

(11)特許出願公開番号

特開2010-67243

(P2010-67243A)

(43) 公開日 平成22年3月25日(2010.3.25)

(51) Int.Cl.

F I

テーマコード (参考)

G06F 17/30 (2006.01)

G06F 17/30 220Z

5 B 0 7 5

G06F 17/21 (2006.01)

G06F 17/30 170A

5 B 1 0 9

G06F 13/00 (2006.01)

G06F 17/21 570R

G06F 19/00 (2006.01)

G06F 13/00 540A

G06F 19/00 130

審査請求 有 請求項の数 7 書面 (全 14 頁)

(21) 出願番号 特願2008-265751 (P2008-265751)

(22) 出願日 平成20年9月12日 (2008. 9. 12)

(71) 出願人 500257300

ヤフー株式会社

東京都港区赤坂9丁目7番1号

(72) 発明者 小林 竜己

東京都港区六本木六丁目10番1号 ヤフー株式会社内

Fターム(参考) 5B075 ND03 NR12 NS10 UU24

5B109 QA04 SA14

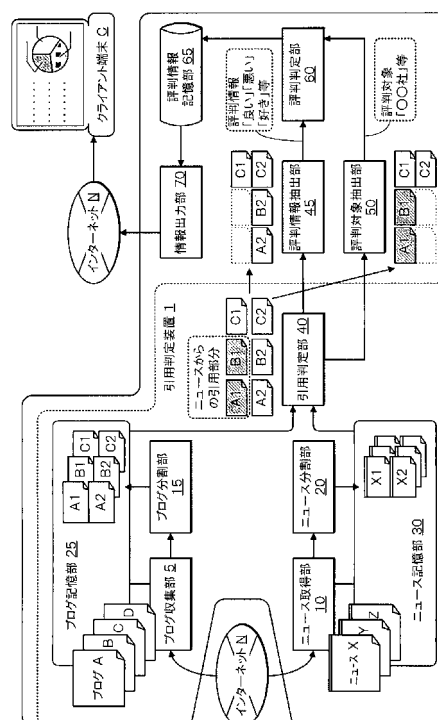
(54) 【発明の名称】 引用判定方法及びそれを用いた評判抽出方法

(57) 【要約】

【課題】教師データ作成や事前学習の負担や弊害無しで、ブログなどユーザ記事との類似性を迅速・高精度に判定可能にする。

【解決手段】ブログ収集部５が、ユーザ記事をウェブ上から、インターネットに代表される通信ネットワークＮ經由でデータとして取得し、ニュース取得部１０が、ニュースを同様にウェブ上からデータとして取得し、それぞれブログ記憶部２５とニュース記憶部３０に記憶させる。引用判定部４０が、クラスタリングなどの機械学習により、ブログ記憶部２５に記憶されている各ブログについて、ニュース記憶部３０に記憶されている各ニュースとの類似度を判断することにより、いずれかのニュースを引用しているか否かを判定する。

【選択図】図 1



【特許請求の範囲】**【請求項 1】**

記事間の引用をコンピュータで判定する引用判定装置であって、

引用する側として引用判定の対象とするユーザ記事をウェブ上からデータとして取得するユーザ記事収集手段と、

引用される側として引用判定の対象とする公表記事をウェブ上からデータとして取得する公表記事取得手段と、

前記ユーザ記事収集手段で収集した前記各ユーザ記事を記憶するユーザ記事記憶手段と、

前記公表記事取得手段で取得した前記各公表記事を記憶する公表記事記憶手段と、

10

前記ユーザ記事記憶手段に記憶されている前記各ユーザ記事の集合に、前記公表記事記憶手段に記憶されている前記各公表記事を教師データとして加えた全体集合にクラスタリングを行うことにより、前記ユーザ記事がいずれかの前記公表記事を引用しているか否かを判定する、引用判定手段と、

を前記コンピュータの演算制御部で実現することを特徴とする引用判定装置。

【請求項 2】

前記ユーザ記事収集手段で収集した前記各ユーザ記事及び前記公表記事取得手段で取得した前記各公表記事を、それぞれ段落ごとに分割する段落分割手段、を前記演算制御部で実現すると共に、

前記引用判定手段は、前記段落分割手段で分割された前記各ユーザ記事の各段落及び前記各公表記事の各段落を対象として、前記引用の判定を行うように構成した

20

ことを特徴とする請求項 1 記載の引用判定装置。

【請求項 3】

請求項 1 又は 2 記載の引用判定装置の構成に加え、

記事から評判対象を抽出する評判対象抽出手段と、

記事から前記評判対象に対する評判を表す評判情報を抽出する評判情報抽出手段と、

前記評判対象抽出手段で抽出した評判対象と、この評判対象に対する評判として前記評判情報抽出手段で抽出した評判情報と、に基いて評判を判定する評判判定手段と、

を前記演算制御部で実現すると共に、

前記引用判定手段で前記公表記事が引用されていないと判定したユーザ記事について、評判対象抽出手段で評判対象を抽出し、評判情報抽出手段で前記評判対象に対する評判を表す評判情報を抽出する

30

ことを特徴とする評判判定装置。

【請求項 4】

ユーザ記事を分割した一部の段落が前記公表記事からの引用部分であると前記引用判定手段が判定した場合に、前記評判対象抽出手段は、その引用部分から評判対象を抽出する一方、前記評判情報抽出手段は、同じユーザ記事のうち前記引用部分を除く段落から評判情報を抽出するように構成したことを特徴とする請求項 3 記載の評判判定装置。

