

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7513396号
(P7513396)

(45)発行日 令和6年7月9日(2024.7.9)

(24)登録日 令和6年7月1日(2024.7.1)

(51)国際特許分類

F I

G 0 6 F 16/33 (2019.01)

G 0 6 F 16/33

G 0 6 F 16/383 (2019.01)

G 0 6 F 16/383

請求項の数 23 (全29頁)

(21)出願番号	特願2019-569437(P2019-569437)	(73)特許権者	510280589
(86)(22)出願日	平成30年6月19日(2018.6.19)		京東方科技集團股▲ふん▼有限公司
(65)公表番号	特表2021-510858(P2021-510858 A)		BOE TECHNOLOGY GROU P CO. , LTD.
(43)公表日	令和3年4月30日(2021.4.30)		中華人民共和國100015北京市朝陽區酒仙橋路10號
(86)国際出願番号	PCT/CN2018/091822		No. 10 Jiuxianqiao R d. , Chaoyang Distri ct, Beijing 100015, CHINA
(87)国際公開番号	WO2019/140863	(74)代理人	100108453
(87)国際公開日	令和1年7月25日(2019.7.25)		弁理士 村山 靖彦
審査請求日	令和3年6月16日(2021.6.16)	(74)代理人	100110364
審判番号	不服2023-12978(P2023-12978/J 1)		弁理士 実広 信哉
審判請求日	令和5年8月3日(2023.8.3)	(72)発明者	ジェンジョン・ジャン
(31)優先権主張番号	201810060942.7		
(32)優先日	平成30年1月22日(2018.1.22)		
(33)優先権主張国・地域又は機関			
	最終頁に続く		最終頁に続く

(54)【発明の名称】 関連性を算出する方法、関連性を算出する装置、データクエリ装置及び非一時的なコンピュータ可読記録媒体

(57)【特許請求の範囲】

【請求項1】

第1のテキストと第2のテキストとの間の関連性を算出する方法であって、前記方法は少なくとも一つのプロセッサによって実行され、かつ、前記方法は、
前記第1のテキスト及び前記第2のテキストをベクトルにそれぞれマッピングすることと、

前記ベクトルに基づいて前記第1のテキストと前記第2のテキストとの間の類似部分及び相違部分を判断することと、

前記類似部分及び前記相違部分の両方を用いて前記第1のテキストと前記第2のテキストとの間の関連性を算出することとを含み、

前記類似部分及び相違部分を判断することは、前記第1のテキストおよび前記第2のテキストに基づいてセマンティックコンテンツ分析を実行することを含み、

前記セマンティックコンテンツ分析は、前記第1のテキストと前記第2のテキストとの間でセマンティックマッチング処理を行って前記第1のテキストと前記第2のテキストとの間のセマンティックオーバーレイを取得することを含み、

前記セマンティックマッチング処理を行うことは、前記第2のテキストの単語のベクトルを用いて前記第1のテキストにおける単語のベクトルを再構成することと、前記再構成の結果に基づいてセマンティックオーバーレイを算出することとを含み、

前記セマンティックオーバーレイは、式(1)に基づいて算出され、

【数1】

$$\min(\left\|S_i - \sum_{j=1}^m \alpha_{i,j} \times T_j\right\|^2 + \lambda \sum_{j=1}^m \alpha_{i,j}^2)$$

—式 (1)

ここで、 S_i は前記第 1 のテキストの $d \times 1$ 列ベクトルであり、 T_j は前記第 2 のテキストの $d \times 1$ 列ベクトルであり、 $\alpha_{i,j}$ はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数であり、

前記セマンティックオーバーレイを算出した後に、前記第 1 のテキスト及び前記第 2 のテキストに対してセマンティック分解を行って、前記類似部分及び前記相違部分を判断し、

前記第 1 のテキストと前記第 2 のテキストとの間の前記類似部分及び前記相違部分を入力として用いて、リカレントニューラルネットワークを用いて前記第 1 のテキストと前記第 2 のテキストとの間の関連性を算出する、

方法。

【請求項 2】

前記類似部分及び前記相違部分を判断することは、

前記第 1 のテキストの類似部分及び相違部分を判断することと、

前記第 2 のテキストの類似部分及び相違部分を判断することを含む、請求項 1 に記載の方法。

【請求項 3】

前記第 1 のテキスト及び前記第 2 のテキストをベクトルにそれぞれマッピングした後、前記ベクトルに対して次元削減を行って、低次元密ベクトルを用いて前記第 1 のテキスト及び前記第 2 のテキストを表すことをさらに含む、請求項 1 に記載の方法。

【請求項 4】

次元削減は、Word2Vec、Sentence2Vec 及び Doc2Vec からなる群より選択される 1 つ以上の方法を用いて行われる、請求項 3 に記載の方法。

【請求項 5】

前記第 1 のテキストの前記類似部分は、

【数 2】

$$S'_i = \sum_{j=1}^m A \times T_j$$

に基づいて算出され、

ここで、 S'_i は前記第 1 のテキストの前記類似部分を表し、 A は $\alpha_{i,j}$ の行列であり、

$S_i - S'_i$ は前記第 1 のテキストの前記相違部分を表す、請求項 1 に記載の方法。

【請求項 6】

前記セマンティックオーバーレイは、式 (1) に加えて、式 (2) に基づいて算出され、

【数 3】

$$\min(\left\|T_i - \sum_{j=1}^m \beta_{i,j} \times S_j\right\|^2 + \lambda \sum_{j=1}^m \beta_{i,j}^2)$$

—式 (2)

ここで、 T_i は前記第 2 のテキストの $d \times 1$ 列ベクトルであり、 S_j は前記第 1 のテキストの $d \times 1$ 列ベクトルであり、 $\beta_{i,j}$ はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数である、請求項 1 に記載の方法。

【請求項 7】

前記第 2 のテキストの前記類似部分は、

【数 4】

10

20

30

40

50

$$T_{j'} = \sum_{j=1}^m A \times S_j$$

に基づいて算出され、

ここで、 $T_{j'}$ は前記第 2 のテキストの前記類似部分を表し、 A は β_j の行列であり、 $T_{j'} - T_{j'}$ は、前記第 2 のテキストの前記相違部分を表す、請求項 6 に記載の方法。

【請求項 8】

前記入力は、

前記第 1 のテキストの類似部分及び相違部分と、

前記第 2 のテキストの類似部分及び相違部分とを含む、請求項 1 に記載の方法。

【請求項 9】

(S_s, T_s, L) で表されるサンプルデータを用いて前記リカレントニューラルネットワークをトレーニングすることをさらに含み、ここで、 S_s は第 1 のサンプルテキストを表し、 T_s は第 2 のサンプルテキストを表し、 L はサンプルの関連性を表す、請求項 1 に記載の方法。

【請求項 10】

前記リカレントニューラルネットワークをトレーニングすることは、

第 1 の範囲にある第 1 の関連性に第 1 の粒度値を、第 2 の範囲にある第 2 の関連性に第 2 の粒度値を割り当てることをさらに含み、

粒度が 2 に設定されるとき、前記第 1 の粒度値は 1 であり、かつ前記第 2 の粒度値は 0 であり、粒度が 3 に設定されるとき、前記第 1 の粒度値は 1 であり、かつ前記第 2 の粒度値は 0.5 であり、かつ前記第 1 の範囲および前記第 2 の範囲は互いに異なっている、請求項 9 に記載の方法。

【請求項 11】

第 1 のテキストと第 2 のテキストとの間の関連性を算出する装置であって、メモリと、

1 つ以上のプロセッサとを備え、

前記メモリと、前記 1 つ以上のプロセッサとは互いに接続され、

前記メモリは、

前記第 1 のテキスト及び前記第 2 のテキストをベクトルにそれぞれマッピングし、

前記ベクトルに基づいて前記第 1 のテキストと前記第 2 のテキストとの間の類似部分及び相違部分を判断し、

前記類似部分及び前記相違部分の両方を用いて前記第 1 のテキストと前記第 2 のテキストとの間の関連性を算出するように、前記 1 つ以上のプロセッサを制御するためのコンピュータ実行可能命令を記憶し、

前記メモリは、前記第 1 のテキスト及び前記第 2 のテキストに基づいてセマンティックコンテンツ分析を行って前記類似部分及び前記相違部分を判断するように、前記 1 つ以上のプロセッサを制御するためのコンピュータ実行可能命令をさらに記憶し、

前記メモリは、前記第 1 のテキストと前記第 2 のテキストとの間でセマンティックマッチング処理を行って前記第 1 のテキストと前記第 2 のテキストとの間のセマンティックオーバーレイを取得するように、前記 1 つ以上のプロセッサを制御するためのコンピュータ実行可能命令をさらに記憶し、

前記メモリは、前記第 2 のテキストの単語のベクトルを用いて前記第 1 のテキストにおける単語のベクトルを再構成し、再構成の結果に基づいてセマンティックオーバーレイを算出するように、前記 1 つ以上のプロセッサを制御するためのコンピュータ実行可能命令をさらに記憶し、

前記セマンティックオーバーレイは、式 (1) に基づいて算出され、

【数 5】

10

20

30

40

50

$$\min(\left\|S_i - \sum_{j=1}^m \alpha_{i,j} \times T_j\right\|^2 + \lambda \sum_{j=1}^m \alpha_{i,j}^2) \quad \text{—式 (1)}$$

ここで、 S_i は前記第 1 のテキストの $d \times 1$ 列ベクトルであり、 T_j は前記第 2 のテキストの $d \times 1$ 列ベクトルであり、 $\alpha_{i,j}$ はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数であり、

前記メモリは、

前記セマンティックオーバーレイを算出した後に、前記第 1 のテキスト及び前記第 2 のテキストに対してセマンティック分解を行って、前記類似部分及び前記相違部分を判断し、

前記第 1 のテキストと前記第 2 のテキストとの間の前記類似部分及び前記相違部分を入力として用いて、リカレントニューラルネットワークを用いて前記第 1 のテキストと前記第 2 のテキストとの間の関連性を算出するように、

前記 1 つ以上のプロセッサを制御するためのコンピュータ実行可能命令をさらに記憶する、

装置。

【請求項 1 2】

前記メモリは、

前記第 1 のテキストの類似部分及び相違部分を判断し、

前記第 2 のテキストの類似部分及び相違部分を判断するように、前記 1 つ以上のプロセッサを制御するためのコンピュータ実行可能命令を記憶する、請求項 1 1 に記載の装置。

