記複数のハッシュ関数のうちの前記所定のハッシュ関数以外の各ハッシュ関数を適用することによって、前記各ハッシュ関数に対応するハッシュ値を計算する第２のハッシュ値計算手段と、
前記インデックス情報に基づいて、前記各集合の個々の要素の前記第１のハッシュ値と、当該第１のハッシュ値に対応する、前記第２のハッシュ値計算手段によって計算されたハッシュ値とを組み合わせることによって、前記各集合の個々の要素に対して、前記複数のハッシュ関数に対応する複数のハッシュ値を決定するハッシュ値決定手段とを含む
付記１に記載の類似度導出システム。

【０１１６】
（付記３）
前記ハッシュ値計算手段は、
前記複数の集合から、順次、１つの集合を選択する集合選択手段と、
選択された集合から、順次、１つの要素を選択する要素選択手段と、
選択された要素に前記所定のハッシュ関数を適用することによって前記選択された要素の第１のハッシュ値を計算する第１のハッシュ値計算手段と、
前記選択された要素と、第１のハッシュ値が一致し、かつ、要素自体が一致する要素である一致要素が既に選択されているか否かを判定する判定手段と、
前記一致要素が既に選択されている場合に、前記選択された要素の前記所定のハッシュ関数以外の各ハッシュ関数に対応するハッシュ値を、前記一致要素の前記各ハッシュ関数に対応するハッシュ値と同一であると定め、
前記一致要素が選択されていない場合に、前記選択された要素の前記所定のハッシュ関数以外の各ハッシュ関数に対応するハッシュ値を計算する第２のハッシュ値計算手段とを含む
付記１に記載の類似度導出システム。

【０１１７】
（付記４）
前記所定のハッシュ関数は、前記複数のハッシュ関数の中で最も値域が広いハッシュ関数である
付記１から付記３のうちのいずれか１項に記載の類似度導出システム。

【０１１８】
（付記５）
前記類似度導出手段は、
個々のハッシュ関数に対応する最小ハッシュ値同士が一致している数を、ペアをなす２つの集合の類似度として定める
付記１から付記４のうちのいずれか１項に記載の類似度導出システム。

【０１１９】
（付記６）
複数の集合に含まれる各集合の個々の要素に対して複数のハッシュ関数を適用して得られる複数のハッシュ値を求める際に、前記複数のハッシュ関数のうちの所定のハッシュ関数によって得られるハッシュ値が一致し、かつ、要素自体が一致する複数の要素に関しては、前記所定のハッシュ関数以外の各ハッシュ関数の計算の重複を排除し、前記各集合の個々の要素に対して前記複数のハッシュ値を求めるハッシュ値計算処理と、
前記複数の集合に含まれるそれぞれの集合に関して、個々のハッシュ関数毎に、ハッシュ値の最小値である最小ハッシュ値を特定する最小ハッシュ値特定処理と、
前記複数の集合から得られる１組以上の集合のペアに関して、ペアをなす２つの集合の類似度を、個々のハッシュ関数に対応する最小ハッシュ値に基づいて導出する類似度導出処理とを含む
ことを特徴とする類似度導出方法。