【請求項 5】

前記公表記事ごとに、一部の段落がその公表記事からの引用部分であると前記引用判定手段で判定した前記各ユーザ記事と、各ユーザ記事から前記評判情報抽出手段で抽出した前記評判情報と、を記憶する評判情報記憶手段と、

40

前記コンピュータに設けた画面表示装置もしくは通信ネットワーク経由で接続される他のコンピュータに、前記評判情報を出力する情報出力手段と、

を前記コンピュータの前記演算制御部で実現し、

前記情報出力手段は、前記評判情報記憶手段に記憶されている各情報に基づいて、前記公表記事ごとに、その公表記事からの引用部分を持つ各ユーザ記事と、前記評判情報と、を表示する

ことを特徴とする請求項 4 記載の評判判定装置。

【請求項 6】

50

記事間の引用をコンピュータで判定する引用判定方法であって、

前記コンピュータの演算制御部により、ユーザ記事収集手段と、公表記事取得手段と、ユーザ記事記憶手段と、公表記事記憶手段と、引用判定手段と、を実現するとともに、

前記ユーザ収集手段により、引用する側として引用判定の対象とするユーザ記事をウェブ上からデータとして取得する、ユーザ記事収集処理ステップを実行し、

前記公表記事取得手段により、引用される側として引用判定の対象とする公表記事をウェブ上からデータとして取得する、公表記事取得処理ステップを実行し、

前記ユーザ記事記憶手段に、前記ユーザ記事収集手段で収集した前記各ユーザ記事を記憶し、

前記公表記事記憶手段に、前記公表記事取得手段で取得した前記各公表記事を記憶し、

前記引用判定手段により、前記ユーザ記事記憶手段に記憶されている前記各ユーザ記事について、前記公表記事記憶手段に記憶されている前記各公表記事との類似度を判断することにより、いずれかの前記公表記事を引用しているか否かを判定する、引用判定処理ステップを実行する

ことを特徴とする引用判定方法。

【請求項 7】

コンピュータを制御することにより、記事間の引用を判定させる引用判定プログラムであって、

そのプログラムは、

前記コンピュータの前記演算制御部を制御することにより、ユーザ記事収集手段と、公表記事取得手段と、ユーザ記事記憶手段と、公表記事記憶手段と、引用判定手段と、を実現させるとともに、

前記ユーザ収集手段により、引用する側として引用判定の対象とするユーザ記事をウェブ上からデータとして取得する、ユーザ記事収集処理ステップを実行させ、

前記公表記事取得手段により、引用される側として引用判定の対象とする公表記事をウェブ上からデータとして取得する、公表記事取得処理ステップを実行させ、

前記ユーザ記事収集手段で収集した前記各ユーザ記事を、前記ユーザ記事記憶手段に記憶させ、

前記公表記事取得手段で取得した前記各公表記事を、前記公表記事記憶手段に記憶させ

、
前記引用判定手段により、前記ユーザ記事記憶手段に記憶されている前記各ユーザ記事について、前記公表記事記憶手段に記憶されている前記各公表記事との類似度を判断することにより、いずれかの前記公表記事を引用しているか否かを判定する、引用判定処理ステップを実行させる

ことを特徴とする引用判定プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、ブログなどの記事について、ニュースなどの引用の有無を判定する技術の改良に関する。

【背景技術】

【0002】

近年、インターネットと情報通信機器の普及成熟に伴い、企業など大規模組織からの情報発信にとどまらず、個々人のインターネットユーザから情報を発信する手段も急速に普及した。その一例は、ブログ、SNS（ソーシャル・ネットワーキング・サービス）、電子掲示板などで、新形態も次々登場しているが、これらはUGC（User Generated Content）と総称される。

【0003】

これらUGCの代表例はブログ（ウェブログ）であり、その典型的な形態は、サービス提供会社のウェブサーバ上に、個人が好みのテーマやタイトル、デザインでブログのウェ

10

20

30

40

50

ブサイトを開設し、気の向いたテーマで日記風の記事を記述する。記事に対しては、ブログの設定に応じ、面識ある知人やその他の閲覧者が「コメント」などと呼ばれる応答記事を投稿できる。

【0004】

なお、このようなUGCにおける記事、コメント、トラックバック、レス（レスポンス）などの投稿単位を本出願では「ユーザ記事」と総称する。他方、報道機関などがニュースサイトで配信するようなニュース記事や、教育機関や学会などがウェブサイトに掲載する学術論文などの記事は、本出願では「公表記事」と総称して、前記ユーザ記事と区別することとする。

【0005】

10

そして、上記のようなユーザ記事に企業や商品の好悪など評判が記述された場合の影響増大に伴って、情報の収集や管理などのため、ブログなどのユーザ記事から、製品やサービスの名称、企業名などと共に、それらに対する「よい」「だめ」などの評判情報を抽出して、ポジティブ/ネガティブを判定する技術も登場し、評判判定と呼ばれている。

【0006】

このような評判判定の技術の一例として、特許文献1では、ネットワーク上の文書から対象物と評価表現からなる評価対を抽出して、機械学習により評価対の規則を学習する例や、また、トレーニング用のモデル文書から抽出した評価対を機械学習器に与えて評価対となりうる規則を学習し、未知の文書から評価対を抽出する例を示している。

【特許文献1】特開2007-219880号公報

20

【発明の開示】

【発明が解決しようとする課題】

【0007】

しかし、ブログ等のユーザ記事は、必ずしも全文をユーザが独自に記述したものとは限らず、例えば、ニュース等を引用してコメントを記入するといったものもある。このようなユーザ記事に対して上記のような従来の評判抽出の技術をそのまま適用すると、ニュース等で記者等が記述している評判情報も抽出してしまうため、評判判定の精度が低下するという課題があった。