【請求項 1 3】

前記メモリは、前記第 1 のテキスト及び前記第 2 のテキストをベクトルにそれぞれマッピングした後に、前記ベクトルに対して次元削減を行って低次元密ベクトルを用いて前記第 1 のテキスト及び前記第 2 のテキストを表すように、前記 1 つ以上のプロセッサを制御するためのコンピュータ実行可能命令をさらに記憶する、請求項 1 1 に記載の装置。

【請求項 1 4】

次元削減は、Word 2 Vec、Sentence 2 Vec 及び Doc 2 Vec からなる群から選択される 1 つ以上の方法を用いて行われる、請求項 1 3 に記載の装置。

【請求項 1 5】

前記第 1 のテキストの前記類似部分は、

【数 6】

$$S'_i = \sum_{j=1}^m A \times T_j$$

に基づいて算出され、

ここで、 S'_i は前記第 1 のテキストの前記類似部分を表し、 A は $\alpha_{i,j}$ の行列であり、

$S_i - S'_i$ は前記第 1 のテキストの相違部分を表す、請求項 1 1 に記載の装置。

【請求項 1 6】

前記セマンティックオーバーレイは、式 (1) に加えて、式 (2) に基づいて算出され、

【数 7】

$$\min(\left\|T_i - \sum_{j=1}^m \beta_{i,j} \times S_j\right\|^2 + \lambda \sum_{j=1}^m \beta_{i,j}^2) \quad \text{—式 (2)}$$

ここで、 T_i は前記第 2 のテキストの $d \times 1$ 列ベクトルであり、 S_j は前記第 1 のテキ

ストの $d \times 1$ 列ベクトルであり、 $\beta_{i,j}$ はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数である、請求項 11 に記載の装置。

【請求項 17】

前記第 2 のテキストの前記類似部分は、

【数 8】

$$T_{j'} = \sum_{j=1}^m A \times S_j$$

に基づいて算出され、

ここで、 $T_{j'}$ は前記第 2 のテキストの前記類似部分を表し、 A は $\beta_{i,j}$ の行列であり、 $T_j - T_{j'}$ は、前記第 2 のテキストの前記相違部分を表す、請求項 16 に記載の装置。

【請求項 18】

前記入力は、

前記第 1 のテキストの類似部分及び相違部分と、

前記第 2 のテキストの類似部分及び相違部分とを含む、請求項 11 に記載の装置。

【請求項 19】

前記メモリは、 (S_s, T_s, L) で表されるサンプルデータを用いて前記リカレントニューラルネットワークをトレーニングするように、前記 1 つ以上のプロセッサを制御するためのコンピュータ実行可能命令をさらに記憶し、 S_s は第 1 のサンプルテキストを表し、 T_s は第 2 のサンプルテキストを表し、 L はサンプルの関連性を表す、請求項 11 に記載の装置。

【請求項 20】

前記メモリは、第 1 の範囲にある第 1 の関連性に第 1 の粒度値を、第 2 の範囲にある第 2 の関連性に第 2 の粒度値を割り当てるように、前記 1 つ以上のプロセッサを制御するためのコンピュータ実行可能命令をさらに記憶し、粒度が 2 に設定されるとき、前記第 1 の粒度値は 1 であり、かつ前記第 2 の粒度値は 0 であり、粒度が 3 に設定されるとき、前記第 1 の粒度値は 1 であり、かつ前記第 2 の粒度値は 0.5 であり、かつ前記第 1 の範囲および前記第 2 の範囲は互いに異なっている、請求項 19 に記載の装置。

【請求項 21】

請求項 11 から 20 のいずれか 1 項に記載の関連性を算出する装置を備える、データクエリ装置。

【請求項 22】

請求項 1 から 10 のいずれか 1 項に記載の方法により第 1 のテキストと複数の潜在的なターゲットテキストから選択される第 2 のテキストとの間の関連性を算出することと、

前記複数の潜在的なターゲットテキストと前記第 1 のテキストとの間の関連性の算出に基づいて、前記複数の潜在的なターゲットテキストをランク付けすることと、

前記複数の潜在的なターゲットテキストのランク付けの結果に基づいて、前記複数の潜在的なターゲットテキストからターゲットテキストを選択することを含む、データクエリ方法。

【請求項 23】

コンピュータ可読命令を含むコンピュータプログラムであって、

前記コンピュータ可読命令は、プロセッサによって実行可能であり、前記プロセッサに、第 1 のテキスト及び第 2 のテキストをベクトルにそれぞれマッピングさせ、

前記ベクトルに基づいて前記第 1 のテキストと前記第 2 のテキストとの間の類似部分及び相違部分を判断させ、

前記類似部分及び前記相違部分の両方を用いて第 1 のテキストと第 2 のテキストとの間の関連性を算出させ、

前記類似部分及び相違部分を判断することは、前記第 1 のテキストおよび前記第 2 のテキストに基づいてセマンティックコンテンツ分析を実行することを含み、

10

20

30

40

50

前記セマンティックコンテンツ分析は、前記第 1 のテキストと前記第 2 のテキストとの間でセマンティックマッチング処理を行って前記第 1 のテキストと前記第 2 のテキストとの間のセマンティックオーバーレイを取得することを含み、

前記セマンティックマッチング処理を行うことは、前記第 2 のテキストの単語のベクトルを用いて前記第 1 のテキストにおける単語のベクトルを再構成することと、前記再構成の結果に基づいてセマンティックオーバーレイを算出することを含み、

前記セマンティックオーバーレイは、式 (1) に基づいて算出され、

【数 9】

$$\min(\|S_i - \sum_{j=1}^m \alpha_{i,j} \times T_j\|^2 + \lambda \sum_{j=1}^m \alpha_{i,j}^2) \quad \text{—式 (1)}$$

ここで、 S_i は前記第 1 のテキストの $d \times 1$ 列ベクトルであり、 T_j は前記第 2 のテキストの $d \times 1$ 列ベクトルであり、 $\alpha_{i,j}$ はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数であり、

前記コンピュータ可読命令は、プロセッサによって実行可能であり、さらに前記プロセッサに、

前記セマンティックオーバーレイを算出した後に、前記第 1 のテキスト及び前記第 2 のテキストに対してセマンティック分解を行って、前記類似部分及び前記相違部分を判断させ、

前記第 1 のテキストと前記第 2 のテキストとの間の前記類似部分及び前記相違部分を入力として用いて、リカレントニューラルネットワークを用いて前記第 1 のテキストと前記第 2 のテキストとの間の関連性を算出させる、

コンピュータプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

〔関連出願の相互参照〕

本願は、2018年1月22日に提出された中国特許出願201810060942.7号の優先権を主張し、そのすべての内容を参照によりここに援用する。

【0002】

本発明は、データ処理技術に関し、特に、関連性を算出する方法、関連性を算出する装置、データクエリ装置及び非一時的なコンピュータ可読記録媒体に関する。

【背景技術】

【0003】

インターネットやビッグデータ技術が発展するに伴い、膨大な情報量が得られるようになっている。ユーザがキーワード入力に基づいて有用な情報を検索しやすいように、様々な検索エンジンが開発されている。例えば、医療従事者は、サーチクエリに基づいて様々な医療文献、書籍、インターネットのウェブサイトを検索することができる。患者は、健康情報ウェブサイト健康関連情報をさがすことができる。

【発明の概要】

【課題を解決するための手段】

【0004】

1つの方面において、本発明は、第1のテキストと第2のテキストとの間の関連性を算出する方法であって、前記第1のテキスト及び前記第2のテキストをベクトルにそれぞれマッピングすることと、前記ベクトルに基づいて前記第1のテキストと前記第2のテキストとの間の類似部分及び相違部分を判断することと、前記類似部分及び前記相違部分の両方を用いて前記第1のテキストと前記第2のテキストとの間の関連性を算出することを含む、方法を提供する。

【0005】

或いは、前記類似部分及び前記相違部分を判断することは、前記第1のテキストの類似部分及び相違部分を判断することと、前記第2のテキストの類似部分及び相違部分を判断することを含んでもよい。

【0006】

或いは、前記第1のテキスト及び前記第2のテキストをベクトルにそれぞれマッピングした後、前記ベクトルに対して次元削減を行って、低次元密ベクトルを用いて前記第1のテキスト及び前記第2のテキストを表すことをさらに含んでもよい。

【0007】

或いは、次元削減は、Word2Vec、Sentence2Vec及びDoc2Vecからなる群より選択される1つ以上の方法を用いて行われてもよい。

10

【0008】

或いは、前記類似部分及び相違部分を判断することは、前記第1のテキスト及び前記第2のテキストに基づいてセマンティックコンテンツ分析を行うことを含んでもよい。

【0009】

或いは、前記セマンティックコンテンツ分析は、前記第1のテキストと前記第2のテキストとの間でセマンティックマッチング処理を行って前記第1のテキストと前記第2のテキストとの間のセマンティックオーバーレイを取得することを含んでもよい。

【0010】

或いは、前記セマンティックマッチング処理を行うことは、前記第2のテキストの単語のベクトルを用いて前記第1のテキストにおける単語のベクトルを再構成することと、前記再構成の結果に基づいてセマンティックオーバーレイを算出することを含んでもよい。

20

【0011】

或いは、前記類似部分及び相違部分を判断することは、前記第1のテキスト及び前記第2のテキストに対してセマンティック分解を行うことを含んでもよい。

【0012】

或いは、前記セマンティックオーバーレイは、式(1)に基づいて算出されてもよい。

【数1】

$$\min(\left\|S_i - \sum_{j=1}^m \alpha_{ij} \times T_j\right\|^2 + \lambda \sum_{j=1}^m \alpha_{ij}^2) \quad (1)$$

30

【0013】

ここで、 S_i は前記第1のテキストの $d \times 1$ 列ベクトルであり、 T_j は前記第2のテキストの $d \times 1$ 列ベクトルであり、 α_{ij} はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数である。

【0014】

或いは、第1のテキストの類似部分は、

40

【数2】

$$S_i' = \sum_{j=1}^m A \times T_j$$

に基づいて算出され、ここで、 S_i' は前記第1のテキストの前記類似部分を表し、 A は α_{ij} の行列であり、 $S_i - S_i'$ は前記第1のテキストの前記相違部分を表してもよい。