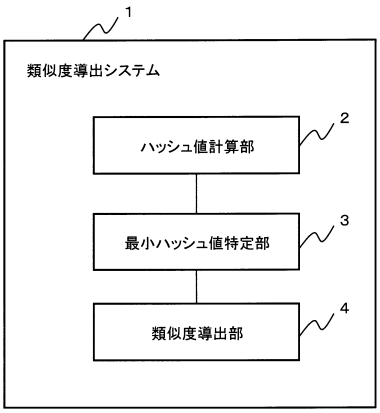
【０１２０】
（付記７）
前記ハッシュ値計算処理は、
前記各集合の個々の要素に対して前記所定のハッシュ関数を適用することによって第１のハッシュ値を計算する第１のハッシュ値計算処理と、
前記各集合の個々の要素のうち、前記第１のハッシュ値が一致し、かつ、要素自体が一致する複数の要素からは１つの要素だけを取り出すとともに、前記複数の要素に該当しない各要素を取り出すことによって、重複なく全ての種類の要素を含む１つの集合である全体集合を生成する全体集合生成処理と、
前記複数の集合に含まれるどの集合のどの要素が前記全体集合のどの要素に該当するかを示すインデックス情報を生成するインデックス情報生成処理と、
前記全体集合に属する各要素に対して、前記複数のハッシュ関数のうちの前記所定のハッシュ関数以外の各ハッシュ関数を適用することによって、前記各ハッシュ関数に対応するハッシュ値を計算する第２のハッシュ値計算処理と、
前記インデックス情報に基づいて、前記各集合の個々の要素の前記第１のハッシュ値と、当該第１のハッシュ値に対応する、前記第２のハッシュ値計算処理で計算されたハッシュ値とを組み合わせることによって、前記各集合の個々の要素に対して、前記複数のハッシュ関数に対応する複数のハッシュ値を決定するハッシュ値決定処理とを含む
付記６に記載の類似度導出方法。
【０１２１】
（付記８）
前記ハッシュ値計算処理は、
前記複数の集合から、順次、１つの集合を選択する集合選択処理と、
選択された集合から、順次、１つの要素を選択する要素選択処理と、
選択された要素に前記所定のハッシュ関数を適用することによって前記選択された要素の第１のハッシュ値を計算する第１のハッシュ値計算処理と、
前記選択された要素と、第１のハッシュ値が一致し、かつ、要素自体が一致する要素である一致要素が既に選択されているか否かを判定する判定処理と、
前記一致要素が既に選択されている場合に、前記選択された要素の前記所定のハッシュ関数以外の各ハッシュ関数に対応するハッシュ値を、前記一致要素の前記各ハッシュ関数に対応するハッシュ値と同一であると定め、
前記一致要素が選択されていない場合に、前記選択された要素の前記所定のハッシュ関数以外の各ハッシュ関数に対応するハッシュ値を計算する第２のハッシュ値計算処理とを含む
付記６に記載の類似度導出方法。
【０１２２】
（付記９）
コンピュータに、
複数の集合に含まれる各集合の個々の要素に対して複数のハッシュ関数を適用して得られる複数のハッシュ値を求める際に、前記複数のハッシュ関数のうちの所定のハッシュ関数によって得られるハッシュ値が一致し、かつ、要素自体が一致する複数の要素に関しては、前記所定のハッシュ関数以外の各ハッシュ関数の計算の重複を排除し、前記各集合の個々の要素に対して前記複数のハッシュ値を求めるハッシュ値計算処理、
前記複数の集合に含まれるそれぞれの集合に関して、個々のハッシュ関数毎に、ハッシュ値の最小値である最小ハッシュ値を特定する最小ハッシュ値特定処理、および、
前記複数の集合から得られる１組以上の集合のペアに関して、ペアをなす２つの集合の類似度を、個々のハッシュ関数に対応する最小ハッシュ値に基づいて導出する類似度導出処理
を実行させるための類似度導出プログラムを記録したコンピュータ読取可能な記録媒体。
【０１２３】

（付記１０）
前記コンピュータに、
前記ハッシュ値計算処理で、
前記各集合の個々の要素に対して前記所定のハッシュ関数を適用することによって第１のハッシュ値を計算する第１のハッシュ値計算処理、
前記各集合の個々の要素のうち、前記第１のハッシュ値が一致し、かつ、要素自体が一致する複数の要素からは１つの要素だけを取り出すとともに、前記複数の要素に該当しない各要素を取り出すことによって、重複なく全ての種類の要素を含む１つの集合である全体集合を生成する全体集合生成処理、
前記複数の集合に含まれるどの集合のどの要素が前記全体集合のどの要素に該当するかを示すインデックス情報を生成するインデックス情報生成処理、
前記全体集合に属する各要素に対して、前記複数のハッシュ関数のうちの前記所定のハッシュ関数以外の各ハッシュ関数を適用することによって、前記各ハッシュ関数に対応するハッシュ値を計算する第２のハッシュ値計算処理、および、
前記インデックス情報に基づいて、前記各集合の個々の要素の前記第１のハッシュ値と、当該第１のハッシュ値に対応する、前記第２のハッシュ値計算処理で計算されたハッシュ値とを組み合わせることによって、前記各集合の個々の要素に対して、前記複数のハッシュ関数に対応する複数のハッシュ値を決定するハッシュ値決定処理を実行させる
類似度導出プログラムを記録した付記９に記載のコンピュータ読取可能な記録媒体。
【０１２４】
（付記１１）
前記コンピュータに、
前記ハッシュ値計算処理で、
前記複数の集合から、順次、１つの集合を選択する集合選択処理、
選択された集合から、順次、１つの要素を選択する要素選択処理、
選択された要素に前記所定のハッシュ関数を適用することによって前記選択された要素の第１のハッシュ値を計算する第１のハッシュ値計算処理、
前記選択された要素と、第１のハッシュ値が一致し、かつ、要素自体が一致する要素である一致要素が既に選択されているか否かを判定する判定処理、および、
前記一致要素が既に選択されている場合に、前記選択された要素の前記所定のハッシュ関数以外の各ハッシュ関数に対応するハッシュ値を、前記一致要素の前記各ハッシュ関数に対応するハッシュ値と同一であると定め、
前記一致要素が選択されていない場合に、前記選択された要素の前記所定のハッシュ関数以外の各ハッシュ関数に対応するハッシュ値を計算する第２のハッシュ値計算処理を実行させる
類似度導出プログラムを記録した付記９に記載のコンピュータ読取可能な記録媒体。
【０１２５】
以上、実施形態を参照して本願発明を説明したが、本願発明は上記の実施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、本願発明のスコープ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。
【産業上の利用の可能性】
【０１２６】
本発明は、集合同士の類似度を導出する類似度導出システムに好適に適用される。
【符号の説明】
【０１２７】
１　類似度導出システム
２　ハッシュ値計算部
３　最小ハッシュ値特定部
４　類似度導出部
５１　第１のハッシュ値計算部