【0008】

このため、ユーザ記事ごとに、ニュースなどが引用されているか否かの判定を行う必要があり、このような引用の有無を判定する基本的な手段としては、ニュース1件1件と、ブログなど全てのユーザ記事とを、文字列として比較照合することも考えられるが、この場合、その組合せ数の多さから処理量が膨大になるうえ、一部引用を考慮して一部一致の判定まで加えると、処理負荷はさらに飛躍的に膨張し、適用が現実的とはいえない。

30

【0009】

また、SVM（サポート・ベクター・マシン）などの機械学習を用い、ニュース等の特徴を表す教師データを作成して事前学習をさせたうえ、その教師データへの類似性を判定させる手法もありうるが、教師データの作成は非常に手間であり、事前のオフライントレーニングを要するため迅速性やリアルタイム性にも劣り、さらに、教師データに過剰適合して本番データの判定精度が落ちる等の弊害や、ネット上に流れる様々なニュース形態や記載スタイルの今後の変遷に対応するのは困難といった問題もあった。

40

【0010】

本発明は、上記のような従来技術の課題を解決するもので、その目的は、教師データ作成や事前学習の負担や弊害無しで、ブログなどユーザ記事との類似性を迅速・高精度に判定可能にすることである。

【課題を解決するための手段】

【0011】

上記の目的をふまえ、本発明は以下のような各態様を含むものであり、また、装置に加え、同様な方法及びコンピュータ・プログラムについても、以下の各態様に準ずるものである。

50

【 0 0 1 2 】

(1) 本発明の一態様は、記事間の引用をコンピュータで判定する引用判定装置であって、引用する側として引用判定の対象とするユーザ記事をウェブ上からデータとして取得するユーザ記事収集手段と、引用される側として引用判定の対象とする公表記事をウェブ上からデータとして取得する公表記事取得手段と、前記ユーザ記事収集手段で収集した前記各ユーザ記事を記憶するユーザ記事記憶手段と、前記公表記事取得手段で取得した前記各公表記事を記憶する公表記事記憶手段と、前記ユーザ記事記憶手段に記憶されている前記各ユーザ記事の集合に、前記公表記事記憶手段に記憶されている前記各公表記事を教師データとして加えた全体集合にクラスタリングを行うことにより、前記ユーザ記事がいずれかの前記公表記事を引用しているか否かを判定する、引用判定手段と、を前記コンピュータの演算制御部で実現することを特徴とする。

10

【 0 0 1 3 】

このように、本番の引用判定において、判定対象のユーザ記事群についてニュースなどの公表記事自体との類似度を判断することにより、教師データや事前学習無しでユーザ記事とニュース等の引用有無を迅速かつ高精度に判定可能となる。しかも、実際のニュース等の本番データで教師データを兼ねることにより、人為的に教師データを作成する負担も無く、教師データに過剰適合し本番データの判定精度が落ちる等の弊害も避けながら、ネット上に登場する様々な記載スタイルのニュース形式にも柔軟に対応可能となる。

また、引用有無判定の対象とするユーザ記事群に、引用される側すなわちターゲットクラスである公表記事を加えた上で全体をクラスタリングし、公表記事を含むクラスタを取得することにより、公表記事への類似度すなわち引用可能性の高い高純度なユーザ記事を抽出できるうえ、SVMなどオフライントレーニングタイプの機械学習よりも、新規の様々なニュースへの対応が迅速容易になる。

20

【 0 0 1 4 】

(2) 本発明の他の態様は、上記いずれかの態様において、前記ユーザ記事収集手段で収集した前記各ユーザ記事及び前記公表記事取得手段で取得した前記各公表記事を、それぞれ段落ごとに分割する段落分割手段、を前記演算制御部で実現すると共に、前記引用判定手段は、前記段落分割手段で分割された前記各ユーザ記事の各段落及び前記各公表記事の各段落を対象として、前記引用の判定を行うように構成したことを特徴とする。

【 0 0 1 5 】

30

このように、ユーザ記事と公表記事の双方を段落単位に分割のうえ類似度判断による引用判定を行うことにより、ニュース等の一部引用についても高精度に判定可能となる。

【 0 0 1 6 】

(3) 本発明の他の態様は、上記いずれかの引用判定装置の構成に加え、記事から評判対象を抽出する評判対象抽出手段と、記事から前記評判対象に対する評判を表す評判情報を抽出する評判情報抽出手段と、前記評判対象抽出手段で抽出した評判対象と、この評判対象に対する評判として前記評判情報抽出手段で抽出した評判情報と、に基いて評判を判定する評判判定手段と、を前記演算制御部で実現すると共に、前記引用判定手段で前記公表記事が引用されていないと判定したユーザ記事について、評判対象抽出手段で評判対象を抽出し、評判情報抽出手段で前記評判対象に対する評判を表す評判情報を抽出することを

40

【 0 0 1 7 】

このように、本発明の引用判定で引用でないと判定したユーザ記事について、評判対象と評判情報の抽出対象とすることにより、ニュースなどに含まれる好悪等の評判情報を除いて評判抽出を行うこととなり、評判判定の精度を効果的に向上させることが可能となる。

【 0 0 1 8 】

(4) 本発明の他の態様は、上記態様の評判判定装置において、ユーザ記事を分割した一部の段落が前記公表記事からの引用部分であると前記引用判定手段が判定した場合に、前記評判対象抽出手段は、その引用部分から評判対象を抽出する一方、前記評判情報抽出手