【0015】

或いは、前記セマンティックオーバーレイは、式(2)に基づいて算出されてもよい。

【数3】

50

$$\min(\left\|T_i - \sum_{j=1}^m \beta_{ij} \times S_j\right\|^2 + \lambda \sum_{j=1}^m \beta_{ij}^2) \quad (2)$$

【0016】

ここで、 T_i は前記第2のテキストの $d \times 1$ 列ベクトルであり、 S_j は前記第1のテキストの $d \times 1$ 列ベクトルであり、 $\beta_{i,j}$ はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数である。

【0017】

或いは、前記第2のテキストの前記類似部分は、

【数4】

$$T_j' = \sum_{j=1}^m A \times S_j$$

に基づいて算出され、ここで、 T_j' は前記第2のテキストの前記類似部分を表し、 A は $\beta_{i,j}$ の行列であり、 $T_j - T_j'$ は前記第2のテキストの前記相違部分を表してもよい。

【0018】

或いは、前記第1のテキストと前記第2のテキストとの間の前記類似部分及び前記相違部分を入力として用いて、リカレントニューラルネットワークを用いて前記第1のテキストと前記第2のテキストとの間の関連性を算出してもよい。

【0019】

或いは、前記入力、前記第1のテキストの類似部分及び相違部分と、前記第2のテキストの類似部分及び相違部分とを含んでもよい。

【0020】

或いは、この方法は、 (S_s, T_s, L) で表されるサンプルデータを用いて前記リカレントニューラルネットワークをトレーニングすることをさらに含み、ここで、 S_s は第1のサンプルテキストを表し、 T_s は第2のサンプルテキストを表し、 L はサンプルの関連性を表してもよい。

【0021】

或いは、前記リカレントニューラルネットワークをトレーニングすることは、第1の範囲にある第1の関連性に第1の粒度値を、第2の範囲にある第2の関連性に第2の粒度値を割り当てることをさらに含んでもよい。

【0022】

別の方面において、本発明は、第1のテキストと第2のテキストとの間の関連性を算出する装置であって、メモリと、1つ以上のプロセッサとを備え、前記メモリと、前記1つ以上のプロセッサとは互いに接続され、前記メモリは、前記第1のテキスト及び前記第2のテキストをベクトルにそれぞれマッピングし、前記ベクトルに基づいて前記第1のテキストと前記第2のテキストとの間の類似部分及び相違部分を判断し、前記類似部分及び前記相違部分の両方を用いて前記第1のテキストと前記第2のテキストとの間の関連性を算出するように、前記1つ以上のプロセッサを制御するためのコンピュータ実行可能命令を記憶する、装置を提供する。

【0023】

或いは、前記メモリは、前記第1のテキストの類似部分及び相違部分を判断し、前記第2のテキストの類似部分及び相違部分を判断するように、前記1つ以上のプロセッサを制御するためのコンピュータ実行可能命令を記憶してもよい。

【0024】

或いは、前記メモリは、前記第1のテキスト及び前記第2のテキストをベクトルにそれぞれマッピングした後に、前記ベクトルに対して次元削減を行って低次元密ベクトルを用

いて前記第 1 のテキスト及び前記第 2 のテキストを表すように、前記 1 つ以上のプロセッサを制御するためのコンピュータ実行可能命令をさらに記憶してもよい。

【0025】

或いは、次元削減は、Word 2 Vec、Sentence 2 Vec 及び Doc 2 Vec からなる群から選択される 1 つ以上の方法を用いて行われてもよい。

【0026】

或いは、前記メモリは、前記第 1 のテキスト及び前記第 2 のテキストに基づいてセマンティックコンテンツ分析を行って前記類似部分及び前記相違部分を判断するように、前記 1 つ以上のプロセッサを制御するためのコンピュータ実行可能命令をさらに記憶してもよい。

【0027】

或いは、前記メモリは、前記第 1 のテキストと前記第 2 のテキストとの間でセマンティックマッチング処理を行って前記第 1 のテキストと前記第 2 のテキストとの間のセマンティックオーバーレイを取得するように、前記 1 つ以上のプロセッサを制御するためのコンピュータ実行可能命令をさらに記憶してもよい。

【0028】

或いは、前記メモリは、前記第 2 のテキストの単語のベクトルを用いて前記第 1 のテキストにおける単語のベクトルを再構成し、再構成の結果に基づいてセマンティックオーバーレイを算出するように、前記 1 つ以上のプロセッサを制御するためのコンピュータ実行可能命令をさらに記憶してもよい。

【0029】

或いは、前記メモリは、前記第 1 のテキスト及び前記第 2 のテキストに対してセマンティック分解を行って前記類似部分及び前記相違部分を判断するように、前記 1 つ以上のプロセッサを制御するためのコンピュータ実行可能命令をさらに記憶してもよい。

【0030】

或いは、前記セマンティックオーバーレイは、式 (1) に基づいて算出されてもよい。

【数 5】

$$\min(\left\|S_i - \sum_{j=1}^m \alpha_{i,j} \times T_j\right\|^2 + \lambda \sum_{j=1}^m \alpha_{i,j}^2) \quad (1)$$

ここで、 S_i は前記第 1 のテキストの $d \times 1$ 列ベクトルであり、 T_j は前記第 2 のテキストの $d \times 1$ 列ベクトルであり、 $\alpha_{i,j}$ はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数である。

【0031】

或いは、前記第 1 のテキストの前記類似部分は、

【数 6】

$$S'_i = \sum_{j=1}^m A \times T_j$$

に基づいて算出され、ここで、 S'_i は第 1 のテキストの類似部分を表し、 A は $\alpha_{i,j}$ の行列であり、 $S_i - S'_i$ は前記第 1 のテキストの前記相違部分を表してもよい。

【0032】

或いは、前記セマンティックオーバーレイは、式 (2) に基づいて算出されてもよい。

【数 7】

10

20

30

40

50

$$\min(\left\|T_i - \sum_{j=1}^m \beta_{i,j} \times S_j\right\|^2 + \lambda \sum_{j=1}^m \beta_{i,j}^2) \quad (2)$$

【0033】

ここで、 T_i は前記第2のテキストの $d \times 1$ 列ベクトルであり、 S_j は前記第1のテキストの $d \times 1$ 列ベクトルであり、 $\beta_{i,j}$ はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数である。

【0034】

或いは、前記第2のテキストの前記類似部分は、

【数8】

$$T_j' = \sum_{j=1}^m A \times S_j$$

に基づいて算出され、ここで、 T_j' は前記第2のテキストの前記類似部分を表し、 A は $\beta_{i,j}$ の行列であり、 $T_j - T_j'$ は前記第2のテキストの前記相違部分を表してもよい。

【0035】

或いは、前記装置は、前記第1のテキストと前記第2のテキストとの間の前記類似部分及び前記相違部分を入力として用いて、前記第1のテキストと前記第2のテキストとの間の関連性を算出するリカレントニューラルネットワークをさらに備えてもよい。

【0036】

或いは、前記入力は、前記第1のテキストの類似部分及び相違部分と、前記第2のテキストの類似部分及び相違部分とを含んでもよい。

【0037】

或いは、前記メモリは、 (S_s, T_s, L) で表されるサンプルデータを用いて前記リカレントニューラルネットワークをトレーニングするように、前記1つ以上のプロセッサを制御するためのコンピュータ実行可能命令をさらに記憶し、 S_s は第1のサンプルテキストを表し、 T_s は第2のサンプルテキストを表し、 L はサンプルの関連性を表してもよい。

【0038】

或いは、前記メモリは、第1の範囲にある第1の関連性に第1の粒度値を、第2の範囲にある第2の関連性に第2の粒度値を割り当てるように、前記1つ以上のプロセッサを制御するためのコンピュータ実行可能命令をさらに記憶してもよい。

【0039】

別の方面において、本発明は、本明細書に記載の関連性を算出する装置を備える、データクエリ装置を提供する。

【0040】

別の方面において、本発明は、本明細書に記載の関連性を算出する方法により、第1のテキストと複数の潜在的なターゲットテキストから選択される第2のテキストとの間の関連性を算出することと、前記複数の潜在的なターゲットテキストと前記第1のテキストとの間の関連性の算出に基づいて、前記複数の潜在的なターゲットテキストをランク付けすることと、前記複数の潜在的なターゲットテキストのランク付けの結果に基づいて、前記複数の潜在的なターゲットテキストからターゲットテキストを選択することを含む、データクエリ方法を提供する。

【0041】

別の方面において、本発明は、コンピュータ可読命令を記憶する非一時的なコンピュータ可読記録媒体であって、前記コンピュータ可読命令は、プロセッサによって実行可能であり、前記プロセッサに、第1のテキスト及び第2のテキストをベクトルにそれぞれマッ

10

20

30

40

50

ピングさせ、前記ベクトルに基づいて前記第1のテキストと前記第2のテキストとの間の類似部分及び相違部分を判断させ、前記類似部分及び前記相違部分の両方を用いて第1のテキストと第2のテキストとの間の関連性を算出させる、非一時的なコンピュータ可読記録媒体を提供する。

【0042】

以下の図面は開示する様々な実施形態による例示を目的とした例にすぎず、本発明の範囲を限定するものではない。

【図面の簡単な説明】

【0043】

【図1】本開示のいくつかの実施形態における関連性を算出する方法を示すフローチャートである。

10

【図2】本開示のいくつかの実施形態における関連性を算出する装置の構造を示す模式図である。

【図3】本開示のいくつかの実施形態における関連性を算出する装置の構造を示す模式図である。

【図4】本開示のいくつかの実施形態におけるデータクエリ装置の構造を示す模式図である。

【図5】本開示のいくつかの実施形態における電子装置を示す模式図である。

【発明を実施するための形態】

【0044】

20

以下では、実施形態を参照しつつ、本開示について具体的に説明する。なお、いくつかの実施形態に関する以下の説明は例示及び説明としてのものにすぎない。それは、網羅的であること、又は開示された正確な形態に限定されることを意図するものではない。

【0045】

関連性を算出する従来の方法では、通常、テキスト間の類似度が考慮されるが、テキスト間の相違については考慮されない。本開示は、テキスト間の相違も、関連性算出処理における正確な情報マッチングを容易にする重要なセマンティック情報を提供することを初めて見出した。テキスト間の相違も考慮することで、関連性を算出する際により優れた正確な結果が得られ、ユーザのクエリに応答するにあたってより有用な情報がレンダリングされる。

