

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号
特許第5963310号
(P5963310)

(45) 発行日 平成28年8月3日 (2016.8.3)

(24) 登録日 平成28年7月8日 (2016.7.8)

(51) Int.Cl.

F I

GO6F 17/22 (2006.01)

GO6F 17/30 (2006.01)

GO6F 17/22 664

GO6F 17/30 220Z

GO6F 17/30 170A

請求項の数 23 (全 20 頁)

(21) 出願番号	特願2013-15626 (P2013-15626)	(73) 特許権者	390009531
(22) 出願日	平成25年1月30日 (2013.1.30)		インターナショナル・ビジネス・マシーンズ・コーポレーション
(65) 公開番号	特開2014-146257 (P2014-146257A)		INTERNATIONAL BUSINESS MACHINES CORPORATION
(43) 公開日	平成26年8月14日 (2014.8.14)		アメリカ合衆国10504 ニューヨーク州 アーモンク ニュー オーチャードロード
審査請求日	平成27年11月17日 (2015.11.17)		New Orchard Road, Armonk, New York 10504, United States of America
早期審査対象出願		(74) 代理人	100108501 弁理士 上野 剛史
			最終頁に続く

(54) 【発明の名称】 情報処理装置、情報処理方法、及び、情報処理プログラム

(57) 【特許請求の範囲】

【請求項 1】

複数のテキストの中から他のテキストを引用した引用部分を検出する検出部と、
前記複数のテキスト中の引用部分を予め定められた文字列に置換して複数の変換済テキストを生成する変換部と、
前記複数の変換済テキストをテキストマイニングするテキストマイニング部と、
を備える情報処理装置。

【請求項 2】

前記検出部は、一のテキストに含まれる文字列が、一のテキストに含まれる参照先情報により指定される参照先をアクセスして得られる情報の中に含まれる場合に、当該文字列を前記参照先からの引用部分であると判断する照合部を有する
請求項 1 に記載の情報処理装置。

【請求項 3】

複数のテキストの中から他のテキストを引用した引用部分を検出する検出部と、
前記複数のテキスト中の引用部分を削除または予め定められた文字列に置換して複数の変換済テキストを生成する変換部と、
前記複数の変換済テキストをテキストマイニングするテキストマイニング部と、
を備え、
前記検出部は、
一のテキストに含まれる文字列が、一のテキストに含まれる参照先情報により指定され

る参照先をアクセスして得られる情報の中に含まれる場合に、当該文字列を前記参照先からの引用部分であると判断する照合部と、

前記複数のテキストに含まれる参照先情報により指定される参照先を辿り、異なる２以上の参照先情報から同一の情報に辿りつくことを検出する参照先検出部と

を有し、

前記変換部は、前記参照先検出部の検出結果に応じて当該２以上の参照先情報を同一の文字列に置換する参照先変換部を有する情報処理装置。

【請求項４】

前記参照先検出部は、参照先情報により指定される参照先をアクセスして他の参照先にリダイレクトする旨のリダイレクト情報が得られたことに応じて、リダイレクト先をさらに辿る

10

請求項３に記載の情報処理装置。

【請求項５】

前記参照先変換部は、参照先情報により指定される参照先をアクセスして得られた情報に正規の参照先を示す参照先情報が含まれている場合に、参照先情報を正規の参照先情報に置換する

請求項３または４に記載の情報処理装置。

【請求項６】

複数のテキストの中から他のテキストを引用した引用部分を検出する検出部と、

前記複数のテキスト中の引用部分を削除または予め定められた文字列に置換して複数の変換済テキストを生成する変換部と、

20

前記複数の変換済テキストをテキストマイニングするテキストマイニング部と、

を備え、

前記検出部は、

一のテキストに含まれる文字列が、一のテキストに含まれる参照先情報により指定される参照先をアクセスして得られる情報の中に含まれる場合に、当該文字列を前記参照先からの引用部分であると判断する照合部と、

前記複数のテキスト中に同一の文字列を検出したことに応じて、当該文字列を引用部分と判断する判断部と

を有する情報処理装置。

30

【請求項７】

前記判断部は、前記複数のテキスト中から検出した同一の文字列が予め定められた基準文字数以上の長さであることを条件として、当該文字列を引用部分と判断する

請求項６に記載の情報処理装置。

【請求項８】

前記判断部は、前記複数のテキスト中に同一の文字列を予め定められた基準個数以上検出したことを条件として、当該文字列を引用部分と判断する

請求項６または７に記載の情報処理装置。

【請求項９】

前記変換部は、前記複数のテキスト中から検出した同一の文字列が、前記複数のテキストの一のテキストの全体と一致しないことを条件として、前記一のテキスト中の引用部分を削除または予め定められた文字列に置換する

40

請求項６から８のいずれか一項に記載の情報処理装置。

【請求項１０】

前記照合部は、検出済みの２以上の引用部分が共通部分を含んでいる場合に、同一の情報からの引用部分であると判断する

請求項６から９のいずれか一項に記載の情報処理装置。

【請求項１１】

複数のテキストの中から他のテキストを引用した引用部分を検出する検出部と、

前記複数のテキスト中の引用部分を削除または予め定められた文字列に置換して複数の

50

変換済テキストを生成する変換部と、

前記複数の変換済テキストをテキストマイニングするテキストマイニング部と、
を備え、

前記変換部は、前記複数のテキスト中における同一の引用部分を、当該引用部分を識別する識別情報に置換する情報処理装置。

【請求項 1 2】

前記テキストマイニング部は、前記複数の変換済テキスト中における互いに引用内容が異なる引用部分のそれぞれの出現回数を測定する

請求項 1 1 に記載の情報処理装置。

【請求項 1 3】

複数のテキストの中から他のテキストを引用した引用部分を検出する検出部と、

前記複数のテキスト中の引用部分を削除または予め定められた文字列に置換して複数の変換済テキストを生成する変換部と、

前記複数の変換済テキストをテキストマイニングするテキストマイニング部と、
を備え、

前記テキストマイニング部は、互いに引用内容が異なる引用部分同士の類似度を算出し、類似度に基づいて引用部分をグループ化する情報処理装置。

【請求項 1 4】

複数のテキストの中から他のテキストを引用した引用部分を検出する検出部と、

前記複数のテキスト中の引用部分を削除または予め定められた文字列に置換して複数の変換済テキストを生成する変換部と、

前記複数の変換済テキストをテキストマイニングするテキストマイニング部と、
を備え、

前記検出部は、一のテキストに含まれる文字列が、一のテキストに含まれる参照先情報により指定される参照先をアクセスして得られる情報の中に含まれる場合に、当該文字列を前記参照先からの引用部分であると判断する照合部を有し、

前記テキストマイニング部は、互いに引用内容が異なる引用部分に対応付けられた参照先の情報同士の類似度を算出し、類似度に基づいて引用部分をグループ化する情報処理装置。

【請求項 1 5】

複数のテキストの中から他のテキストを引用した引用部分を検出する検出部と、

前記複数のテキスト中の引用部分を削除または予め定められた文字列に置換して複数の変換済テキストを生成する変換部と、

前記複数の変換済テキストをテキストマイニングするテキストマイニング部と、
を備え、

前記検出部は、一のテキストに含まれる文字列が、一のテキストに含まれる参照先情報により指定される参照先をアクセスして得られる情報の中に含まれる場合に、当該文字列を前記参照先からの引用部分であると判断する照合部を有し、

前記テキストマイニング部は、互いに引用内容が異なる 2 以上の引用部分に対応付けられた参照先の情報同士の中に、同一の参照先を指定する参照先情報が含まれている場合に、当該 2 以上の引用部分をグループ化する情報処理装置。