50

段は、同じユーザ記事のうち前記引用部分を除く段落から評判情報を抽出するように構成したことを特徴とする。

【0019】

このように、一部の段落が公表記事の引用と判定した場合に、引用されているニュース等から主題となる企業名、製品名、人名等の評判対象を抽出することで評判対象が明確となることに加え、引用以外の段落からはその対象についての評判情報を抽出するので、ニュース等を引用しているユーザ記事も評判判定に有効活用可能となる。

【0020】

(5) 本発明の他の態様は、上記態様の評判判定装置において、前記公表記事ごとに、一部の段落がその公表記事からの引用部分であると前記引用判定手段で判定した前記各ユーザ記事と、各ユーザ記事から前記評判情報抽出手段で抽出した前記評判情報と、を記憶する評判情報記憶手段と、前記コンピュータに設けた画面表示装置もしくは通信ネットワーク経由で接続される他のコンピュータに、前記評判情報を出力する情報出力手段と、を前記コンピュータの前記演算制御部で実現し、前記情報出力手段は、前記評判情報記憶手段に記憶されている各情報に基いて、前記公表記事ごとに、その公表記事からの引用部分を持つ各ユーザ記事と、前記評判情報と、を表示することを特徴とする。

10

【0021】

このように、個々の公表記事に対して、その公表記事を引用した各ユーザ記事と、各ユーザ記事から抽出した評判情報と、を一画面で表示するユーザインタフェースにより、各公表記事に対するユーザの記述や評判が一見把握容易になり、広報等の情報政策への有効活用が容易になる。

20

【発明の効果】

【0022】

以上のように、本発明によれば、教師データ作成や事前学習の負担や弊害無しで、ブログなどユーザ記事との類似性を迅速・高精度に判定可能となる。

【発明を実施するための最良の形態】

【0023】

次に、本発明を実施するための最良の形態（以下「本実施形態」と呼ぶ）について、図に沿って説明する。なお、背景技術や課題などで既に述べた内容と共通の前提事項については適宜省略する。

30

【0024】

本実施形態は、図1の構成図に示すように、記事間の引用をコンピュータで判定する引用判定装置1を含む評判判定装置（以下「本装置」と呼ぶ）2に関するものであるが、本発明は、引用判定装置1や評判判定装置2に対応する情報処理の方法及びコンピュータ・プログラムとしても把握可能である。

【0025】

なお、本実施形態では、引用する側として引用判定の対象とするユーザ記事としてブログを例にとり、引用される側として引用判定の対象とする公表記事としてニュースを例にとるものとする。なお、ここで「ブログ」「ニュース」は、それぞれウェブサイトではなくウェブサイトにおける個別の記事を意味するものとする。

40

【0026】

〔構成〕

まず、本装置は、ユーザ記事収集手段としてのブログ収集部5と、公表記事取得手段としてのニュース取得部10と、段落分割手段としてのブログ分割部15並びにニュース分割部20と、ユーザ記事記憶手段としてのブログ記憶部25と、公表記事記憶手段としてのニュース記憶部30と、引用判定手段としての引用判定部40と、評判情報抽出手段としての評判情報抽出部45と、評判対象抽出手段としての評判対象抽出部50と、評判判定手段としての評判判定部60と、評判情報記憶手段としての評判情報記憶部65と、情報出力手段としての情報出力部70と、を有する。

【0027】

50

これら各手段としての各部は、コンピュータの図示しない演算制御部（CPUなど）を所定のコンピュータ・プログラム（引用判定プログラムや評判判定プログラム）で制御することにより実現するもので、本発明や本実施形態を構成する以下のような各機能や処理ステップを実現・実行する処理手段である。

【0028】

なお、本装置を実現している前記コンピュータは、HDDや主メモリ等の記憶装置、マウスや液晶表示装置などの入出力装置、ネットワークとの通信回路などを有し、前記演算制御部はこれらハードウェアを利用して前記各部や各手段を実現するが、これらハードウェア事態は一般的であるため、図示や詳説は省略する。

【0029】

〔基本的な作用効果〕

上記のように構成した本装置における基本的な作用効果としては、まず、ブログ収集部5が、ユーザ記事であるブログA, B, C...をウェブ(WWW: World Wide Web)上から、インターネットに代表される通信ネットワークN経由でデータとして取得し(ユーザ記事収集処理ステップ)、また、ニュース取得部10が、ニュースX, Y...を同様にウェブ上からデータとして取得し(公表記事取得処理ステップ)、それぞれブログ記憶部25とニュース記憶部30に記憶させる。

【0030】

そして、引用判定部40が、クラスタリングにより、ブログ記憶部25に記憶されている各ブログについて、ニュース記憶部30に記憶されている各ニュースとの類似度を判断することにより、いずれかのニュースを引用しているか否かを判定する(引用判定処理ステップ)。

【0031】

このように、本番の引用判定において、判定対象のブログ群についてニュース自体との類似度を判断することにより、教師データや事前学習無しでブログとニュースの引用有無を迅速かつ高精度に判定可能となる。

【0032】

しかも、実際のニュースの本番データで教師データを兼ねることにより、人為的に教師データを作成する負担も無く、教師データに過剰適合し本番データの判定精度が落ちる等の弊害も避けながら、ネット上に登場する様々な記載スタイルのニュース形式にも柔軟に対応可能となる。

【0033】

〔引用判定の手法〕

また、引用判定には広義の機械学習を用いることも考えられるが、潜在的意味インデクシング(LSI: Latent Semantic Indexing)を用いたクラスタリングが最も望ましい。ここで、機械学習は、教師あり機械学習、教師なし機械学習、強化学習に分けることができ、教師あり機械学習の例は、バックプロパゲーション、サポートベクターマシン、ID3、単純ベイズ分類器、事例ベース推論、ブースティングなどがある。