30

【0046】

そこで、本開示は、特に、関連技術における制限及び欠点に起因する1つ以上の課題を実質的に解消する、関連性を算出する方法、関連性を算出する装置、データクエリ装置及び電子装置を提供する。1つの方面において、本開示は関連性を算出する方法を提供する。いくつかの実施形態において、この方法は、第1のテキスト及び第2のテキストをベクトルにそれぞれマッピングすることと、ベクトルに基づいて第1のテキストと第2のテキストとの間の類似部分及び相違部分を判断することと、類似部分及び相違部分の両方を用いて第1のテキストと第2のテキストとの間の関連性を算出することを含む。1つの実施例において、この方法は、第1のテキストをベクトルの第1のセットにマッピングすることと、第2のテキストをベクトルの第2のセットにマッピングすることを含む。或いは、類似部分及び相違部分を判断するステップは、第1のテキストの類似部分及び相違部分を判断することと、第2のテキストの類似部分及び相違部分を判断することを含んでもよい。

40

【0047】

本明細書において「類似部分」という用語は、類似尺度がしきい値以上であるテキストの部分の指し、「相違部分」という用語は、類似尺度がしきい値未満であるテキストの部分の指す。或いは、第1のテキストの類似部分とは、第2のテキストにおいて同一の出現があるか、又は第1のテキストと第2のテキストとの間でセマンティックオーバーレイがある第1のテキストの部分（例えば、単語、文、文書）を指し、第2のテキストの類似部分とは、第1のテキストにおいて同一の出現があるか、又は第1のテキストと第2のテキ

50

ストとの間でセマンティックオーバーレイがある第2のテキストの部分（例えば、単語、文、文書）を指してもよい。或いは、第1のテキストの相違部分とは、第2のテキストにおいて同一の出現がなく、第1のテキストと第2のテキストとの間でセマンティックオーバーレイのない第1のテキストの部分（例えば、単語、文、文書）を指し、第2のテキストの相違部分とは、第1のテキストにおいて同一の出現がなく、第1のテキストと第2のテキストとの間でセマンティックオーバーレイのない第2のテキストの部分（例えば、単語、文、文書）を指してもよい。

【0048】

本明細書において「関連性」という用語は、2つのテキスト間の関連性の程度、例えば、テキストが別のテキスト（例えば、クエリテキスト）に対して如何に適切であるか、関連されているか、又は適用可能であるかの尺度を指す。或いは、関連性の値は、粒度により表してもよく、連続関係（例えば、0、0.5又は1の関係）がなくてもよい。或いは、関連性は、関連性アルゴリズムにより算出してもよい。或いは、関連性アルゴリズムは、サンプルデータを用いてトレーニングされて、関連性スコアを算出してもよい。或いは、関連性は、複数の基準に基づいて、例えば、テキストの類似部分及び相違部分の両方を考慮して判断してもよい。

【0049】

1つの実施例において、ユーザは「風邪の紹介」というクエリを入力する。データベースでは、3つの結果が見つかる。テキスト1：この記事は、副鼻腔炎について紹介するものである。テキスト2：この記事は、通常の風邪ではなく、インフルエンザについて紹介するものである。テキスト3：この記事は、肺炎ではなく、インフルエンザについて紹介するものである。1つの実施例において、テキスト間の類似度を表すのに関数 $sim(.,.)$ が用いられる。ユーザが入力したクエリテキストとデータベースで見つかった3つのテキストとの間の類似度が算出されると、 $sim(“クエリテキスト”, “テキスト2”) > sim(“クエリテキスト”, “テキスト1”)$ と $sim(“クエリテキスト”, “テキスト3”) > sim(“クエリテキスト”, “テキスト1”)$ ということが比較的簡単に判断される。しかし、 $sim(“クエリテキスト”, “テキスト2”)$ と $sim(“クエリテキスト”, “テキスト3”)$ とを比較するのは相対的により難しい。しかし、テキスト間の相違も考慮すれば、 $sim(“クエリテキスト”, “テキスト2”)$ と $sim(“クエリテキスト”, “テキスト3”)$ との比較を行うことができる。例えば、テキスト2は、通常の風邪と共通しない、例えば、インフルエンザ独特の特徴といったインフルエンザと通常の風邪の違いを一層強調する。一方で、テキスト3はインフルエンザと肺炎の違いを強調する。インフルエンザと肺炎に共通する特徴は少ないため、テキスト3は風邪の共通の特徴に関する情報をより多く含む場合がある。したがって、テキスト2と比較して、テキスト3はクエリテキストにより関連しているということになる。

【0050】

別の実施例において、ユーザは（例えば、種 i 及び種 ii 等の複数の種のある）属 A に関する情報を求めるクエリを入力する。データベースでは、3つの結果が見つかる。テキスト1は、密接に関連する属 B 又は密接に関連する属 B の種に関する情報を含む。テキスト2は、種 i と種 ii との間の違いに関する情報を含む。テキスト3は、種 i と関連する属 C との間の違いに関する情報を含む。1つの実施例において、テキスト間の類似度を表すのに関数 $sim(.,.)$ が用いられる。ユーザが入力したクエリテキストとデータベースで見つかった3つのテキストとの間の類似度が算出されると、 $sim(“クエリテキスト”, “テキスト2”) > sim(“クエリテキスト”, “テキスト1”)$ と $sim(“クエリテキスト”, “テキスト3”) > sim(“クエリテキスト”, “テキスト1”)$ ということが比較的簡単に判断される。しかし、 $sim(“クエリテキスト”, “テキスト2”)$ と $sim(“クエリテキスト”, “テキスト3”)$ とを比較するのは相対的により難しい。しかし、テキスト間の相違も考慮すれば、 $sim(“クエリテキスト”, “テキスト2”)$ と $sim(“クエリテキスト”, “テキスト3”)$ との比較を行うことができる。例えば、テキスト2は、種 ii と共通しない、例えば、種 i 独特の特徴といった種 i と種 ii の違いを一層強調する。一方で

、テキスト3は種iと関連する属Cの違いを強調する。種iと関連する属Cに共通する特徴は少ないため、テキスト3は属Aの共通の特徴に関する情報をより多く含む場合がある。したがって、テキスト2と比較して、テキスト3はクエリテキストにより関連しているということになる。

【0051】

さらに別の実施例において、ユーザは（例えば、癌性神経疼痛や癌性骨疼痛等の様々な種類のある）「癌性疼痛」に関する情報を求めるクエリを入力する。データベースでは、3つの結果が見つかる。テキスト1は、偏頭痛に関する情報を含む。テキスト2は、癌性神経疼痛と癌性骨疼痛の違いに関する情報を含む。テキスト3は、癌性骨疼痛と関節炎疼痛の見分け方に関する情報を含む。1つの実施例において、テキスト間の類似度を表すのに関数 $\text{sim}(\cdot, \cdot)$ が用いられる。ユーザが入力したクエリテキストとデータベースで見つかった3つのテキストとの間の類似度が算出されると、 $\text{sim}(\text{“クエリテキスト”}, \text{“テキスト2”}) > \text{sim}(\text{“クエリテキスト”}, \text{“テキスト1”})$ と $\text{sim}(\text{“クエリテキスト”}, \text{“テキスト3”}) > \text{sim}(\text{“クエリテキスト”}, \text{“テキスト1”})$ ということが比較的簡単に判断される。しかし、 $\text{sim}(\text{“クエリテキスト”}, \text{“テキスト2”})$ と $\text{sim}(\text{“クエリテキスト”}, \text{“テキスト3”})$ とを比較するのは相対的により難しい。しかし、テキスト間の相違も考慮すれば、 $\text{sim}(\text{“クエリテキスト”}, \text{“テキスト2”})$ と $\text{sim}(\text{“クエリテキスト”}, \text{“テキスト3”})$ との比較を行うことができる。例えば、テキスト2は、一般に癌性骨疼痛と共通しない、例えば、癌性神経疼痛独特の特徴といった癌性神経疼痛と癌性骨疼痛の違いを強調する。一方で、テキスト3は癌性骨疼痛と関節炎疼痛の違いを強調する。癌性骨疼痛及び関節炎疼痛は、2つの全く異なる病的状態によって引き起こされるため、テキスト3は、癌性疼痛に共通する特徴についての情報をより多く含む場合がある。したがって、テキスト2と比較して、テキスト3はクエリテキストにより関連しているということになる。

【0052】

図1は、本開示のいくつかの実施形態における関連性を算出する方法を示すフローチャートである。図1を参照すると、いくつかの実施形態において、関連性を算出する方法は、少なくとも第1のテキストSを取得し、第2のテキストTを取得するステップと、第1のテキストS及び第2のテキストTをベクトルにそれぞれマッピングするステップと、第1のテキストSと第2のテキストTとの間の類似部分及び相違部分を判断するステップと、類似部分及び相違部分を用いて第1のテキストSと第2のテキストTとの間の関連性を算出するステップとを含む。或いは、類似部分及び相違部分を判断するステップは、第1のテキストの類似部分及び相違部分を判断することと、第2のテキストの類似部分及び相違部分を判断することを含んでもよい。

【0053】

いくつかの実施形態において、第1のテキストS又は第2のテキストTは、複数の単語を含む文である。或いは、第1のテキストS又は第2のテキストTは、複数の文を含む段落であってもよい。或いは、第1のテキストS又は第2のテキストTは、複数の段落を含む文書であってもよい。或いは、第1のテキストS又は第2のテキストTは、膨大な数の文書を含んでもよい。

【0054】

いくつかの実施形態において、第1のテキストS及び第2のテキストTは、単一のデータソースから取得される。いくつかの実施形態において、第1のテキストS及び第2のテキストTは、複数のデータソースから取得される。或いは、第1のテキストS及び第2のテキストTのソースには、ローカルデバイス（例えば、携帯電話、パッド及びコンピュータ）に記憶された、又はローカルデバイスにより出力されたテキストが含まれてもよい。或いは、第1のテキスト及び第2のテキストのソースには、インターネットにより送信されたテキストを含んでもよい。