【請求項 1 6】

複数のテキストの中から他のテキストを引用した引用部分を検出する検出段階と、

前記複数のテキスト中の引用部分を予め定められた文字列に置換して複数の変換済テキストを生成する変換段階と、

前記複数の変換済テキストをテキストマイニングするテキストマイニング段階と、
を備える情報処理方法。

【請求項 1 7】

複数のテキストの中から他のテキストを引用した引用部分を検出する検出段階と、

前記複数のテキスト中の引用部分を削除または予め定められた文字列に置換して複数の

10

20

30

40

50

変換済テキストを生成する変換段階と、

前記複数の変換済テキストをテキストマイニングするテキストマイニング段階と、
を備え、

前記検出段階は、

一のテキストに含まれる文字列が、一のテキストに含まれる参照先情報により指定される参照先をアクセスして得られる情報の中に含まれる場合に、当該文字列を前記参照先からの引用部分であると判断する照合段階と、

前記複数のテキストに含まれる参照先情報により指定される参照先を辿り、異なる2以上の参照先情報から同一の情報に辿りつくことを検出する参照先検出段階と

を有し、

前記変換段階は、前記参照先検出段階の検出結果に応じて当該2以上の参照先情報を同一の文字列に置換する参照先変換段階を有する情報処理方法。

【請求項18】

複数のテキストの中から他のテキストを引用した引用部分を検出する検出段階と、

前記複数のテキスト中の引用部分を削除または予め定められた文字列に置換して複数の変換済テキストを生成する変換段階と、

前記複数の変換済テキストをテキストマイニングするテキストマイニング段階と、
を備え、

前記検出段階は、

一のテキストに含まれる文字列が、一のテキストに含まれる参照先情報により指定される参照先をアクセスして得られる情報の中に含まれる場合に、当該文字列を前記参照先からの引用部分であると判断する照合段階と、

前記複数のテキスト中に同一の文字列を検出したことに応じて、当該文字列を引用部分と判断する判断段階と

を有する情報処理方法。

【請求項19】

複数のテキストの中から他のテキストを引用した引用部分を検出する検出段階と、

前記複数のテキスト中の引用部分を削除または予め定められた文字列に置換して複数の変換済テキストを生成する変換段階と、

前記複数の変換済テキストをテキストマイニングするテキストマイニング段階と、
を備え、

前記変換段階は、前記複数のテキスト中における同一の引用部分を、当該引用部分を識別する識別情報に置換する情報処理方法。

【請求項20】

複数のテキストの中から他のテキストを引用した引用部分を検出する検出段階と、

前記複数のテキスト中の引用部分を削除または予め定められた文字列に置換して複数の変換済テキストを生成する変換段階と、

前記複数の変換済テキストをテキストマイニングするテキストマイニング段階と、
を備え、

前記テキストマイニング段階は、互いに引用内容が異なる引用部分同士の類似度を算出し、類似度に基づいて引用部分をグループ化する情報処理方法。

【請求項21】

複数のテキストの中から他のテキストを引用した引用部分を検出する検出段階と、

前記複数のテキスト中の引用部分を削除または予め定められた文字列に置換して複数の変換済テキストを生成する変換段階と、

前記複数の変換済テキストをテキストマイニングするテキストマイニング段階と、
を備え、

前記検出段階は、一のテキストに含まれる文字列が、一のテキストに含まれる参照先情報により指定される参照先をアクセスして得られる情報の中に含まれる場合に、当該文字列を前記参照先からの引用部分であると判断する照合段階を有し、

10

20

30

40

50

前記テキストマイニング段階は、互いに引用内容が異なる引用部分に対応付けられた参照先の情報同士の類似度を算出し、類似度に基づいて引用部分をグループ化する情報処理方法。

【請求項 2 2】

複数のテキストの中から他のテキストを引用した引用部分を検出する検出段階と、
前記複数のテキスト中の引用部分を削除または予め定められた文字列に置換して複数の変換済テキストを生成する変換段階と、
前記複数の変換済テキストをテキストマイニングするテキストマイニング段階と、
を備え、

前記検出段階は、一のテキストに含まれる文字列が、一のテキストに含まれる参照先情報により指定される参照先をアクセスして得られる情報の中に含まれる場合に、当該文字列を前記参照先からの引用部分であると判断する照合段階を有し、

前記テキストマイニング段階は、互いに引用内容が異なる 2 以上の引用部分に対応付けられた参照先の情報同士の中に、同一の参照先を指定する参照先情報が含まれている場合に、当該 2 以上の引用部分をグループ化する情報処理方法。

【請求項 2 3】

コンピュータを、請求項 1 ～ 1 5 のいずれか一項に記載の情報処理装置として機能させるための情報処理プログラム。

【発明の詳細な説明】

【技術分野】

【0 0 0 1】

本発明は、情報処理装置、情報処理方法、及び、情報処理プログラムに関する。

【背景技術】

【0 0 0 2】

ユーザにより作成された多数のテキストを分析することが知られている（例えば、特許文献 1）。

〔特許文献 1〕 特開 2 0 1 1 － 3 1 5 7 号公報

【発明の概要】

【発明が解決しようとする課題】

【0 0 0 3】

しかし、例えば、インターネット上に投稿等されたテキストには、投稿者が自ら作成していない引用箇所が多数含まれることがある。このような場合、テキストを分析するための計算量が増加し、又、その様な引用箇所が多数存在すると引用内容が支配的な情報となり、テキストの正確な分析の妨げとなることがあった。

【課題を解決するための手段】

【0 0 0 4】

本発明の第 1 の態様においては、複数のテキストの中から他のテキストを引用した引用部分を検出する検出部と、複数のテキスト中の引用部分を削除または予め定められた文字列に置換して複数の変換済テキストを生成する変換部と、複数の変換済テキストをテキストマイニングするテキストマイニング部とを備える情報処理装置、当該情報処理装置に実行される方法、及び、コンピュータを当該情報処理装置として機能させるプログラムを提供する。

【0 0 0 5】

なお、上記の発明の概要は、本発明の必要な特徴の全てを列挙したものではない。また、これらの特徴群のサブコンビネーションもまた、発明となりうる。

【図面の簡単な説明】

【0 0 0 6】

【図 1】 本実施形態の情報処理装置 1 0 の構成を示す。

【図 2】 本実施形態の情報処理装置 1 0 の処理フローを示す。

【図 3】 S 1 0 0 において情報処理装置 1 0 が取得する複数のテキストを例示する。

10

20

30

40

50

【図４】Ｓ１０２において参照先検出部１２２が生成する参照テーブルを例示する。
【図５】Ｓ１０４において判断部１２４が実行するＮグラム索引を例示する。
【図６】Ｓ１０４において判断部１２４が実行するＮグラム索引を例示する。
【図７】Ｓ１０４において判断部１２４が生成する引用文字列テーブルを例示する。
【図８】Ｓ１０８及びＳ１１０において照合部１２６が生成する照合テーブルを例示する。

【図９】Ｓ１１２において変換部１４０が生成する複数の変換済テキストを例示する。

【図１０】コンピュータ１９００のハードウェア構成の一例を示す。

【発明を実施するための形態】

【０００７】

10

以下、発明の実施の形態を通じて本発明を説明するが、以下の実施形態は特許請求の範囲にかかる発明を限定するものではない。また、実施形態の中で説明されている特徴の組み合わせの全てが発明の解決手段に必須であるとは限らない。

【０００８】

図１は、本実施形態の情報処理装置１０の構成を示す。情報処理装置１０は、サーバ２０及びサーバ３０から複数のテキストを取得し、複数のテキストから引用部分を検出し、当該引用部分を予め定められた文字列に変換する。情報処理装置１０は、通信部１１０、検出部１２０、記憶部１３０、変換部１４０、及び、テキストマイニング部１５０を備える。