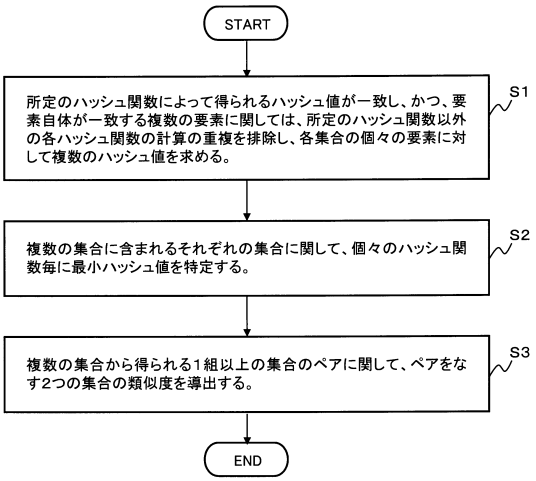
５２　全体集合生成部
５３　インデックス情報生成部
５４　第２のハッシュ値計算部
５５　ハッシュ値決定部
６１　集合選択部
６２　要素選択部
６３　第１のハッシュ値計算部
６４　判定部
６５　第２のハッシュ値計算部

【図面】

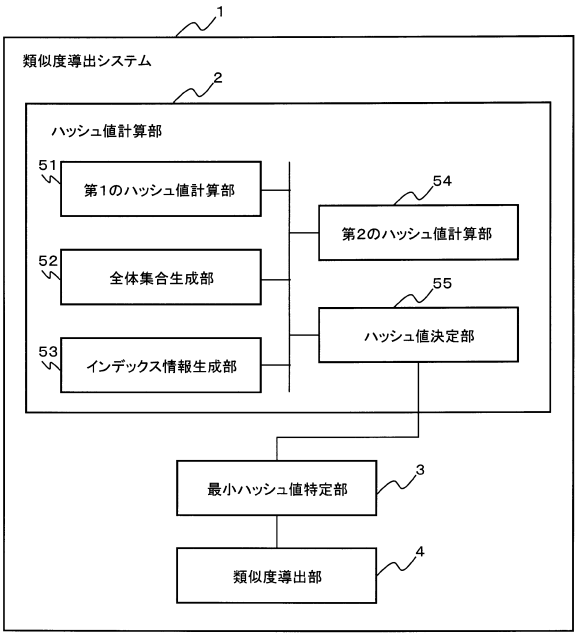
【図１】



【図２】



【図３】



【図４】

集合A

要素	h1
"the"	389
"highest"	12
"mountain"	666
"Fuji"	920

集合B

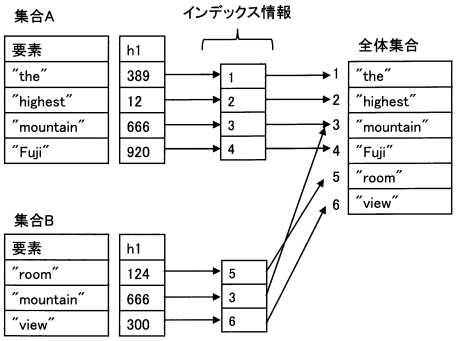
要素	h1
"room"	124
"mountain"	666
"view"	300