【0055】

いくつかの実施形態において、第1のテキストS及び第2のテキストTは、それぞれ別

々のベクトルにマッピングされる。例えば、第1のテキストSは、 $S_1, S_2, \dots, S_n$ にマッピングされる。第2のテキストTは、 $T_1, T_2, \dots, T_m$ にマッピングされる。したがって、 $S_1 \sim S_n$ は第1のテキストSを構成する単語の配列である。例えば、 $S_1 \sim S_n$ は第1のテキストSに対応するベクトルである。1つの実施例において、 $S = \text{“風邪の症状”}$ 、 $S_1 = \text{“症状”}$ 、 $S_2 = \text{“の”}$ 、 $S_3 = \text{“風邪”}$ である。同様に、 $T_1 \sim T_m$ は第2のテキストTを構成する単語の配列であり、 $T_1 \sim T_m$ は第2のテキストTに対応するベクトルである。1つの実施例において、 $T = \text{“インフルエンザの症状”}$ 、 $T_1 = \text{“症状”}$ 、 $T_2 = \text{“の”}$ 、 $T_3 = \text{“インフルエンザ”}$ である。

【0056】

様々な適切な方法によりテキストをベクトル化してよい。テキストをベクトル化する適切な方法の例として、ベクトル空間モデル（VSM）及び単語の分散表現が挙げられる。潜在セマンティック解析（LSA）、確率的潜在セマンティック解析（PLSA）及び潜在的ディリクレ配分（LDA）を含む、様々な適切な算出モデルを用いて単語の分散表現によってテキストをベクトルにマッピングしてもよい。

10

【0057】

いくつかの実施形態において、この方法は、第1のテキストS又は第2のテキストTをベクトルにマッピングした後、ベクトルに対して次元削減を行って、低次元密ベクトルを用いてテキストを表すことをさらに含む。或いは、次元削減は、主成分分析（PCA）により行ってもよい。多変量データセットが高次元データ空間（1つの変数あたり1つの軸）内の座標のセットとして可視化される場合、PCAは、最も有益な視点から見たときのこの対象物の投影であるより低次元の画像をユーザに提供することができる。これは、変換されたデータの次元が低下するように、初めのいくつかの主成分のみを用いて行われる。したがって、PCAは多重分散を線形に変換させていくつかの主分散を取得できる。或いは、Word2Vec、Sentence2Vec及びDoc2Vec等の方法により次元削減を行ってもよい。

20

【0058】

いくつかの実施形態において、次元削減は単語埋め込み処理により行われる。単語埋め込み処理は、Word2Vec等の方法により行ってもよい。1つの実施例において、次元削減の後、第1のテキストSのベクトル $S_1 \sim S_n$ 及び第2のテキストTのベクトル $T_1 \sim T_m$ はそれぞれ2つのセットの低次元ベクトルにマッピングされる。次元数がdである場合、テキストの各単語は、 $d \times 1$ 列ベクトルに対応する。第1のテキストSは、 $d \times n$ 行列（つまり、行列P）に対応し、第2のテキストTは、 $d \times m$ 行列（つまり、行列Q）に対応する。

30

【0059】

いくつかの実施形態において、第1のテキストと第2のテキストとの間の類似部分及び相違部分が判断される。或いは、類似部分及び相違部分は、Simhash、Shingling及びMinhash等のテキスト符号化方式により判断されてもよい。1つの実施例において、Simhashは、第1のテキストと第2のテキストをバイナリ符号化するのに使用され、第1のテキストと第2のテキストとの間のハミング距離は、XOR演算を実行する（つまり、それら2つのテキストのSimhashの結果に対してXOR演算を実行することによって判断される。XOR演算の結果において、結果「1」の数が3を超える場合、2つのテキストは異なる（例えば、相違部分）であると判断され、結果「1」の数が3以下である場合、2つのテキストは類似する（例えば、類似部分である）と判断される。

40

【0060】

いくつかの実施形態において、第1のテキストSと第2のテキストTとの間の類似部分及び相違部分を判断するステップは、第1のテキストSと第2のテキストTに基づいてセマンティックコンテンツ分析を行うことを含む。或いは、セマンティックコンテンツ分析は、第1のテキストと第2のテキストとの間でセマンティックマッチング処理を行って、第1のテキストSと第2のテキストTとの間のセマンティックオーバーレイを取得するこ

50

とを含んでもよい。

【0061】

いくつかの実施形態において、第1のテキストSと第2のテキストTとの間の類似部分を算出するために、S（T）における各単語が、T（S）における単語により意味的にオーバーレイ可能か否かを知る必要がある。例えば、クエリテキストにおける「風邪」という単語は、テキスト3における「インフルエンザ」という単語を意味的にオーバーレイできるが、テキスト1における「副鼻腔炎」という用語をオーバーレイすることはほぼできない。このため、 $\text{sim}(\text{“クエリテキスト”}, \text{“テキスト3”}) > \text{sim}(\text{“クエリテキスト”}, \text{“テキスト1”})$ である。

【0062】

いくつかの実施形態において、第1のテキストSと第2のテキストTとの間のセマンティックオーバーレイを判断するステップは、第1のテキストSにおける単語（例えば、単語 S_i ）のベクトルを第2のテキストTの単語のベクトルを用いて再構成することと、再構成の結果に基づいてセマンティックオーバーレイを算出することを含む。例えば、式（1）を用いて、第2のテキストTに対する第1のテキストSの単語のセマンティックオーバーレイを算出してもよい。

【0063】

【数9】

$$\min \left(\left\| S_i - \sum_{j=1}^m \alpha_{i,j} \times T_j \right\|^2 + \lambda \sum_{j=1}^m \alpha_{i,j}^2 \right) \quad (1)$$

【0064】

ここで、 S_i 及び T_j は各々 $d \times 1$ 列ベクトルであり、 S_i 及び T_j は各々単語埋め込み処理で得られた単語のベクトルであり、 $\alpha_{i,j}$ はパラメータであり、 $\lambda > 0$ であり、 λ は正の実数である。或いは、 S_i は第1のテキストの $d \times 1$ 列ベクトルであり、 T_j は第2のテキストの $d \times 1$ 列ベクトルであり、 $\alpha_{i,j}$ はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数であってもよい。

【0065】

例えば、パラメータ $\alpha_{i,j}$ は、以下の方法により解くことができる。

【0066】

まず、

【数10】

$$f_i = \left\| S_i - \sum_{j=1}^m \alpha_{i,j} \times T_j \right\|^2 + \lambda \sum_{j=1}^m \alpha_{i,j}^2$$

を満たす凸関数 f_i を設定する。 f_i の最小値を得るために、

【数11】

$$\alpha_{i,j} \text{ は } \frac{\partial f_i}{\partial \alpha_{i,j}} = 0$$

という条件を満たさなければならない。

【0067】

こうして、以下が導かれる。

【数12】

10

20

30

40

50

$$\begin{aligned}
\frac{\partial f_i}{\partial \alpha_{i,j}} &= -2(S_i - \sum_{k=1}^m \alpha_{i,k} T_k)^T T_j + 2\lambda \alpha_{i,j} \\
&= -2 \left[S_i^T T_j - \sum_{k=1}^m \alpha_{i,k} T_k^T T_j - \lambda \alpha_{i,j} \right] = 0 \quad (2) \\
\Rightarrow S_i^T T_j - \sum_{k=1}^m \alpha_{i,k} T_k^T T_j - \lambda \alpha_{i,j} &= 0
\end{aligned}$$

10

【0068】

ここで、上付きのTは、元の行列の転置を表す。上記の式は、 $i = 1, 2, \dots, n$ のとき、及び $j = 1, 2 \dots m$ のときに、等しく適用可能であるため、合わせて $n \times m$ 個の方程式を生成して、 $n \times m$ 個のパラメータの値を解くことができる。

【0069】

$n \times m$ 個のパラメータの値を解くプロセスを説明するために、

【数13】

$$S_i^T T_j - \sum_{k=1}^m \alpha_{i,k} T_k^T T_j - \lambda \alpha_{i,j} = 0$$

式

20

は、行列形式でより良く表すことができる。まず、 $n \times m$ 行列Rが提供され、 i 番目の行及び j 番目の列の要素 $R_{i,j}$ は、 $R_{i,j} = S_i^T T_j$ である。行列Rは、 $R = P^T Q$ として表すことができ、ここで、P及びQは、単語の埋め込み後に構成された行列に対応する。次に、 $m \times m$ 行列Uが提供され、 k 番目の行及び j 番目の列の要素 $U_{k,j}$ は、 $U_{k,j} = T_k^T T_j$ である。行列Uは、 $U = Q^T Q$ として表すことができ、ここで、Qは、単語の埋め込み後に構成された行列に対応する。第3に、 $n \times m$ 行列Aが提供され、 i 番目の行及び j 番目の列の要素は、 $\alpha_{i,j}$ である。第4に、行列R、U及びAを用いて、

【数14】

30

$$S_i^T T_j - \sum_{k=1}^m \alpha_{i,k} T_k^T T_j - \lambda \alpha_{i,j} = 0$$

式

を

【数15】

$$\begin{aligned}
R - AU - \lambda A &= 0 \\
\Rightarrow R &= AU + \lambda A \\
&= A(U + \lambda I) \\
\Rightarrow A &= R(U + \lambda I)^{-1} \quad (3)
\end{aligned}$$

40

として表すことができ、ここで、Iは、 $m \times m$ の単位行列である。

【0070】

式(2)を解くプロセスにおいて、行列 $(U + \lambda I)$ は非特異行列でなければならず、即ち、行列 $(U + \lambda I)$ の逆行列が存在する。 $U_{k,j} = T_k^T T_j$ であるため、Uは、 $m \times m$ の実対称行列である。行列Uに類似する対角行列 Λ があり、即ち、 $V^{-1} U V = \Lambda$ で

50

ある。こうして、

【数 1 6】

$$V^{-1}(U + \lambda I)V = \Lambda + \lambda I$$

が導かれる。 λ が行列 U の負の固有値に等しくない限り、行列 $(U + \lambda I)$ の固有値（即ち、 $\Lambda + \lambda I$ の対角線に沿ったの要素）はゼロとはならず、行列 $(U + \lambda I)$ は非特異行列である。行列 U の固有値は、実数フィールドにおける任意の実数であることができる。行列 U の分布関数を $F(x)$ とすると、 x の値は連続するため、任意の設定された実数 a について、 a の確率値 $F(a) = 0$ である。したがって、先験的な知識なしに、 λ の値が行列 U の負の固有値と等しいと事前に判断される確率は実質的にゼロであり、即ち、行列 $(U + \lambda I)$ の逆行列が存在する。

【0 0 7 1】

同様に、同じ方法により、第 1 のテキスト S に対する第 2 のテキスト T の単語のセマンティックオーバーレイも取得することができる。或いは、式 (3) を用いて第 1 のテキスト S に対する第 2 のテキスト T の単語のセマンティックオーバーレイを算出してもよい。