【０００９】

20

通信部１１０は、インターネット等のネットワークに接続し、ネットワークを介して外部の機器と通信する。例えば、通信部１１０は、サーバ２０及びサーバ３０等の外部の機器から複数のテキストを取得する。通信部１１０は、取得した複数のテキストを検出部１２０及び変換部１４０に供給する。

【００１０】

検出部１２０は、複数のテキストの中から、他のテキストを引用した引用部分を検出する。検出部１２０は、参照先検出部１２２、判断部１２４、及び照合部１２６を有する。

【００１１】

参照先検出部１２２は、複数のテキストに含まれる引用部分として参照先情報を検出する。そして、参照先検出部１２２は、検出した異なる２以上の参照先情報から同一の情報に辿りつくか否かを検出する。参照先検出部１２２は、参照先情報としてファイルの場所を示す情報を検出してよく、例えば、Uniform Resource Locator（URL）を検出してよい。また例えば、参照先検出部１２２は、正規のURLと正規のURLをリダイレクト技術により短縮表示した短縮URLとから、同一のウェブサイト等の情報に辿りつくことを検出してよい。

30

【００１２】

参照先検出部１２２は、同一のウェブサイト等の情報に辿りつく２以上の参照先情報のうち、リダイレクト先となる最終的な参照先情報と、最終的な参照先情報の直接的／間接的なリダイレクト元となる１又は複数の参照先情報からなる他の参照先情報とを対応付けた参照テーブルを作成する。参照先検出部１２２は、作成した参照テーブルを記憶部１３０に格納する。

40

【００１３】

判断部１２４は、複数のテキスト中に共通して含まれる同一の文字列を検出したことに応じて、当該文字列を引用部分と判断する。判断部１２４は、検出した同一の文字列が「所定の文字数以上の長さであること」等の予め定められた条件を満たすことを条件として、当該文字列を引用部分と判断してよい。判断部１２４は、複数のテキストから引用部分として検出した文字列から構成される引用文字列テーブルを生成し、当該引用文字列テーブルを記憶部１３０に格納する。

【００１４】

照合部１２６は、記憶部１３０から引用部分として参照先情報を含む参照テーブル、及

50

び、引用部分として文字列を含む引用文字列テーブルを読み出して、これらのテーブルから引用部分ごとに異なる識別情報を付与した照合テーブルを作成する。

【0015】

また、照合部126は、引用文字列テーブルに含まれる検出済みの2以上の文字列が共通部分を含んでいる場合に、同一の情報からの引用部分であると判断してよい。この場合、照合部126は、照合テーブルにおいて当該共通部分を含む2以上の文字列に同一の識別情報を付与してよい。

【0016】

また、照合部126は、一のテキストに含まれる引用部分である文字列が、一のテキストに含まれる参照先情報により指定される参照先をアクセスして得られる情報の中に含まれるか否か判断する。照合部126は、含まれる場合には、当該文字列を参照先からの引用部分であると判断して、照合テーブル中の当該文字列及び参照先情報のレコードを同一の引用部分として統合する。照合部126は、照合テーブルを記憶部130に格納する。

【0017】

記憶部130は、参照先検出部122から受け取った参照テーブル及び判断部124から受け取った引用文字列テーブルを記憶し、これらのテーブルを照合部126に供給する。また、記憶部130は、照合部126から受け取った照合テーブルを記憶し、照合テーブルを変換部140に供給する。記憶部130は、情報処理装置10の主記憶装置、補助記憶装置、又は、情報処理装置10の外部に設けられた記憶装置であってもよい。

【0018】

変換部140は、複数のテキスト中の引用部分を予め定められた文字列に置換して、複数の変換済テキストを生成する。例えば、変換部140は、複数のテキスト中における参照先情報及び／又は同一の文字列を同一の引用部分として、当該引用部分を識別する識別情報に置換する。変換部140は、参照先変換部142及び文字列変換部144を有する。

【0019】

参照先変換部142は、参照先検出部122の検出結果に応じて当該2以上の参照先情報を同一の文字列に置換する。例えば、参照先変換部142は、テキスト中の参照先情報を、照合テーブル中の当該最終的な参照先情報又は「NEWS_TITLE1」等の識別情報に置換する。

【0020】

文字列変換部144は、複数のテキスト中の同一の文字列を識別情報に置換する。例えば、文字列変換部144は、照合テーブル中の文字列と同一の文字列がテキストに含まれる場合、テキストに含まれる同一の文字列を、照合テーブルの当該文字列に対応する「NEWS_TITLE1」等の識別情報に置換する。

【0021】

変換部140は、参照先変換部142及び／又は文字列変換部144において、複数のテキストの引用部分を識別情報に置換する代わりに、引用部分を削除してもよい。変換部140は、参照先変換部142及び／又は文字列変換部144が変換した変換済テキストをテキストマイニング部150に供給する。

【0022】

テキストマイニング部150は、変換部140から複数の変換済テキストを受け取り、当該複数の変換済テキストをテキストマイニングする。例えば、テキストマイニング部150は、複数の変換済テキスト中における互いに引用内容が異なる引用部分のそれぞれの出現回数を測定する。

【0023】

このように、本実施形態の情報処理装置10は、複数のテキスト中に出現する同一の文字列及び参照先情報から検出される同一の引用部分を識別情報に対応付け、複数のテキスト中の引用部分を識別情報に置換すること等により、複数のテキストから引用部分を取り除く。これにより、情報処理装置10は、複数のテキストのうち引用でないオリジナルの

10

20

30

40

50

部分のみをテキストマイニングすることができる。さらに情報処理装置10は、識別情報を分析することにより複数のテキストの引用部分の数の分布など引用の傾向等を分析することができる。

【0024】

図2に、本実施形態の情報処理装置10の処理フローを示す。本実施形態において、情報処理装置10は、S100からS114までの処理を実行する。

【0025】

まず、S100において、通信部110がサーバ20等の外部の機器と通信して複数のテキストを取得する。例えば、通信部110は、インターネット等のネットワークを介して、サーバ20等に格納されるブログ及びソーシャルネットワークサービス等の投稿サイト、及び／又は、ウェブニュース及びメールニュース等のニュースサイトにアクセスして、これらのウェブサイト上で公開されたテキストを取得する。通信部110は、取得した複数のテキストを検出部120及び変換部140に供給する。

10

【0026】

次に、S102において、参照先検出部122は、複数のテキストに含まれる参照先情報を検出し、異なる2以上の参照先情報から同一の情報に辿りつくことを検出する。具体的には、まず参照先検出部122は、複数のテキストからURL等の参照先情報を検出する。

【0027】

参照先検出部122は、通信部110を介して、検出した参照先情報により指定されるウェブページ等の参照先にアクセスし、参照先のHTTPヘッダに含まれるLocation値を検出することにより、他の参照先にリダイレクトする旨のリダイレクト情報を検出する。参照先検出部122は、当該リダイレクト情報が得られたことに応じて、リダイレクト先のURLに対して呼び出しを実行し、リダイレクト先を辿る。

20

【0028】

参照先検出部122は、参照先情報の参照先にリダイレクト情報が検出されない場合は当該参照先情報を最終的な参照先情報とする。また、参照先検出部122は、参照先情報により指定される参照先をアクセスして得られた情報に正規の参照先を示す参照先情報が含まれている場合に、当該正規の参照先を示す参照先情報を最終的な参照先情報としてよい。

30

【0029】

一例として、参照先検出部122は、参照先のウェブページの<Meta>要素に「canonical href」又は「og:url」等のタグで示されるURLが含まれている場合には、当該URLを正規の参照先を示す参照先情報としてよい。

【0030】

参照先検出部122は、同一のウェブサイト等の情報に辿りつく2以上の参照先情報のうち、最後のリダイレクト先となる最終的な参照先情報と、最終的な参照先情報の直接的／間接的なリダイレクト元となる1又は複数の参照先情報からなる他の参照先情報とを対応付けた参照テーブルを作成する。参照先検出部122は、作成した参照テーブルを記憶部130に格納する。