【図５】

全体集合

"the"
"highest"
"mountain"
"Fuji"
"room"
"view"

【図６】



【図７】

全体集合	h2	h3	…	hm
"the"	56	143	…	513
"highest"	323	902	…	322
"mountain"	92	792	…	30
"Fuji"	820	43	…	325
"room"	329	469	…	832
"view"	810	521	…	292

【図８】

集合A

要素	h1	h2	h3	…	hm
"the"	389	56	143	…	513
"highest"	12	323	902	…	322
"mountain"	666	92	792	…	30
"Fuji"	920	820	43	…	325

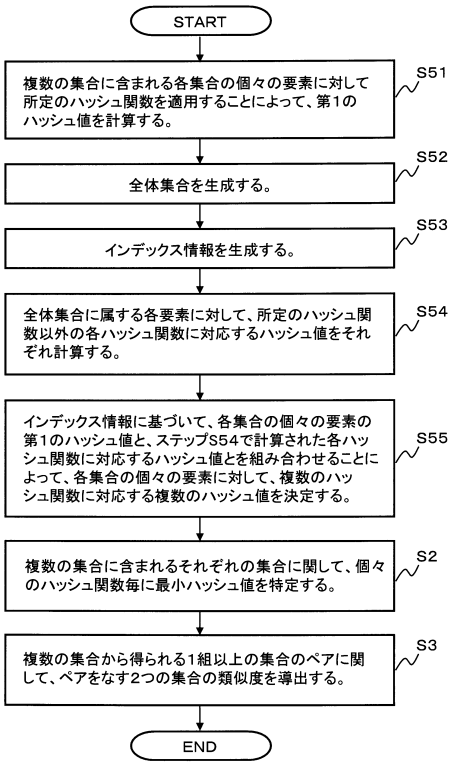
集合B

要素	h1	h2	h3	…	hm
"room"	124	329	469	…	832
"mountain"	666	92	792	…	30
"view"	300	810	521	…	292