【0 0 7 2】

【数 1 7】

$$\min(\left\|T_i - \sum_{j=1}^m \beta_{ij} \times S_j\right\|^2 + \lambda \sum_{j=1}^m \beta_{ij}^2) \quad (3)$$

ここで、 T_i 及び S_j は各々 $d \times 1$ 列ベクトルであり、 T_i 及び S_j は各々単語埋め込み処理で得られた単語のベクトルであり、 β_{ij} はパラメータであり、 $\lambda > 0$ であり、 λ は正の実数である。或いは、 T_i は第 1 のテキストの $d \times 1$ 列ベクトルであり、 S_j は第 2 のテキストの $d \times 1$ 列ベクトルであり、 β_{ij} はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数であってもよい。

【0 0 7 3】

様々な代替方法を用いて上式 (1) 及び式 (3) を解くことができる。適切な方法の例として、数値的方法及び代数的方法が挙げられる。

【0 0 7 4】

いくつかの実施形態において、セマンティック分解を用いて第 1 のテキスト S と第 2 のテキスト T との間の類似部分及び相違部分（例えば、第 1 のテキスト S を第 2 のテキスト T と比較する場合は、第 1 のテキスト S の類似部分及び相違部分、第 2 のテキスト T を第 1 のテキスト S と比較する場合は、第 2 のテキスト T の類似部分及び相違部分）を取得することができる。

【0 0 7 5】

いくつかの実施形態において、行列 A （式 (2) を参照）は、第 1 のテキスト S における各単語が第 2 のテキスト T でどのように表されるか又はオーバーレイされるかに基づいており、ここで、行列 A の i 番目の行は、単語 S_i がテキスト T でどのように表されるか又はオーバーレイされるかを示す。

【数 1 8】

$$S_i' = \sum_{j=1}^m A \times T_j$$

とすると、 S_i' は第 1 のテキスト S の類似部分、例えば、第 2 のテキスト T でオーバーレイ可能な第 1 のテキスト S の単語 S_i の一部を表す。そして、 $S_i - S_i'$ は、第 1 のテキ

ストSの相違部分を表す。第1のテキストSは、類似部分（ $d \times n$ 行列により表され、図1の類似部分 S'_{+} を参照）と相違部分（ $d \times n$ 行列により表され、図2の相違部分 S'_{-} を参照）の2つの部分により表される。したがって、テキストは、 $d \times n$ 行列ではなく、 $d \times 2n$ 行列により表すことができる。この方法を採用することで、第1のテキストSと第2のテキストTを比較したとき、第1のテキストSの類似部分及び相違部分を算出することができる。

【0076】

同様に、第2のテキストTと第1のテキストSとの間の類似部分及び相違部分を算出することは、式

【数19】

$$T_j' = \sum_{j=1}^m A \times S_j$$

を採用することを含み、ここで、 A は $\beta_{i,j}$ の行列であり、 T_j' は第2のテキストTの類似部分を表し、 $T_j - T_j'$ は第2のテキストTの相違部分を表す。第2のテキストTは、類似部分（ $d \times m$ 行列により表され、図1の類似部分 T'_{+} を参照）と相違部分（ $d \times m$ 行列により表され、図2の相違部分 T'_{-} を参照）の2つの部分を含む。この方法を採用することで、第1のテキストSと第2のテキストTを比較したとき、第2のテキストTの類似度と相違度も算出することができる。

【0077】

いくつかの実施形態において、第1のテキストS及び第2のテキストT各々の類似部分及び相違部分の判断に基づいて、第1のテキストS及び第2のテキストTとの間の関連性を算出することができる。或いは、関連性を算出することは、テキスト符号化方式によりJaccard類似度を判断することを含んでもよい。或いは、テキストセマンティック算出により関連性を算出してもよい。或いは、機械学習により関連性を算出してもよい。

【0078】

いくつかの実施形態において、第1のテキストと第2のテキストとの間の関連性は、リカレントニューラルネットワーク（RNN）を用いて算出することができる。或いは、RNNを用いることは、RNNをトレーニングするのにサンプルデータを用いることを含んでもよい。1つの実施例において、サンプルデータは（ S_s, T_s, L ）として表され、ここで、 S_s と T_s はサンプルテキストを表し（例えば、 S_s は第1のサンプルテキストを表し、 T_s は第2のサンプルテキストを表す）、 L は関連性を表す。或いは、 S_s 及び T_s は、元々入力されていた第1のテキストS及び第2のテキストTであってもよい。或いは、 S_s 及び T_s は、第1のテキストS及び第2のテキストTと異なるサンプルテキストであってもよい。或いは、関連性の粒度を割当ててもよい。1つの実施例において、粒度は2、 $L \in \{0, 1\}$ であり、ここで、0は無関係であることを表し、1は関係のあることを表す。別の実施例において、粒度は3、 $L \in \{0, 0.5, 1\}$ であり、ここで、0は無関係であることを表し、0.5はいくらか関係のあることを表し、1は大いに関係のあることを表す。

【0079】

或いは、サンプルデータでRNNをトレーニングするために、RNNにサンプルデータが入力されてもよい。そして、第1のテキストSの類似部分と、第2のテキストTの類似部分と、第1のテキストSの相違部分と、第2のテキストTの相違部分に基づく複合的関連性が算出される。1つの実施例において、粒度が2に設定されると、第1の範囲における複合的関連性は粒度1を有するように定義され、第1の範囲と異なる第2の範囲における複合的関連性は粒度0を有するように定義される。別の実施例において、粒度が3に設定されると、第1の範囲における複合的関連性は粒度1を有するように定義され、第2の範囲における複合的関連性は粒度0.5を有するように定義され、第3の範囲における複合的関連性は粒度0を有するように定義され、第1の範囲、第2の範囲及び第3の範囲は

10

20

30

40

50

互いに異なる。RNNは、一旦サンプルデータを用いてトレーニングされれば、非サンプルデータにおける関連性を算出するのに用いることができる。第1のテキストSを表す $d \times 2n$ 行列と第2のテキストTを表す $d \times 2m$ 行列をトレーニングされたRNNに入力する（例えば、第1のテキストS及び第2のテキストTの類似部分及び相違部分を入力する）ことで、入力情報に基づいて関連性を算出するのにRNNを用いることができる。

【0080】

本願の関連性を算出する方法では、2つのテキスト間の類似部分及び相違部分両方の影響を考慮する。それ故、本願の方法は、より細かい粒度分類で関連性算出を可能にすることで、テキストマッチングの精度を向上させ、高精度の検索又はクエリ結果をレンダリングする。

【0081】

別の方面において、本開示は関連性を算出する装置をさらに提供する。図2及び図3は、本開示のいくつかの実施形態における関連性を算出する装置の構造を示す模式図である。図2を参照すると、関連性を算出する装置は、第1のテキスト及び第2のテキストを取得するように構成されたデータ取得部200と、第1のテキスト及び第2のテキストそれぞれをベクトルにマッピングするように構成されたマッピングエンジン部210と、ベクトルに基づいて第1のテキストと第2のテキストとの間の類似部分及び相違部分を判断し、類似部分及び相違部分に基づいて第1のテキストと第2のテキストとの間の関連性を算出するように構成された関連性算出部220とを備えている。或いは、関連性算出部220は、第1のテキストの類似部分及び相違部分を判断し、第2のテキストの類似部分及び相違部分を判断するように構成されてもよい。

【0082】

いくつかの実施形態において、図3を参照すると、マッピングエンジン部210は、第1のテキスト及び第2のテキストにそれぞれ対応するベクトルの次元を削減するように構成された次元削減処理部211を備えている。

【0083】

いくつかの実施形態において、図3を参照すると、関連性算出部220は、第1のテキストと第2のテキストとの間のセマンティックマッチングを行うように構成されたセマンティックマッチング処理部221と、第1のテキストと第2のテキストとの間の類似部分及び相違部分を判断するように構成されたセマンティック分解処理部222と、類似部分及び相違部分に基づいて第1のテキストと第2のテキストとの間の関連性を算出するように構成された算出部223とを備えている。或いは、セマンティック分解処理部222は、第1のテキストの類似部分及び相違部分を判断し、第2のテキストの類似部分及び相違部分を判断するように構成されてもよい。

【0084】

或いは、セマンティックマッチング処理部221は、第2のテキストのベクトルを用いて第1のテキストのベクトルを再構成することによって、例えば、第2のテキストの単語のベクトルを用いて第1のテキストにおける単語のベクトルを再構成することによって、第1のテキストと第2のテキストとの間のセマンティックオーバーレイを判断するように構成されてもよい。

【0085】

いくつかの実施形態において、セマンティックマッチング処理部221は、

【数20】

$$\text{式(1)} \quad \min \left(\left\| S_i - \sum_{j=1}^m \alpha_{i,j} \times T_j \right\|^2 + \lambda \sum_{j=1}^m \alpha_{i,j}^2 \right)$$

を用いて第2のテキストTに対する第1のテキストSの単語のセマンティックオーバーレイを算出するように構成され、ここで、 S_i は第1のテキストの列ベクトル（例えば、 d

× 1 列ベクトル) であり、 T_j は第 2 のテキストの列ベクトル (例えば、 $d \times 1$ 列ベクトル) であり、 $\alpha_{i,j}$ はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数である。或いは、 S_i は第 1 のテキストの $d \times 1$ 列ベクトルであり、 T_j は第 2 のテキストの $d \times 1$ 列ベクトルであり、 $\alpha_{i,j}$ はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数であってもよい。

【0086】

いくつかの実施形態において、セマンティックマッチング処理部 221 は、

【数 21】

$$\min \left(\left\| T_i - \sum_{j=1}^m \beta_{i,j} \times S_j \right\|^2 + \lambda \sum_{j=1}^m \beta_{i,j}^2 \right)$$

式 (3)

10

を用いて第 1 のテキスト S に対する第 2 のテキスト T の単語のセマンティックオーバーレイを算出するように構成され、ここで、 T_i は第 2 のテキストの列ベクトル (例えば、 $d \times 1$ 列ベクトル) であり、 S_j は第 1 のテキストの列ベクトル (例えば、 $d \times 1$ 列ベクトル) であり、 $\beta_{i,j}$ はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数である。或いは、 T_i は第 1 のテキストの $d \times 1$ 列ベクトルであり、 S_j は第 2 のテキストの $d \times 1$ 列ベクトルであり、 $\beta_{i,j}$ はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数であってもよい。