40

【0031】

次に、S104において、判断部124は、複数のテキスト中に同一の文字列を検出する。例えば、判断部124は、複数のテキストに対してNグラム索引を生成し、複数のテキストに共通して含まれる同一の文字列を引用部分として検出する。判断部124が、Nグラム索引を生成する具体的方法等については後述する。

【0032】

判断部124は、複数のテキスト中から検出した同一の文字列が予め定められた基準文字数以上の長さであることを条件として、当該文字列を引用部分と判断してよい。一例として、判断部124は、20文字以上の長さの文字列のみを引用部分と判断してよい。

【0033】

50

これにより、判断部124は、単語の単位で文字列を引用として検出しないので、単に同一の単語及び慣用句等を用いた複数のテキストを引用関係にあると誤認することを防ぐ。また、これにより、判断部124は、引用の程度が低い文字列の処理を回避して情報処理装置10の処理リソースを節約することができる。

【0034】

次に、S106において、判断部124は、検出した同一の文字列が予め定められた条件を満たすことを条件として、当該文字列を引用部分と判断する。例えば、判断部124は、複数のテキスト中に同一の文字列を予め定められた基準個数（例えば、10個）以上検出したことを条件として、当該文字列を引用部分と判断してよい。

【0035】

これにより、判断部124は、例えば、引用回数の少ない重要性の低い文字列を引用部分から除外することができ、情報処理装置10の変換部140による処理負荷を低減することができる。判断部124は、引用部分と判断した文字列から構成される引用文字列テーブルを生成し、当該引用文字列テーブルを記憶部130に格納する。

【0036】

次に、S108において、照合部126は、記憶部130から引用部分として参照先情報を含む参照テーブル、及び、引用部分として文字列を含む引用文字列テーブルを読み出して、これらのテーブルから引用部分ごとに異なる識別情報を付与した照合テーブルを作成する。例えば、照合部126は、引用文字列テーブル中の文字列及び参照テーブル中の参照先情報のそれぞれに対して「NEWS_TITLE1」及び「NEWS_TITLE2」等の異なる識別情報を付与した照合テーブルを作成する。

【0037】

また、照合部126は、照合テーブル中の複数の文字列が共通部分を含むか照合する。共通部分を含む場合、照合部126は、共通部分を含む文字列が同一の情報からの引用部分であると判断して、照合テーブルにおいてこれらの文字列に同一の識別情報を付与する。

【0038】

一例として、照合部126は、共通部分「IBMはPureSystemsを新時代のIT製品として発表」を含む、照合テーブル中の文字列「IBMはPureSystemsを新時代のIT製品として発表した。」と文字列「日本IBMはPureSystemsを新時代のIT製品として発表」とに対して同一の識別情報を付与してよい。

【0039】

次に、S110において、照合部126は、通信部110を介して、照合テーブル中の最終的な参照先情報の参照先にアクセスして、参照先に照合テーブル中の文字列のいずれかが含まれるかを照会する。例えば、照合部126は、照合テーブル中の文字列が、参照先のテキストの少なくとも一部と一致する場合、当該文字列を参照先からの引用部分であると判断する。

【0040】

照合部126は、文字列が引用部分であると判断した場合、照合テーブル中の当該文字列及び参照先情報のレコードを同一の引用部分として統合することにより、当該文字列及び参照先情報に同一の識別情報を付与する。照合部126は、照合テーブルを記憶部130に格納する。

【0041】

次に、S112において、変換部140が複数のテキスト中の引用部分を識別情報等に置換して変換済テキストを生成する。具体的には、参照先変換部142は、記憶部130から照合テーブルを読み出し、テキスト中の参照先情報が照合テーブルの最終的な参照先情報又は他の参照先情報に一致する場合、当該テキスト中の参照先情報を、照合テーブル中の当該最終的な参照先情報に置換するか、又は、参照先情報に対応する「NEWS_TITLE1」等の識別情報に置換する。

【0042】

また、参照先変換部142は、複数のテキスト中の参照先情報が照合テーブルの他の参照先情報に含まれる場合、複数のテキストに含まれる他の参照先情報を当該他の参照先情報に対応する正規の参照先情報に置換してよい。

【0043】

文字列変換部144は、複数のテキスト中の同一の文字列を識別情報に置換する。例えば、文字列変換部144は、照合テーブル中の文字列と同一の文字列がテキストに含まれる場合、テキストに含まれる同一の文字列を、照合テーブルの当該文字列に対応する「NEWS_TITLE1」等の識別情報に置換する。

【0044】

文字列変換部144は、照合テーブルに含まれる文字列が、複数のテキストの一のテキスト全体と一致するか否か判断し、一致しないことを条件として、当該一のテキスト中の引用部分を削除または予め定められた文字列に置換してよい。文字列変換部144は、照合テーブルに含まれる文字列が、複数のテキストの一のテキスト全体と一致する場合は、当該テキストを置換等しなくてよい。これにより、文字列変換部144は、例えば、他のユーザの投稿全体をそのまま再投稿するツイッター（登録商標）のリツイート等のテキストを、引用部分と区別して扱うことができる。

【0045】

また、文字列変換部144は、複数のテキストに含まれる文字列のうち重要性の低いものを削除又は別の文字列に置換してもよい。例えば、文字列変換部144は、正規表現を利用して、宛先を示す文字列（例えば、@とユーザ名を連結した「@Hogehoge」）を検出し、当該文字列を宛先があったことを示す識別情報（例えば、「To_User」）に変換してよい。また、例えば、文字列変換部144は、正規表現を利用してテキストの話題を示す文字列（例えば、#と話題を連結した「#IBM_News」等のタグ）を検出し、当該文字列を削除してよい。

【0046】

なお、変換部140は、通信部110から受け取った複数のテキストを、最初に参照先変換部142において変換し、次に当該複数の変換済テキストを文字列変換部144で変換してよい。これに代えて、変換部140は、複数のテキストを文字列変換部144で変換し、次に参照先変換部142で変換してよい。

【0047】

変換部140は、参照先変換部142及び文字列変換部144のいずれかでのみ複数のテキストの引用部分を変換してもよい。また、変換部140は、参照先変換部142及び／又は文字列変換部144において、複数のテキストの引用部分を識別情報に置換する代わりに削除してもよい。

【0048】

また、変換部140は、参照先変換部142及び文字列変換部144が複数のテキストの引用部分を識別情報に変換した結果、1つの変換済テキストに同一の識別情報が重複して存在することになる場合、いずれか一方の識別情報を削除してよい。変換部140は、参照先変換部142及び／又は文字列変換部144が変換した変換済テキストをテキストマイニング部150に供給する。

【0049】

次に、S114において、テキストマイニング部150は、変換部140から複数の変換済テキストを受け取り、当該複数の変換済テキストをテキストマイニングすることにより、複数のテキストの内容を分析する。例えば、テキストマイニング部150は、IBM Context Analytics (ICA)、Text Network Analysis (TENA)、又はIBM SPSS Text Analytics等の分析ツールにより、テキストマイニングを実行してよい。

【0050】

例えば、テキストマイニング部150は、複数のテキストに含まれる夫々の識別情報の数をカウントすることにより、複数の変換済テキスト中における互いに引用内容が異なる

10

20

30

40

50

引用部分のそれぞれの出現回数を測定する。

【0051】

また、例えば、テキストマイニング部150は、互いに引用内容が異なる引用部分同士の類似度を算出し、類似度に基づいて引用部分をグループ化することにより、複数の変換済テキストをグループ化してもよい。具体的には、テキストマイニング部150は、記憶部130から照合テーブルを読み出し、照合テーブルに含まれる文字列の類似度を、文字列に含まれる単語の意味空間上の距離等に基づいて算出する。