【0034】

教師あり学習の中でも、特に代表的なアルゴリズムであるサポートベクターマシン(SVM: Support Vector Machine)は、ソフトウェアで実現され、高次元特徴空間において線形関数の仮説空間を用いる学習システムであり、その学習結果は、妥当データの集合と非妥当データの集合とを識別するための識別面、及びサポートベクターを含む。

【0035】

また、教師なし機械学習の例は、クラスタリング(クラスター分析)、主成分分析、ベクトル量子化、自己組織化マップなどがあり、クラスタリングのなかでも、階層クラスタリングとして最短距離法(単連結法)、最長距離法(完全連結法)、群平均法、ワード法などがあり、分割最適化クラスタリングとしてK平均法(k-means)などがあり

10

20

30

40

50

、ワンパス・クラスタリングなども含め、任意の公知の技術を用いることができる。

【0036】

そして、ここでクラスタリングに用いる潜在的意味インデキシング (LSI: Latent Semantic Indexing) は、大規模なテキストデータ群から単語の意味を、意味空間上のベクトルとして表現するための数学的・統計的手法であり、文書ごとの特徴的な語句集合、類似の文脈で使用されている語句集合を抽出し、抽出された特徴語句でタグ付けされた類似文書ごとにクラスタリングするものである。

【0037】

特に、階層的クラスタリングによって、ベクトルの類似度を用い、類似している文書 (あるいは語句) をクラスタリングすることができ、図2 (概念図) に例示するように、見たい所望のクラスタ階層で切り取ることができる。なお、各文書には、その文書の特徴付ける特徴語 (概念語) が付与されているが、特徴語は、必ずしも文書内に直接明記されているものには限定されない。

【0038】

但し、潜在的意味インデキシングを用いたクラスタリングは必須ではなく、他の機械学習アルゴリズムに置き換えたり組み合わせれば、それら他の機械学習アルゴリズムの利点も活用可能となる。すなわち、クラスタリングと他の機械学習は、いずれか一方のみを用いてもよいし、双方の手段を用意しておき、使い分けたり、組み合わせ併用してもよい。

【0039】

例えば、図3の概念図は、引用判定部40が、潜在的意味インデキシングを用いたクラスタリングを行うクラスタリング部42と、他の機械学習を行う機械学習部44と、を併有する可能性も示している。

【0040】

そして、潜在的意味インデキシングを用いたクラスタリングでは、引用判定部40 (特にクラスタリング部42) は、各ブログに各ニュースX1などを合わせた対象アイテム41から、上記のような潜在的意味インデキシングを用いたクラスタリングにより、各ニュースを含むクラスタを取得することにより、ニュースに対し所定以上の類似度を持つ各ブログを抽出する。

【0041】

例えば、ニュースX1と同じクラスタx1に含まれる各ブログは、ニュースX1の引用の可能性が高く、同様に、ニュースY1と同じクラスタy1に含まれる各ブログは、ニュースY1の引用の可能性が高い。これら各ブログは、ニュースからの引用と判定できる。一方、いずれのニュースとも異なるクラスタbを構成する各ブログは、ニュースからの引用の可能性は低く、例えば、親記事であるブログや、ブログに対するコメントと判定できる。

【0042】

このように、引用有無判定の対象とするブログ群に、引用される側すなわちターゲットクラスであるニュースを、いわば従来の教師データの役割を兼ねる本番データとして加えた上で全体をクラスタリングし、ニュースを含むクラスタを取得することにより、ニュースへの類似度すなわち引用可能性の高い高純度なブログを抽出できるうえ、SVMなどオンライントレーニングタイプの従来型機械学習よりも、新規の様々な内容や記載スタイルのニュースへの対応が迅速容易になる。

【0043】

〔評判の抽出〕

本装置では、上記のような引用判定の結果を、評判抽出の高精度化に利用し、引用判定部40で引用でないと判定したブログについて、通常の評判抽出を行う。すなわち、評判対象抽出部50が、ブログから評判対象を抽出し (評判対象抽出処理ステップ)、評判情報抽出部45が、前記評判対象に対する評判を表す評判情報をブログから抽出し (評判情報抽出処理ステップ)、評判判定部60が、これら抽出した評判対象と評判情報と、に基

10

20

30

40

50

いて評判を判定する（評判判定処理ステップ）。

【0044】

評判の判定は、多数のブログをもとに、評判対象である会社名、組織名、サービス名、商品名、政策名などごとに、「良い」「悪い」「好き」「嫌い」「ヤバイ」「イマイチ」などの評判情報を、予め分類された肯定的、否定的評判情報の別ごとに集計したり、評判情報のレベル別に予め付与したスコアを集計するなどによって行う。

【0045】

そして、このような評判情報やその判定結果は、表計算ワークシートなどのファイルや、グラフなどの表示等の形で、所定の評判情報記憶部65に保存し、また、操作に応じて外部へ出力する。

10

【0046】

このように、本発明の引用判定で引用でないと判定したユーザ記事について、評判対象と評判情報の抽出対象とすることにより、ニュースなどに含まれる好悪等の評判情報を除いて評判抽出を行うこととなり、評判判定の精度を効果的に向上させることが可能となる。

【0047】

〔段落への分割〕

以上のような処理は、ブログの全体、ニュースの全体に対して行っても、もちろん本発明の優れた効果を奏するが、ブログやニュースを段落単位に分割して適用すれば、一部引用についても優れた効果を奏する。