20

【0087】

或いは、セマンティック分解処理部 222 は、

【数 22】

$$\text{式 } S'_i = \sum_{j=1}^m A \times T_j$$

30

に基づいて第 1 のテキストの類似部分及び相違部分を取得するように構成され、ここで、 A は $\alpha_{i,j}$ の行列であり、 S_i は第 1 のテキストの類似部分を表し、 $S_i - S_i'$ は第 2 のテキストの相違部分を表してもよい。

【0088】

或いは、セマンティック分解処理部 222 は、

【数 23】

$$\text{式 } T'_j = \sum_{i=1}^m A \times S_i$$

40

に基づいて第 2 のテキストの類似部分及び相違部分を取得するように構成され、ここで、 A は $\beta_{i,j}$ の行列であり、 T_j は第 1 のテキストの類似部分を表し、 $T_j - T_j'$ は第 2 のテキストの相違部分を表してもよい。

【0089】

いくつかの実施形態において、関連性を算出する装置はメモリと、1つ以上のプロセッサとを備えている。メモリと、1つ以上のプロセッサとは互いに接続されている。メモリは、第 1 のテキスト及び第 2 のテキストをベクトルにそれぞれマッピングし、ベクトルに基づいて第 1 のテキストと第 2 のテキストとの間の類似部分及び相違部分を判断し、類似部分及び相違部分の両方を用いて第 1 のテキストと第 2 のテキストとの間の関連性を算出するように、1つ以上のプロセッサを制御するためのコンピュータ実行可能命令を記憶す

50

る。或いは、メモリは、第1のテキスト及び第2のテキストをベクトルにそれぞれマッピングした後に、ベクトルに対して次元削減を行って低次元密ベクトルを用いてテキストを表すように、1つ以上のプロセッサを制御するためのコンピュータ実行可能命令をさらに記憶してもよい。或いは、次元削減は、Word2Vec、Sentence2Vec及びDoc2Vecからなる群より選択される1つ以上の方法を用いて行われてもよい。或いは、メモリは、第1のテキストの類似部分及び相違部分を判断し、第2のテキストの類似部分及び相違部分を判断するように、1つ以上のプロセッサを制御するためのコンピュータ実行可能命令を記憶してもよい。

【0090】

いくつかの実施形態において、メモリは、第1のテキスト及び第2のテキストに基づいてセマンティックコンテンツ分析を行って類似部分及び相違部分を判断するように、1つ以上のプロセッサを制御するためのコンピュータ実行可能命令をさらに記憶する。或いは、メモリは、第1のテキストと第2のテキストとの間でセマンティックマッチング処理を行って、第1のテキストと第2のテキストとの間のセマンティックオーバーレイを取得するように、1つ以上のプロセッサを制御するためのコンピュータ実行可能命令をさらに記憶してもよい。或いは、メモリは、第2のテキストの単語のベクトルを用いて第1のテキストSにおける単語のベクトルを再構成し、再構成の結果に基づいてセマンティックオーバーレイを算出するように、1つ以上のプロセッサを制御するためのコンピュータ実行可能命令をさらに記憶してもよい。或いは、第2のテキストTに対する第1のテキストSの単語のセマンティックオーバーレイは、式(1)に基づいて算出してもよい。

【0091】

【数24】

$$\min(\left\|S_i - \sum_{j=1}^m \alpha_{i,j} \times T_j\right\|^2 + \lambda \sum_{j=1}^m \alpha_{i,j}^2) \quad (1)$$

ここで、 S_i は第1のテキストの $d \times 1$ 列ベクトルであり、 T_j は第2のテキストの $d \times 1$ 列ベクトルであり、 $\alpha_{i,j}$ はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数であってもよい。或いは、 S_i は第1のテキストの $d \times 1$ 列ベクトルであり、 T_j は第2のテキストの $d \times 1$ 列ベクトルであり、 $\alpha_{i,j}$ はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数であってもよい。

【0092】

或いは、第1のテキストSに対する第2のテキストTの単語のセマンティックオーバーレイは、式(3)に基づいて算出してもよい。

【数25】

$$\min(\left\|T_i - \sum_{j=1}^m \beta_{i,j} \times S_j\right\|^2 + \lambda \sum_{j=1}^m \beta_{i,j}^2) \quad (3)$$

【0093】

ここで、 T_i は第2のテキストの列ベクトル（例えば、 $d \times 1$ 列ベクトル）であり、 S_j は第1のテキストの列ベクトル（例えば、 $d \times 1$ 列ベクトル）であり、 $\beta_{i,j}$ はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数である。或いは、 T_i は第1のテキストの $d \times 1$ 列ベクトルであり、 S_j は第2のテキストの $d \times 1$ 列ベクトルであり、 $\beta_{i,j}$ はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数であってもよい。

【0094】

10

20

30

40

50

或いは、第 1 のテキストの類似部分は、
【数 2 6】

$$S_i' = \sum_{j=1}^m A \times T_j$$

に基づいて算出され、 S_i' は第 1 のテキストの類似部分を表し、 A は $\alpha_{i,j}$ の行列であってもよい。或いは、 $S_i - S_i'$ は第 1 のテキストの相違部分を表してもよい。

【0095】

或いは、第 2 のテキストの類似部分は、
【数 2 7】

$$T_j' = \sum_{i=1}^m A \times S_i$$

に基づいて算出され、 T_j' は第 2 のテキストの類似部分を表し、 A は $\beta_{i,j}$ の行列であってもよい。或いは、 $T_j - T_j'$ は、第 2 のテキストの相違部分を表してもよい。

【0096】

いくつかの実施形態において、メモリは、第 1 のテキスト及び第 2 のテキストに対してセマンティック分解を行って類似部分及び相違部分を判断するように、1 つ以上のプロセッサを制御するためのコンピュータ実行可能命令をさらに記憶する。

【0097】

いくつかの実施形態において、この装置は、第 1 のテキストと第 2 のテキストとの間の類似部分及び相違部分を入力として用いて、第 1 のテキストと第 2 のテキストとの間の関連性を算出するリカレントニューラルネットワークをさらに備えている。或いは、このメモリは、 (S_s, T_s, L) で表されるサンプルデータを用いてリカレントニューラルネットワークをトレーニングするように、1 つ以上のプロセッサを制御するためのコンピュータ実行可能命令をさらに記憶し、ここで、 S_s は第 1 のサンプルテキストを表し、 T_s は第 2 のサンプルテキストを表し、 L はサンプルの関連性を表してもよい。或いは、 S_s 及び T_s は、元々入力されていた第 1 のテキスト S 及び第 2 のテキスト T であってもよい。 S_s 及び T_s は、第 1 のテキスト S 及び第 2 のテキスト T と異なるサンプルテキストであってもよい。或いは、メモリは、第 1 の範囲にある第 1 の関連性に第 1 の粒度値を、第 2 の範囲にある第 2 の関連性に第 2 の粒度値を割り当てるように、1 つ以上のプロセッサを制御するためのコンピュータ実行可能命令をさらに記憶してもよい。

【0098】

別の方面において、本開示はコンピュータ可読命令を記憶する非一時的なコンピュータ可読記録媒体をさらに提供する。いくつかの実施形態において、コンピュータ可読命令はプロセッサによって実行可能であり、本明細書に述べる方法の 1 つ以上のステップをプロセッサに実行させる。いくつかの実施形態において、コンピュータ可読命令はプロセッサによって実行可能であり、プロセッサに、第 1 のテキスト及び第 2 のテキストをベクトルにそれぞれマッピングさせ、ベクトルに基づいて第 1 のテキストと第 2 のテキストとの間の類似部分及び相違部分を判断させ、類似部分及び相違部分の両方を用いて第 1 のテキストと第 2 のテキストとの間の関連性を算出させる。或いは、コンピュータ可読命令はプロセッサによって実行可能であり、プロセッサに、第 1 のテキスト及び第 2 のテキストをベクトルにそれぞれマッピングした後に、さらにベクトルに対して次元削減を行って、低次元密ベクトルを用いてテキストを表すようにさせてもよい。或いは、次元削減は、Word2Vec、Sentence2Vec 及び Doc2Vec からなる群より選択される 1 つ以上の方法を用いて行われてもよい。或いは、コンピュータ可読命令はプロセッサによって実行可能であり、プロセッサに、第 1 のテキストの類似部分及び相違部分を判断させ、第 2 のテキストの類似部分及び相違部分を判断させてもよい。

10

20

30

40

50

【 0 0 9 9 】

いくつかの実施形態において、コンピュータ可読命令はプロセッサによって実行可能であり、プロセッサに、第 1 のテキスト及び第 2 のテキストに基づいてセマンティックコンテンツ分析を行って類似部分及び相違部分を判断させる。或いは、コンピュータ可読命令はプロセッサによって実行可能であり、プロセッサに、第 1 のテキストと第 2 のテキストとの間でセマンティックマッチング処理を行わせて、第 1 のテキストと第 2 のテキストとの間のセマンティックオーバーレイを取得させてもよい。或いは、コンピュータ可読命令はプロセッサによって実行可能であり、プロセッサに、第 2 のテキストの単語のベクトルを用いて第 1 のテキスト S における単語のベクトルを再構成させ、再構成の結果に基づいてセマンティックオーバーレイを算出させてもよい。或いは、第 2 のテキスト T に対する第 1 のテキスト S の単語のセマンティックオーバーレイは、式 (1) に基づいて算出してもよい。

10

【 0 1 0 0 】

【数 2 8】

$$\min(\left\|S_i - \sum_{j=1}^m \alpha_{i,j} \times T_j\right\|^2 + \lambda \sum_{j=1}^m \alpha_{i,j}^2) \quad (1)$$

20

【 0 1 0 1 】

ここで、 S_i は第 1 のテキストの $d \times 1$ 列ベクトルであり、 T_j は第 2 のテキストの $d \times 1$ 列ベクトルであり、 $\alpha_{i,j}$ はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数であってもよい。或いは、 S_i は第 1 のテキストの $d \times 1$ 列ベクトルであり、 T_j は第 2 のテキストの $d \times 1$ 列ベクトルであり、 $\alpha_{i,j}$ はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数であってもよい。