【0052】

次に、テキストマイニング部150は、類似度が予め定められた値以下の文字列をグループ化し、一のグループに含まれる文字列を含む複数の変換済テキストを同一のグループにグループ化する。これにより、テキストマイニング部150は、引用元が異なるが話題が類似する複数のテキストをまとめて分析することができる。

10

【0053】

また、テキストマイニング部150は、互いに引用内容が異なる2以上の引用部分に対応付けられた参照先の情報同士の中に、同一の参照先を指定する参照先情報が含まれている場合に、当該2以上の引用部分をグループ化してよい。具体的には、テキストマイニング部150は、照合テーブルにおいて複数の異なる文字列に対して同一の参照先情報が対応付けられていた場合、これらの異なる文字列を同一のグループとなるにグループ化する。これにより、テキストマイニング部150は、具体的な引用部分が異なっても、内容が類似する可能性が高い引用元が共通する複数のテキストをまとめて分析することができる。

20

【0054】

また、テキストマイニング部150は、互いに引用内容が異なる引用部分に対応付けられた参照先の情報同士の類似度を算出し、類似度に基づいて引用部分をグループ化してもよい。具体的には、テキストマイニング部150は、通信部110を介して、照合テーブルに含まれる複数の参照先情報の参照先にアクセスし、複数の参照先のウェブページ等に含まれるテキスト同士の類似度を、テキストに含まれる単語の意味空間上の距離等に基づいて算出する。

【0055】

次に、テキストマイニング部150は、類似度が予め定められた値以下のテキストを含む複数の参照先情報を同一のグループにグループ化する。これにより、テキストマイニング部150は、内容が類似するウェブサイトを引用する複数のテキストをまとめて分析することができる。

30

【0056】

テキストマイニング部150は、複数のテキストのそれぞれの発信者の影響力の分析、引用部分に対する評価の分析（例えば、賛成又は反対等の感情の分析）、及び／又は、話題となっているトピックの分析（例えば、注目されている単語、ニュース又は人物等の分析）を実行する。

【0057】

このように、本実施形態の情報処理装置10は、複数のテキストの引用部分を識別情報等に変換又は削除して変換済テキストを生成し、当該変換済テキストをテキストマイニングする。これにより、本実施形態の情報処理装置10は、複数のテキストをテキストマイニングする際に、引用部分に対して計算能力を消費することを防ぐことができる。また、情報処理装置10は、テキストマイニングの結果から引用部分の影響を排除することができる。

40

【0058】

なお、図2で説明した本実施形態の処理フローにおいて、文字列変換部144は、複数のテキスト中における重要性の低い文字列の置換等を、S112で実行する代わりにS100の後に実行してもよい。この場合、文字列変換部144は、置換後の複数のテキストを検出部120に供給する。これにより、検出部120は、宛先等が微妙に異なる引用部

50

分を同一の引用部分として検出することができるので、引用部分の検出の精度を向上させることができる。

【0059】

図3は、本実施形態の処理フローのS100において情報処理装置10が取得する複数のテキストを例示する。本実施形態では、図3に示すように、ニュースウェブサイト「IT News (URL: <http://www.XXXXXXitnews.co.jp/news1111>)」において、「日本IBMはPureSystemsを新時代のIT製品として発表した。同社代表取締役によると…(後略)…」という記事が公開され、当該記事の内容を引用する複数のテキスト1～5が投稿される場合を想定する。

【0060】

図示するように、複数のテキスト1～5は、記事の内容を引用した引用部分「日本IBMはPureSystemsを新時代のIT製品として発表した。」を含むが、当該部分は投稿されるテキストにおいてオリジナルな部分ではないので、分析対象としての価値が低い。例えば、テキストマイニング部150が、図3に示す複数のテキストをテキストマイニングすると、IBM、PureSystems、IT及び発表等の単語を頻出文字として集計してしまう。

【0061】

また、テキスト1～5は、記事のURL及び記事のURLの短縮URL (<http://XXX.XX/123XYZ>及び<http://YYY.YY/987AB>)を含むが、これらのURLも投稿されたテキストにおいて本質的にオリジナルな部分ではないので、分析対象としての価値が低い。

【0062】

図4は、S102において参照先検出部122が生成する参照テーブルを例示する。図4に示すように、参照先検出部122は、正規の参照先を示す「最終的な参照先情報(例えば、<http://www.XXXXXXitnews.co.jp/news1111>)」と「他の参照先情報(例えば、最終的な参照先情報の短縮アドレスである<http://XXX.XX/123XYZ>及び<http://YYY.YY/987AB>)」とを対応付けた参照テーブルを生成する。参照先検出部122は、さらに参照先情報のそれぞれに、複数のテキストにおける参照先情報の出現位置を対応付けてもよい。

【0063】

図5及び図6は、S104において判断部124が複数のテキスト1～5に対して実行するNグラム索引を例示する。図5は、判断部124がテキスト1のNグラム索引を生成した例を示す。

【0064】

例えば、図5の表の2行目に示すように、判断部124は、テキスト1「日本IBMはPureSystemsを新時代のIT製品として発表した。：これどんなシステム？」の「1文字目の1グラム索引」としてテキスト1の1文字目からの1文字「日」の索引を生成する。また、判断部124は、当該「日」の前後の1文字を検出する。「日」は1文字目であるので、判断部124は「日」の前の文字を検出しない。判断部124は「日」の後の文字として「本」を検出する。

【0065】

また、表の3行目に示すように、判断部124は、テキスト1の「2文字目の1グラム索引」として「本」の索引を生成し、「前の文字」及び「後の文字」として「日」及び「I」を検出する。同様に、判断部124は、テキスト1の「2文字目の4グラム索引」として「本IBM」の索引を生成し、「前の文字」及び「後の文字」として「日」及び「は」を検出する。

【0066】

このように、判断部124は、文字数nのテキスト1において、 $1 \leq i \leq n-1$ を満たす全ての自然数iに対して、1文字目～n-i+1文字目までのiグラムの索引を生成する。判断部124は、同様にテキスト2～5のNグラム索引を生成する。

【0067】

判断部124は、文字数nのテキストに対してnグラムの索引を生成しなくてよい。これにより、判断部124は、例えば、他のユーザの投稿の全体をそのまま再投稿するリッ

10

20

30

40

50

イート等のテキストを、引用部分として検出することがない。この場合、S 1 1 2 の処理において、文字列変換部 1 4 4 は、複数のテキストの一のテキスト全体と一致するか判断しなくてよい。

【0068】

また、判断部 1 2 4 は、テキストの索引として既にサンプリングされた文字列と同一の文字列を、当該テキストの別の索引としてサンプリングしなくてよい。これにより、判断部 1 2 4 は、同一の文字列について重複して索引を生成することを回避するので、情報処理装置 1 0 の処理リソースを節約できる。

【0069】

図 6 は、判断部 1 2 4 が複数のテキスト 1 ～ 5 の N グラム索引を生成した例を示す。図 6 の上表は、判断部 1 2 4 がテキスト 1 ～ 5 に対して生成した N グラム索引のうち、テキスト 1 の 2 文字目、テキスト 2 の 1 2 文字目、テキスト 4 の 2 文字目、及び、テキスト 5 の 1 5 文字目の 4 グラム索引として生成した文字列「本 IBM」の索引部分を示す。すなわち、判断部 1 2 4 は、N グラム索引により、テキスト 1、2、4 及び 5 に共通して含まれる同一の文字列「本 IBM」を検出する。