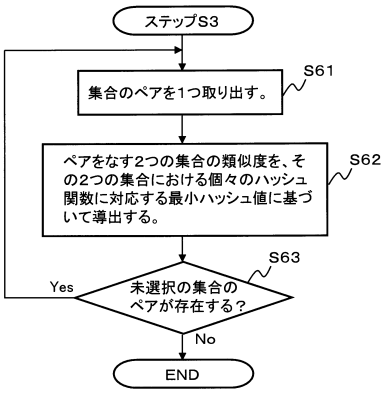
【図９】

	h1	h2	h3	…	hm
集合Aの最小ハッシュ値	12	56	43	…	30
集合Bの最小ハッシュ値	124	92	469	…	30

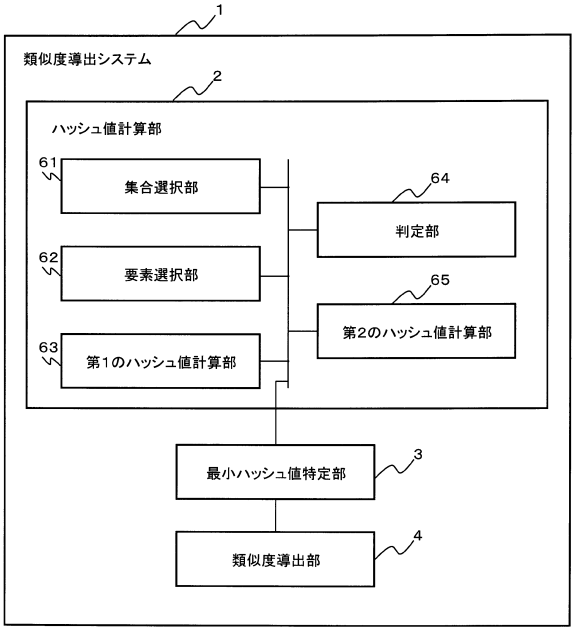
【図１０】



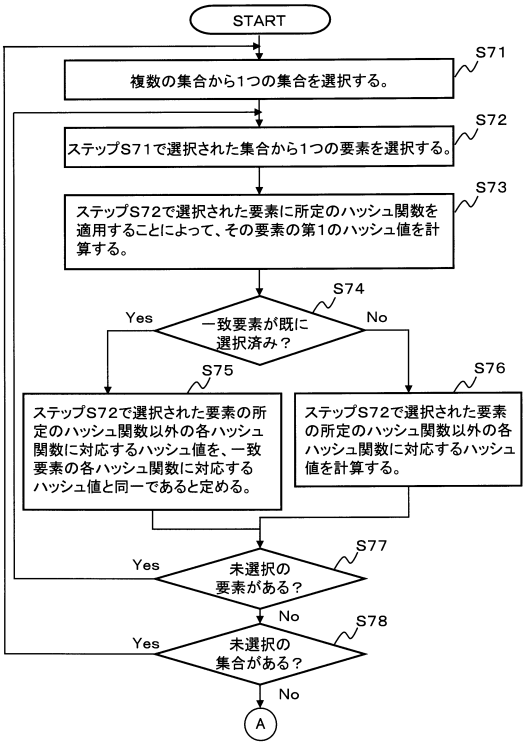
【図１１】



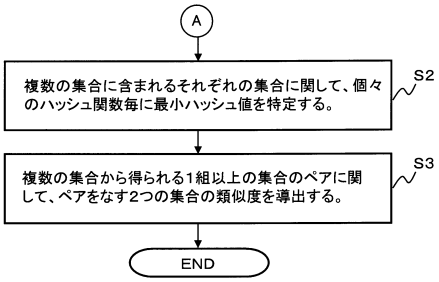
【図１２】



【図１３】



【図１４】



【図１５】

集合A

要素
"the"
"highest"
"mountain"
"Fuji"

【図１６】

集合A

要素	h1	h2	h3	…	hm
"the"	389				
"highest"					
"mountain"					
"Fuji"					

【図１７】

集合A

要素	h1	h2	h3	…	hm
"the"	389	56	143	…	513
"highest"					
"mountain"					
"Fuji"					

【図１８】

集合A

要素	h1	h2	h3	…	hm
"the"	389	56	143	…	513
"highest"	12	323	902	…	322
"mountain"	666	92	792	…	30
"Fuji"	920	820	43	…	325

【図１９】

集合B

要素
"room"
"mountain"
"view"

【図２０】

集合A

要素	h1	h2	h3	…	hm
"the"	389	56	143	…	513
"highest"	12	323	902	…	322
"mountain"	666	92	792	…	30
"Fuji"	920	820	43	…	325

集合B

要素	h1	h2	h3	…	hm
"room"	124	329	469	…	832
"mountain"	666				
"view"					

【図２１】

集合A

要素	h1	h2	h3	…	hm
"the"	389	56	143	…	513
"highest"	12	323	902	…	322
"mountain"	666	92	792	…	30
"Fuji"	920	820	43	…	325

集合B

要素	h1	h2	h3	…	hm
"room"	124	329	469	…	832
"mountain"	666	92	792	…	30
"view"					

【図２２】

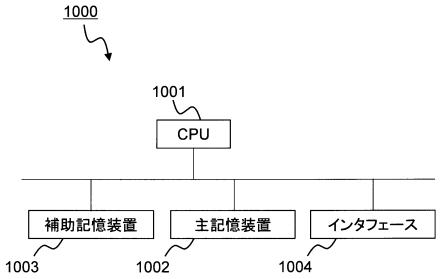
集合A

要素	h1	h2	h3	…	hm
"the"	389	56	143	…	513
"highest"	12	323	902	…	322
"mountain"	666	92	792	…	30
"Fuji"	920	820	43	…	325

集合B

要素	h1	h2	h3	…	hm
"room"	124	329	469	…	832
"mountain"	666	92	792	…	30
"view"	300	810	521	…	292

【図２３】



フロントページの続き

(56)参考文献　国際公開第２０２１／０３８８８７（WO，Ａ１）
米国特許出願公開第２０１８／０１８１６０９（ＵＳ，Ａ１）
米国特許出願公開第２０１８／００９５９４１（ＵＳ，Ａ１）
米国特許出願公開第２０１７／０３２２９３０（ＵＳ，Ａ１）
米国特許出願公開第２０１７／０１６１３７５（ＵＳ，Ａ１）
米国特許出願公開第２０１７／００７８２８６（ＵＳ，Ａ１）

(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１６／００－１６／９５８
Ｇ０９Ｃ　１／００