20

【0048】

この場合、例えば、ブログ収集部5で収集した各ブログAやニュース取得部10で取得した各ニュースXは、それぞれブログ分割部15及びニュース分割部20が段落ごとにブログA1とA2、ニュースX1とX2のように分割して（段落分割処理ステップ）、引用判定部40は、このように分割された各ブログの各段落及び各ニュースの各段落を対象として、前記引用の判定を行う。

【0049】

このように、ブログとニュースの双方を段落単位に分割のうえ類似度判断による引用判定を行うことにより、ニュースの一部引用についても高精度に判定可能となる。

【0050】

また、ブログを分割した一部の段落がニュースからの引用部分であると引用判定部40が判定した場合に、評判対象抽出部50は、その引用部分から評判対象を抽出する一方、評判情報抽出部45は、同じユーザ記事のうち前記引用部分を除く段落から評判情報を抽出する。

30

【0051】

例えば、図1の例において、段落に分割されたブログA1、A2、B1、B2、C1、C2のうち、ブログA1、B1がニュースからの引用部分と判定されたと仮定すると、評判情報抽出部45が評判情報の抽出対象とするのは、引用でないブログC1、C2のほか、引用でない部分のブログA2、B2である。一方、評判対象抽出部50が評判対象の抽出を行うのは、引用でないブログC1、C2のほか、引用部分であるブログA1、B1である。

40

【0052】

〔フローの例〕

以上のような機能作用を実現する情報処理の手順を図4のフローチャートに例示する。この処理手順の例では、ブログの収集（ステップS1）及びニュースの収集を（ステップS2）所定のタイミングまで行ったうえ（ステップS3）、段落単位で処理する設定の場合は（ステップS4）収集・取得したブログ及びニュースを段落単位に分割する（ステップS5）。

【0053】

そして、ブログとニュース全体又は上記のように分割した段落を単位として、各ブログ

50

がいずれかのニュースの引用かを判定し（ステップ S 6）、この判定結果に応じて各ブログから評判情報と評判対象とを抽出し（ステップ S 7）、抽出した評判情報と評判対象から評判を判定する（ステップ S 8）。

【 0 0 5 4 】

また、図 4 において判定結果に応じて各ブログから評判情報と評判対象とを抽出する処理（ステップ S 7）を、より具体化したフローチャートを図 5 に示す。すなわち、引用の判定済みで評判に関する抽出が未処理のブログ（段落単位の処理の場合はブログの段落）を一つ取り出し（ステップ S 7 1）、それがニュースの引用についてどのように判定されたものかを判断する（ステップ S 7 2）。

【 0 0 5 5 】

この判断で、ブログ全体がニュースの引用であれば（ステップ S 7 2）、そのブログについては評判情報も評判対象も抽出しないが、引用無しとの判定であれば（ステップ S 7 2）、そのブログ全体から評判情報と評判対象を抽出する（ステップ S 7 3）。

【 0 0 5 6 】

また、一部段落が引用との判定だった場合は（ステップ S 7 2）、引用でない段落からは評判情報を抽出する（ステップ S 7 4）一方、同じブログのうち引用である段落からは評判対象を抽出する（ステップ S 7 5）。いずれの場合も、その後、判定済みで未処理のブログがまだ残っていれば（ステップ S 7 6）、次のブログ（又はその段落）の処理を続ける（ステップ S 7 1以降）。

【 0 0 5 7 】

このように、一部の段落がニュースの引用と判定した場合に、引用されているニュースの部分からは、主題となる企業名、製品名、人名等の評判対象を抽出することで評判対象が明確となることに加え、引用以外の段落からはその対象についての評判情報を抽出することにより、ニュースを引用しているブログも評判判定に有効活用可能となる。

【 0 0 5 8 】

〔評判情報等の表示〕

上記のように抽出した評判情報は、ニュースごとに、関連する各ブログとともに画面表示することが望ましい。具体的には、評判情報記憶部 6 5 に、ニュースごとに、一部の段落がそのニュースからの引用部分であると引用判定部 4 0 で判定した各ブログ（ブログ内容ではなく、ブログを特定する識別情報でよい）と、各ブログから評判情報抽出部 4 5 で抽出した前記評判情報と、を記憶しておき、情報出力部 7 0 が、コンピュータに設けた図示しない画面表示装置もしくはインターネット N 経由で接続される他のコンピュータであるクライアント端末 C に、前記評判情報を出力する。

【 0 0 5 9 】

この際、情報出力部 7 0 は、評判情報記憶部 6 5 に記憶されている上記各情報に基づいて、図 6 の表示例に示すように、ニュース 8 0 ごとに、そのニュースからの引用部分を持つ各ブログ、すなわち引用ブログ 8 1、8 2、8 3 と、各ブログから評判情報抽出部 4 5 で抽出した前記評判情報と、を表示する。記憶しておいたり表示出力する評判情報の態様は自由で、例えば、肯定的評価を「○」、否定的評価を「×」、中間的評価を「△」のようにブログ単位に表示してもよいし、それらの集計結果 8 5 を数値やグラフなどとして表示してもよい。

【 0 0 6 0 】

このように、個々のニュースに対して、そのニュースを引用した各ブログと、各ブログから抽出した評判情報と、を一画面で表示するユーザインタフェースにより、各ニュースに対するユーザの記述や評判が一見把握容易になり、広報等の情報政策への有効活用が容易になる。