10

【0070】

図 6 の下表は、判断部 1 2 4 がテキスト 1 ～ 5 に対して生成した N グラム索引のうち、テキスト 1 の 1 文字目、テキスト 2 の 1 1 文字目、テキスト 4 の 1 文字目、及び、テキスト 5 の 1 4 文字目の 3 4 グラム索引として生成した文字列「日本 IBM は PureSystems を新時代の IT 製品として発表した。」の索引部分を示す。すなわち、判断部 1 2 4 は、N グラム索引により、テキスト 1、2、4 及び 5 に共通して含まれる同一の文字列「日本 IBM は PureSystems を新時代の IT 製品として発表した。」を検出する。

20

【0071】

判断部 1 2 4 は、複数のテキスト 1 ～ 5 に対して N グラム索引を生成した後で、N グラム索引に含まれる文字列のうち、文字列の前後の文字が複数のテキストにわたって共通していない文字列を、引用部分として検出する。

【0072】

例えば、上表において、複数のテキストに共通する文字列「本 IBM」の前後の文字は「日」及び「は」でテキスト 1、2、4 及び 5 のいずれにおいても共通している。一方で、下表において、複数のテキストに共通する文字列「日本 IBM は PureSystems を新時代の IT 製品として発表した。」の前後の文字は、テキスト 1、2、4 及び 5 のいずれにおいても共通していない。この場合、判断部 1 2 4 は、文字列「本 IBM」を引用部分として検出せず、文字列「日本 IBM は PureSystems を新時代の IT 製品として発表した。」を引用部分として検出する。

30

【0073】

これにより、判断部 1 2 4 は、複数のテキストが共通して含む同一の文字列のうち最長の文字列を引用部分として検出し、最長の文字列より短い文字列を引用部分として検出しない。従って、判断部 1 2 4 は、最長の文字列と実質的に同じ文字列の処理を省略して情報処理装置 1 0 の処理リソースを節約することができる。このように、判断部 1 2 4 は、複数のテキストから引用部分である文字列を検出し、検出した文字列から構成される引用文字列テーブルを生成する。

40

【0074】

図 7 は、S 1 0 4 において判断部 1 2 4 が生成する引用文字列テーブルを例示する。図示するように、例えば、判断部 1 2 4 は、引用部分として文字列 1 「日本 IBM は PureSystems を新時代の IT 製品として発表した。」、文字列 2 「日本 IBM は PureSystems を新時代の IT 製品として発表」、文字列 3 「PureSystems を新時代の IT 製品として発表した。」、文字列 4 「[日光ニュース] A 社が新型スマートフォンを発表。」、文字列 5 「新社長就任のお知らせ」、及び、文字列 6 「娘が全国大会で優勝しました！」を含む引用文字列テーブルを生成する。ここで、文字列 1 は、文字列 2 及び文字列 3 を包含するが、判断部 1 2 4 はこれらを別の引用部分として区別して検出している。

50

【0075】

図8は、S108及びS110において照合部126が生成する照合テーブルを例示する。照合部126は、S108において、共通部分を含む文字列1、文字列2、及び文字列3に係る引用部分を同一の情報からの引用であると判断して、引用文字列テーブルに同一の識別情報「NEWS_TITLE1」を付与した照合テーブルを生成する。

【0076】

また、照合部126は、S110において、最終的な参照先情報「http://www. XXXXXXitnews.co.jp/news1111」の参照先に文字列1～3が含まれることに応じて、照合テーブルにおいて文字列1～3と、最終的な参照先情報及び最終的な参照先情報に対応する他の参照先情報とを対応付ける。このように、照合部126は、識別情報、引用部分、最終的な参照先情報、及び、他の参照先情報に対応付けた照合テーブルを生成する。

10

【0077】

図9は、S112において変換部140が生成する複数の変換済テキストを例示する。図示されるように、文字列変換部144は、複数のテキスト1～5の文字列「日本IBMはPureSystemsを新時代のIT製品として発表した。」を識別情報「NEWS_TITLE1」に変換し、参照先変換部142は参照先情報を識別情報「NEWS_TITLE1」に置換する。

【0078】

ここで、テキスト2、4及び5は、文字列「日本IBMはPureSystemsを新時代のIT製品として発表した。」と参照先情報の両方を有するので、変換部140は、当該文字列及び参照先情報のうち一方を置換せずに削除する。また、文字列変換部144は、宛先を示す「@Hogehoge」を「To_User」に置換し、タグ「#IBM_News」を削除する。

20

【0079】

テキストマイニング部150は、図9に示す変換済テキストをテキストマイニングすることで、例えば「NEWS_TITLE1」が特定のグループのテキスト内で、特定の期間に、何回出現したかを集計することができる。これにより、本実施形態の情報処理装置10は、引用の内容ごとに引用の頻度等を分析することができる。

【0080】

図10は、情報処理装置10として機能するコンピュータ1900のハードウェア構成の一例を示す。本実施形態に係るコンピュータ1900は、ホスト・コントローラ2082により相互に接続されるCPU2000、RAM2020、グラフィック・コントローラ2075、及び表示装置2080を有するCPU周辺部と、入出力コントローラ2084によりホスト・コントローラ2082に接続される通信インターフェイス2030、ハードディスクドライブ2040、及びCD-ROMドライブ2060を有する入出力部と、入出力コントローラ2084に接続されるROM2010、フレキシブルディスク・ドライブ2050、及び入出力チップ2070を有するレガシー入出力部を備える。

30

【0081】

ホスト・コントローラ2082は、RAM2020と、高い転送レートでRAM2020をアクセスするCPU2000及びグラフィック・コントローラ2075とを接続する。CPU2000は、ROM2010及びRAM2020に格納されたプログラムに基づいて動作し、各部の制御を行う。グラフィック・コントローラ2075は、CPU2000等がRAM2020内に設けたフレーム・バッファ上に生成する画像データを取得し、表示装置2080上に表示させる。これに代えて、グラフィック・コントローラ2075は、CPU2000等が生成する画像データを格納するフレーム・バッファを、内部に含んでもよい。

40

【0082】

入出力コントローラ2084は、ホスト・コントローラ2082と、比較的高速な入出力装置である通信インターフェイス2030、ハードディスクドライブ2040、CD-ROMドライブ2060を接続する。通信インターフェイス2030は、有線又は無線によりネットワークを介して他の装置と通信する。また、通信インターフェイスは、通信部110における通信を行うハードウェアとして機能する。ハードディスクドライブ204

50

0は、コンピュータ1900内のCPU2000が使用するプログラム及びデータを格納する。CD-ROMドライブ2060は、CD-ROM2095からプログラム又はデータを読み取り、RAM2020を介してハードディスクドライブ2040に提供する。

【0083】

また、入出力コントローラ2084には、ROM2010と、フレキシブルディスク・ドライブ2050、及び入出力チップ2070の比較的低速な入出力装置とが接続される。ROM2010は、コンピュータ1900が起動時に実行するブート・プログラム、及び／又は、コンピュータ1900のハードウェアに依存するプログラム等を格納する。フレキシブルディスク・ドライブ2050は、フレキシブルディスク2090からプログラム又はデータを読み取り、RAM2020を介してハードディスクドライブ2040に提供10する。入出力チップ2070は、フレキシブルディスク・ドライブ2050を入出力コントローラ2084へと接続するとともに、例えばパラレル・ポート、シリアル・ポート、キーボード・ポート、マウス・ポート等を介して各種の入出力装置を入出力コントローラ2084へと接続する。

【0084】

RAM2020を介してハードディスクドライブ2040に提供されるプログラムは、フレキシブルディスク2090、CD-ROM2095、又はICカード等の記録媒体に格納されて利用者によって提供される。プログラムは、記録媒体から読み出され、RAM2020を介してコンピュータ1900内のハードディスクドライブ2040にインストールされ、CPU2000において実行される。20

【0085】

コンピュータ1900にインストールされ、コンピュータ1900を情報処理装置10として機能させるプログラムは、通信モジュールと、検出モジュールと、参照先検出モジュールと、判断モジュールと、照合モジュールと、変換モジュールと、参照先変換モジュールと、文字列変換モジュールと、テキストマイニングモジュールとを備える。これらのプログラム又はモジュールは、CPU2000等に働きかけて、コンピュータ1900を、通信部110と、検出部120と、参照先検出部122と、判断部124と、照合部126と、変換部140と、参照先変換部142と、文字列変換部144と、テキストマイニング部150としてそれぞれ機能させてよい。

【0086】

これらのプログラムに記述された情報処理は、コンピュータ1900に読込まれることにより、ソフトウェアと上述した各種のハードウェア資源とが協働した具体的手段である通信部110と、検出部120と、参照先検出部122と、判断部124と、照合部126と、変換部140と、参照先変換部142と、文字列変換部144と、テキストマイニング部150として機能する。そして、これらの具体的手段によって、本実施形態におけるコンピュータ1900の使用目的に応じた情報の演算又は加工を実現することにより、使用目的に応じた特有の情報処理装置10が構築される。30

【0087】

一例として、コンピュータ1900と外部の装置等との間で通信を行う場合には、CPU2000は、RAM2020上にロードされた通信プログラムを実行し、通信プログラ40ムに記述された処理内容に基づいて、通信インターフェイス2030に対して通信処理を指示する。通信インターフェイス2030は、CPU2000の制御を受けて、RAM2020、ハードディスクドライブ2040、フレキシブルディスク2090、又はCD-ROM2095等の記憶装置上に設けた送信バッファ領域等に記憶された送信データを読み出してネットワークへと送信し、もしくは、ネットワークから受信した受信データを記憶装置上に設けた受信バッファ領域等へと書き込む。このように、通信インターフェイス2030は、DMA（ダイレクト・メモリ・アクセス）方式により記憶装置との間で送受信データを転送してもよく、これに代えて、CPU2000が転送元の記憶装置又は通信インターフェイス2030からデータを読み出し、転送先の通信インターフェイス2030又は記憶装置へとデータを書き込むことにより送受信データを転送してもよい。50

【0088】

また、CPU2000は、ハードディスクドライブ2040、CD-ROMドライブ2060（CD-ROM2095）、フレキシブルディスク・ドライブ2050（フレキシブルディスク2090）等の外部記憶装置に格納されたファイルまたはデータベース等の中から、全部または必要な部分をDMA転送等によりRAM2020へと読み込ませ、RAM2020上のデータに対して各種の処理を行う。そして、CPU2000は、処理を終えたデータを、DMA転送等により外部記憶装置へと書き戻す。このような処理において、RAM2020は、外部記憶装置の内容を一時的に保持するものとみなせるから、本実施形態においてはRAM2020及び外部記憶装置等をメモリ、記憶部、または記憶装置等と総称し、記憶部130として機能させる。本実施形態における各種のプログラム、データ、テーブル、データベース等の各種の情報は、このような記憶装置上に格納されて、情報処理の対象となる。なお、CPU2000は、RAM2020の一部をキャッシュメモリに保持し、キャッシュメモリ上で読み書きを行うこともできる。このような形態においても、キャッシュメモリはRAM2020の機能の一部を担うから、本実施形態においては、区別して示す場合を除き、キャッシュメモリもRAM2020、メモリ、及び／又は記憶装置に含まれるものとする。

10

【0089】

また、CPU2000は、RAM2020から読み出したデータに対して、プログラムの命令列により指定された、本実施形態中に記載した各種の演算、情報の加工、条件判断、情報の検索・置換等を含む各種の処理を行い、RAM2020へと書き戻す。例えば、CPU2000は、条件判断を行う場合においては、本実施形態において示した各種の変数が、他の変数または定数と比較して、大きい、小さい、以上、以下、等しい等の条件を満たすか否かを判断し、条件が成立した場合（又は不成立であった場合）に、異なる命令列へと分岐し、またはサブルーチンを呼び出す。

20

【0090】

また、CPU2000は、記憶装置内のファイルまたはデータベース等に格納された情報を検索することができる。例えば、第1属性の属性値に対し第2属性の属性値がそれぞれ対応付けられた複数のエントリが記憶装置に格納されている場合において、CPU2000は、記憶装置に格納されている複数のエントリの中から第1属性の属性値が指定された条件と一致するエントリを検索し、そのエントリに格納されている第2属性の属性値を読み出すことにより、所定の条件を満たす第1属性に対応付けられた第2属性の属性値を得ることができる。

30

【0091】

以上に示したプログラム又はモジュールは、外部の記録媒体に格納されてもよい。記録媒体としては、フレキシブルディスク2090、CD-ROM2095の他に、DVD又はCD等の光学記録媒体、MO等の光磁気記録媒体、テープ媒体、ICカード等の半導体メモリ等を用いることができる。また、専用通信ネットワーク又はインターネットに接続されたサーバシステムに設けたハードディスク又はRAM等の記憶装置を記録媒体として使用し、ネットワークを介してプログラムをコンピュータ1900に提供してもよい。

【0092】

以上、本発明を実施の形態を用いて説明したが、本発明の技術的範囲は上記実施の形態に記載の範囲には限定されない。上記実施の形態に、多様な変更または改良を加えることが可能であることが当業者に明らかである。そのような変更または改良を加えた形態も本発明の技術的範囲に含まれ得ることが、特許請求の範囲の記載から明らかである。

40

【0093】

特許請求の範囲、明細書、および図面中において示した装置、システム、プログラム、および方法における動作、手順、ステップ、および段階等の各処理の実行順序は、特段「より前に」、「先立って」等と明示しておらず、また、前の処理の出力を後の処理で用いるのでない限り、任意の順序で実現しうることに留意すべきである。特許請求の範囲、明細書、および図面中の動作フローに関して、便宜上「まず、」、「次に、」等を用いて説

50

明したとしても、この順で実施することが必須であることを意味するものではない。

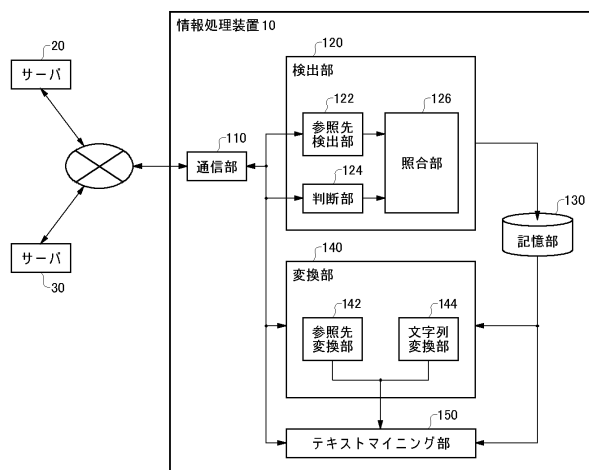
【符号の説明】

【0094】

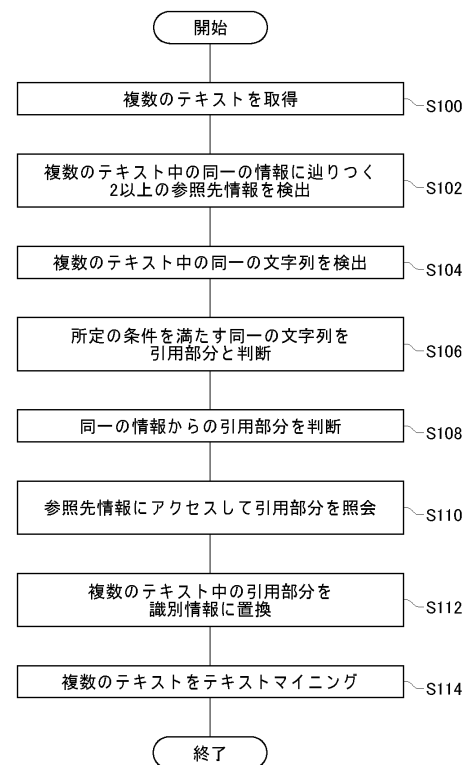
10 情報処理装置、20 サーバ、30 サーバ、110 通信部、120 検出部、
122 参照先検出部、124 判断部、126 照合部、130 記憶部、140 変換部、
142 参照先変換部、144 文字列変換部、150 テキストマイニング部、
1900 コンピュータ、2000 CPU、2010 ROM、2020 RAM、2030 通信インターフェイス、
2040 ハードディスクドライブ、2050 フレキシブルディスク・ドライブ、2060 CD-ROMドライブ、
2070 入出力チップ、2075 グラフィック・コントローラ、2080 表示装置、2082 ホスト・コ
ントローラ、2084 入出力コントローラ、2090 フレキシブルディスク、209
5 CD-ROM