【 0 0 6 1 】

〔他の実施形態〕

なお、上記実施形態は例示に過ぎず、本発明は、以下の例及び他の実施形態も含むものである。例えば、ユーザ記事はブログに限らず電子掲示板の投稿など自由に選択可能であ

10

20

30

40

50

り、同様に、公表記事もニュースに限らず、学術論文や、白書のような公共機関による公表資料など、自由に選択可能である。

【 0 0 6 2 】

また、ユーザ記事や公表記事を段落単位に分割して扱う構成や処理は、必須ではなく省略も可能である。さらに、本発明は引用判定装置 1 のみでも実施可能であり、この場合、図 1 に示した評判判定装置 2 特有の構成である評判情報抽出部 4 5、評判対象抽出部 5 0、評判判定部 6 0、評判情報記憶部 6 5 は省略可能である。

【 0 0 6 3 】

また、クラスタリングにおいて、共通のカテゴリタグを含むなど相互に類似ニュースと予め判っているものが一つのクラスタに存在する数の正確さで、全対象物を最適なクラスタ数、すなわち樹形図状の多段包含（統合）関係を有するクラスタに区分する適切な基準が判定可能である。

10

【 0 0 6 4 】

この場合のアルゴリズムの例としては、例えば、類似度が高いもの同士による部分集合の生成から始めて、ボトムアップ方向に、部分集合同士をまとめるクラスタリングを進める。そして、全てのクラスタを見て、同じカテゴリタグのニュースが全て含まれた時点でそのカテゴリについてはそのときのクラスタで決定とする。

【 0 0 6 5 】

一方、この際でも、他のカテゴリタグのニュースが、まだ異なる複数のカテゴリにわたって配分されていた場合は、そのカテゴリについては、もう一段上で統合されるクラスタに範囲を広げて確認を行う。

20

【 0 0 6 6 】

例えば、図 7 の例において、同じ「オリンピック」というカテゴリタグ（アスタリスク記号で表す）を含む文書を 5 つ含む文書群をクラスタリングした結果、語句集合に基づく各文書間の類似性に基づきボトムアップ式に、図 7 に例示するツリー構造が判明した場合を考える。

【 0 0 6 7 】

この場合、破線 L 1 の位置で他と区切ることとなるクラスタ C L 1（破線で囲んだ範囲）を採用すれば、同じ「オリンピック」カテゴリタグの文書 5 つがちょうど全部含まれることから、適切なクラスタ分けと判断できる。なお、共通のカテゴリタグを持つニュースの必ずしも全数を含むクラスタに限らず、若干の誤差を許容値としてクラスタを決定することも可能である。

30

【図面の簡単な説明】

【 0 0 6 8 】

【図 1】本発明の実施形態の構成を示す機能ブロック図。

【図 2】クラスタリングの結果得た階層的なクラスタを、所望の見たいクラスタ階層で切り取ってまとめる様子を示す概念図。

【図 3】本発明の実施形態における引用判定を示す概念図。

【図 4】本発明の実施形態における処理手順を示すフローチャート。

【図 5】本発明の実施形態において、引用判定の結果に応じてブログから評判情報及び評判対象を抽出する処理手順を示すフローチャート。

40

【図 6】本発明の実施形態において、評判判定結果の画面表示例を示す図。

【図 7】本発明の実施形態におけるクラスタリングの一例を示す概念図。

【符号の説明】

【 0 0 6 9 】

N インターネット

1 引用判定装置

2 評判判定装置

5 ブログ収集部

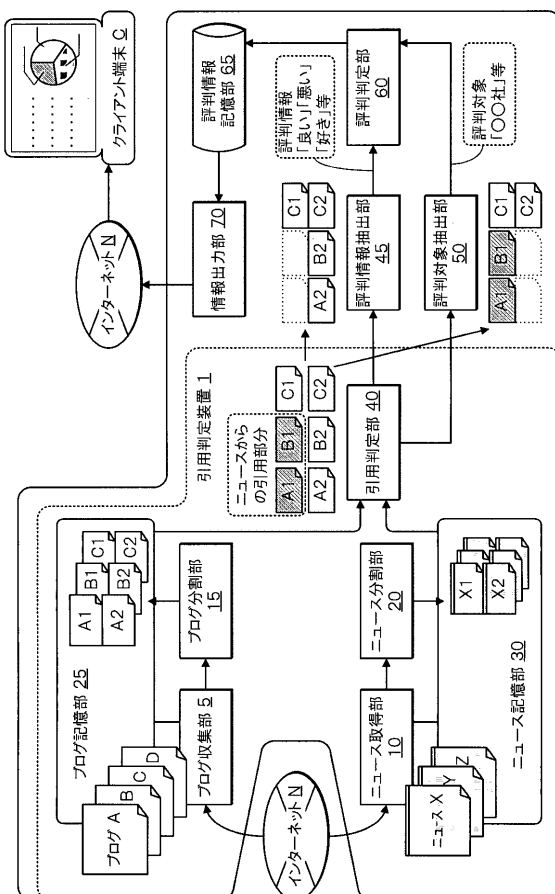
1 0 ニュース取得部

50

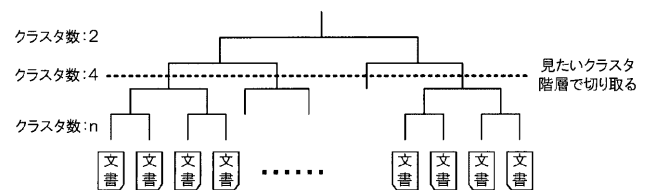
- 15 ブログ分割部
- 20 ニュース分割部
- 25 ブログ記憶部
- 30 ニュース記憶部
- 40 引用判定部
- 45 評判情報抽出部
- 50 評判対象抽出部
- 60 評判判定部
- 65 評判情報記憶部
- 70 結果出力部