10

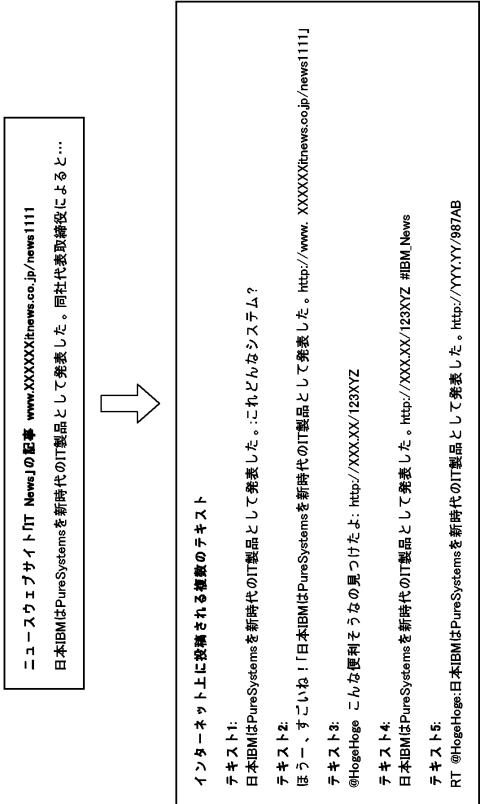
【図1】



【図2】



【図 3】



【図 4】

最終的な参照先情報	他の参照先情報
http://www.XXXXXXitnews.co.jp/news1111	http://XXX.XX/123XYZ http://YYY.YY/987AB
http://nikko.com/news/20120xxxx.html	http://XXX.XX/356CDE
http://www.ibm.com/press/2012/0x/0x/xx	http://XXX.XX/456DEF
http://www.blog.com/hogehoge/2012xxxx	http://XXX.XX/567EFG
...	...

【図 5】

テキスト1:
日本IBMはPureSystemsを新時代のIT製品として発表した。:これどんなシステム?

テキスト1のNグラム索引	前の文字	索引	後の文字
1文字目の1グラム索引	(無し)	日	本
2文字目の1グラム索引	日	本	I
:	:	:	:
2文字目の4グラム索引	日	本IBM	は
:	:	:	:
1文字目の34グラム索引	(無し)	日本IBMはPureSystemsを新時代のIT製品として発表した。	:
:	:	:	:

【図 6】

複数テキストの2グラム索引の一例	前の文字	索引	後の文字
テキスト1の2文字目の4グラム	日	本IBM	は
テキスト2の12文字目の4グラム	日	本IBM	は
テキスト4の2文字目の4グラム	日	本IBM	は
テキスト5の15文字目の4グラム	日	本IBM	は

複数テキストの34グラム索引の一例	前の文字	索引	後の文字
テキスト1の1文字目の34グラム	(無し)	日本IBMはPureSystemsを新時代のIT製品として発表した。	:
テキスト2の11文字目の34グラム	「	日本IBMはPureSystemsを新時代のIT製品として発表した。	h
テキスト4の1文字目の34グラム	(無し)	日本IBMはPureSystemsを新時代のIT製品として発表した。	h
テキスト5の14文字目の34グラム	:	日本IBMはPureSystemsを新時代のIT製品として発表した。	h

【図 7】

No.	引用部分
1	日本IBMはPureSystemsを新時代のIT製品として発表した。
2	日本IBMはPureSystemsを新時代のIT製品として発表
3	PureSystemsを新時代のIT製品として発表した。
4	【日光ニュース】A社が新型スマートフォンを発表。
5	新社長就任のお知らせ
6	娘が全国大会で優勝しました！
⋮	⋮

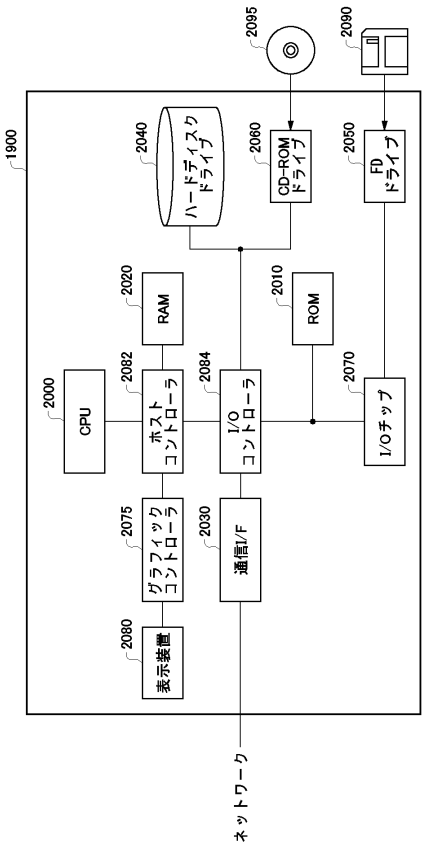
【図 8】

No.	識別情報	引用部分	最終的な参照先情報	他の参照先情報
1	NEWS_TITLE1	日本IBMはPureSystemsを新時代のIT製品として発表した。	http://www.XXXXXXXtnews.co.jp/news/111	https://XXX.XX/123XYZ https://YYY.YY/987AB
2	NEWS_TITLE1	日本IBMはPureSystemsを新時代のIT製品として発表	http://www.XXXXXXXtnews.co.jp/news/111	https://XXX.XX/123XYZ http://YYY.YY/987AB
3	NEWS_TITLE1	PureSystemsを新時代のIT製品として発表した。	http://www.XXXXXXXtnews.co.jp/news/111	https://XXX.XX/123XYZ http://YYY.YY/987AB
4	NIKKO_NEWS.0001	【日光ニュース】A社が新型スマートフォンを発表。	http://nikko.com/news/20120xxxx.html	http://XXX.XX/356ODE
5	IBM_PRESS.0034	新社長就任のお知らせ	http://www.ibm.com/press/2012/0x/0x/xx	http://XXX.XX/456DEF
6	BLOG.0003	娘が全国大会で優勝しました！		
7	URL.0004		http://www.blog.com/hogehoge/2012xxxx	http://XXX.XX/567EFG
⋮	⋮	⋮	⋮	⋮

【図 9】

複数の変換済テキスト
テキスト1: NEW_TITLE1:これどんなシステム？
テキスト2: ほうー、すごいね！「NEW_TITLE1」
テキスト3: To_User こんな便利そうなの見つけたよ: NEW_TITLE1
テキスト4: NEW_TITLE1
テキスト5: RT To_User: NEW_TITLE1

【図 10】



フロントページの続き

(74)代理人 100112690

弁理士 太佐 種一

(72)発明者 米持 幸寿

東京都中央区日本橋箱崎町19番地21 日本アイ・ビー・エム株式会社内

審査官 川▲崎▼ 博章

(56)参考文献 特開2007-087110 (JP, A)

特開2010-231508 (JP, A)

(58)調査した分野(Int.Cl., DB名)

G06F 17/20 — 17/28

G06F 17/30