10

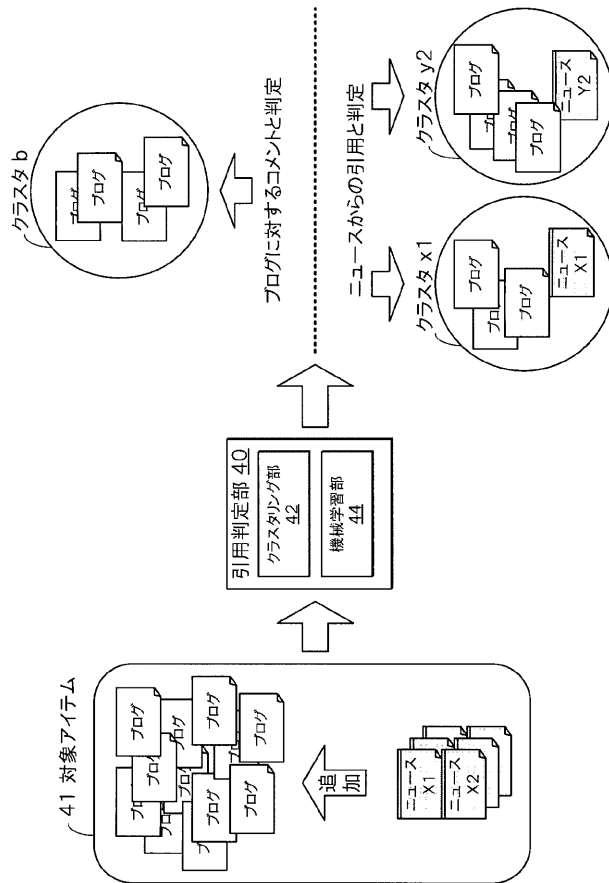
【図1】



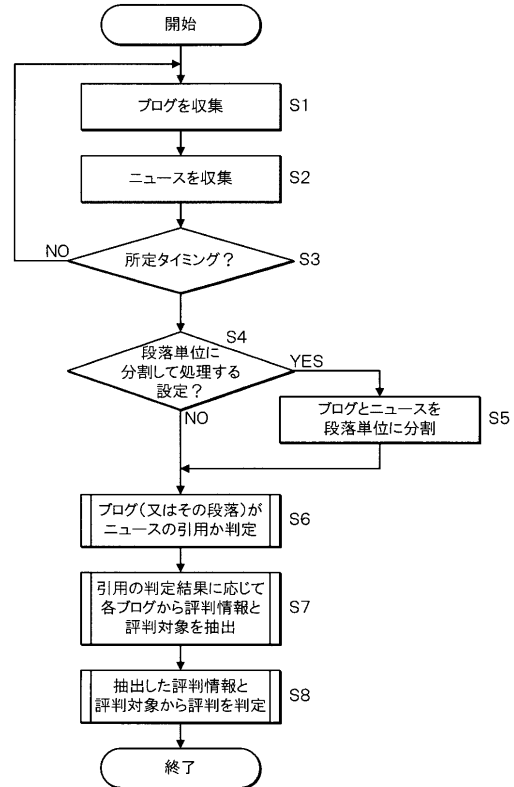
【図2】



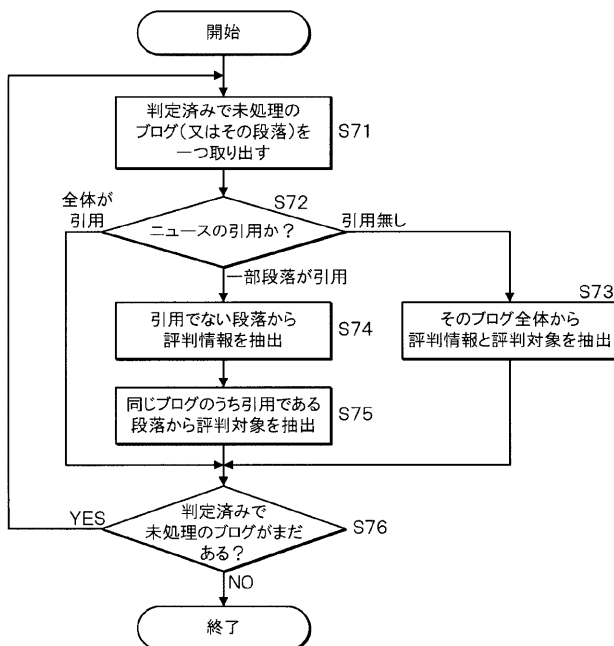
【図 3】



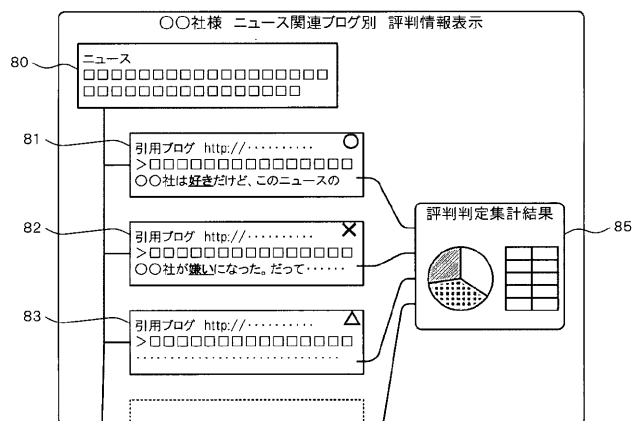
【図 4】



【図 5】



【図 6】



【図 7】