【 0 1 0 2 】

或いは、第 1 のテキスト S に対する第 2 のテキスト T の単語のセマンティックオーバーレイは、式 (3) に基づいて算出してもよい。

30

【 0 1 0 3 】

【数 2 9】

$$\min(\left\|T_i - \sum_{j=1}^m \beta_{i,j} \times S_j\right\|^2 + \lambda \sum_{j=1}^m \beta_{i,j}^2) \quad (3)$$

ここで、 T_i は第 2 のテキストの列ベクトル（例えば、 $d \times 1$ 列ベクトル）であり、 S_j は第 1 のテキストの列ベクトル（例えば、 $d \times 1$ 列ベクトル）であり、 $\beta_{i,j}$ はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数である。或いは、 T_i は第 1 のテキストの $d \times 1$ 列ベクトルであり、 S_j は第 2 のテキストの $d \times 1$ 列ベクトルであり、 $\beta_{i,j}$ はセマンティックオーバーレイを算出するためのパラメータであり、 $\lambda > 0$ であり、 λ は正の実数であってもよい。

40

【 0 1 0 4 】

或いは、第 1 のテキストの類似部分は、

【数 3 0】

$$S'_i = \sum_{j=1}^m A \times T_j$$

50

に基づいて算出され、 S_i は第 1 のテキストの類似部分を表し、 A は $\alpha_{i,j}$ の行列であってもよい。或いは、 $S_i - S_j$ は第 1 のテキストの相違部分を表してもよい。

【0105】

或いは、第 2 のテキストの類似部分は、

【数 3 1】

$$T_j' = \sum_{j=1}^m A \times S_j$$

に基づいて算出され、 T_j は第 2 のテキストの類似部分を表し、 A は $\beta_{i,j}$ の行列であってもよい。或いは、 $T_j - T_j'$ は、第 2 のテキストの相違部分を表してもよい。

【0106】

いくつかの実施形態において、コンピュータ可読命令はプロセッサによって実行可能であり、プロセッサに、第 1 のテキスト及び第 2 のテキストに対してセマンティック分解を行わせて類似部分及び相違部分を判断させる。

【0107】

別の方面において、本開示はデータクエリ装置をさらに提供する。いくつかの実施形態において、データクエリ装置は、本明細書に記載の関連性を算出する装置を備えている。図 4 は、本開示のいくつかの実施形態におけるデータクエリ装置の構造を示す模式図である。図 4 を参照すると、いくつかの実施形態において、データクエリ装置は、ヒューマンコンピュータインタラクション・インターフェース 401 を備えている。或いは、ヒューマンコンピュータインタラクション・インターフェース 401 は、患者の質問を受け、対応する回答を表示することのできる医療質問に回答するプラットフォームであってもよい。或いは、ヒューマンコンピュータインタラクション・インターフェース 401 は、ユーザが探しているコンテンツを受け取り、対応する文献を表示することのできる医学文献検索システムであってもよい。

【0108】

いくつかの実施形態において、データクエリ装置は、コンテンツマッチング処理部 402 と、データベース 403 とをさらに備えている。或いは、コンテンツマッチング処理部 402 は、本明細書に記載の関連性算出部 220 を備えてもよい。

【0109】

いくつかの実施形態において、ユーザがヒューマンコンピュータインタラクション・インターフェース 401 に検索したいコンテンツを入力すると、コンテンツマッチング処理部 402 はデータベース 403 から潜在的なターゲットコンテンツを収集する。或いは、コンテンツマッチング処理部 402 は、潜在的なターゲットコンテンツとユーザが入力したコンテンツとの間の関連性を算出するのに、本明細書に記載の関連性を算出する方法を用いてもよい。或いは、コンテンツマッチング処理部 402 は、関連性のランクに基づいて、潜在的なターゲットコンテンツからユーザがさがしているコンテンツを選択してもよい。

【0110】

別の方面において、本開示はデータクエリ方法をさらに提供する。いくつかの実施形態において、データクエリ方法は、本明細書に記載の関連性算出方法にしたがって第 1 のテキストと第 2 のテキストとの間の関連性を算出することを含む。第 2 のテキストは、複数の潜在的なターゲットテキストから選択されるテキストである。或いは、データクエリ方法は、複数の潜在的なターゲットテキストと第 1 のテキストとの間の関連性の算出に基づいて複数の潜在的なターゲットテキストをランク付けすることと、複数の潜在的なターゲットテキストのランク付けの結果に基づいて、複数の潜在的なターゲットテキストからターゲットテキストを選択することとをさらに含んでもよい。或いは、この方法は、データクエリの結果としてターゲットテキストをユーザに出力することをさらに含んでもよい。

【0111】

別の方面において、本開示は、電子装置 500 をさらに提供する。図 5 は、本開示のいくつかの実施形態における電子装置を示す模式図である。或いは、電子装置 500 は、プロセッサ 502 を備えてもよい。或いは、プロセッサ 502 は、本明細書に記載の関連性算出の少なくとも 1 つのステップを実施するコンピュータ命令を実行するように構成されてもよい。

【0112】

或いは、電子装置 500 は、少なくとも 1 つのプロセッサ 502 に接続されたメモリ 501 を少なくとも備えてもよい。或いは、メモリ 501 は、コンピュータ命令を記憶するように構成されてもよい。

【0113】

いくつかの実施形態において、電子装置 500 はローカルエンド装置であり、これは、電子装置 500 がユーザエンドで関連性を算出できることを意味する。いくつかの実施形態において、電子装置 500 は、ローカル及びリモートインタラクション装置であり、即ち、電子装置 500 は、ユーザエンド装置（つまり、ローカルエンド）において少なくとも第 1 のテキスト及び第 2 のテキストを取得し、リモートインターネットプロセッサ（つまり、リモートエンド）は第 1 のテキスト及び第 2 のテキストを受け取って類似度算出を行う。

【0114】

いくつかの実施形態において、電子装置 500 は、複数のユーザエンド装置と、ユーザエンド装置に接続されている複数のリモートインターネットプロセッサとを備えてもよい。或いは、複数のユーザエンド装置は、少なくとも第 1 のテキスト及び第 2 のテキストを含む情報をリモートインターネットプロセッサにアップロードしてもよい。或いは、アップロードされた情報は、ユーザが入力させた又は記憶させたクエリテキストであってもよい。或いは、リモートインターネットプロセッサは、各ユーザエンド装置によってアップロードされた情報を収集してもよい。或いは、リモートインターネットプロセッサは、続いて関連性算出を行ってもよい。

【0115】

いくつかの実施形態において、メモリ 501 は、任意のタイプの揮発性記憶デバイス或いは不揮発性記憶デバイス、又は揮発性記憶デバイスと不揮発性記憶デバイスの両方の組合せによって構成することができる。揮発性記憶デバイス又は不揮発性記憶デバイスのタイプには、スタティックランダムアクセスメモリ（SRAM）、電氣的消去可能プログラマブル読み取り専用メモリ（EEPROM）、消去可能プログラマブル読み取り専用メモリ（EPROM）、プログラマブル読み取り専用メモリ（PROM）、読み取り専用メモリ（ROM）、磁気メモリ、フラッシュメモリ、ディスク又は CD が含まれる。

【0116】

いくつかの実施形態において、プロセッサ 502 は、中央処理装置（CPU）、フィールドプログラマブルゲートアレイ（FPGA）、マイクロコントローラユニット（MCU）、デジタル信号プロセッサ（DSP）、特定用途向け集積回路（ASIC）、グラフィックス処理ユニット（GPU）等のデータ処理能力及び／又はプログラム実行能力を有する論理演算装置である。或いは、1 つ以上のプロセッサは、並列プロセッサと同時に関連性算出を実行するように構成されてもよい。或いは、1 つ以上のプロセッサは、関連性算出の一部を実行するように構成されてもよい。或いは、他のプロセッサは、関連性算出の残りの部分を実行するように構成されてもよい。

【0117】

いくつかの実施形態において、コンピュータ命令は、プロセッサに対応する命令のセットにより定義される少なくとも 1 つのプロセッサの動作を含んでもよい。或いは、コンピュータ命令は、少なくとも 1 つのコンピュータプログラムにより論理的に包含され、表すことができてもよい。

【0118】

いくつかの実施形態において、電子装置 500 は、ユーザインターフェース、キーボー

10

20

30

40

50

ド等の入力装置に接続することができる。或いは、電子装置500は、スピーカ等の出力装置に接続することができる。或いは、電子装置は、ユーザと電子装置500との間のインタラクションを実行する表示デバイス等の装置に接続することができる。

【0119】

いくつかの実施形態において、電子装置500は、無線ネットワーク、有線ネットワーク、及び／又は無線及び有線ネットワークの任意の組合せを含むネットワークを介して、様々な入力装置、出力装置又は他のインタラクション装置に接続することができる。或いは、ネットワークには、ローカルエリアネットワーク、インターネット、電気通信回路ネットワーク、インターネット及び／又は電気通信ネットワークに基づくモノのインターネット、並びに上記ネットワークの任意の組み合わせが含まれる。或いは、有線ネットワークは、ツイストペア、同軸ケーブル及びファイバを含む、異なる材料を用いて通信してもよい。或いは、無線ネットワークも、3G／4G／5G移動通信ネットワーク、Bluetooth（登録商標）、Zigbee及びWi-Fiを含む、異なる材料を用いて通信してもよい。

10

【0120】

本開示は、コンピュータ命令を記憶するように構成されたコンピュータ可読記録媒体をさらに提供する。コンピュータ命令が実行されると、関連性算出の少なくとも1つのステップが実施される。

【0121】

本発明の実施形態に関する以上の記述は、例示及び説明を目的とする。以上の説明は、網羅的であること、又は開示された正確な形態或いは例示的な実施形態に本発明を限定することを意図していない。それ故、上記記載は限定ではなく例示を目的としていると見なすべきであり、多くの変更や変形は当業者にとって明らかであろう。これらの実施形態は、本発明の原理及びその最良の態様の実際の適用を説明するために選択及び記載されたものであり、それによって、本発明が特定の用途又は想定される実施形態の様々な実施形態及び様々な変形例に適用可能であることを当業者に理解させることを目的としている。本発明の範囲は、本開示に付した請求項及びその均等物により定義することが意図され、別途示唆しない限り、すべての用語は合理的な範囲内で最も広く解釈される。したがって、「本発明」、「本開示」又はこれに類する用語は請求項を必ずしも特定の実施形態に限定せず、本発明の例示的な実施形態に対する参照は本発明への限定を示唆するものではなく、かかる限定を推論すべきではない。本発明は添付する請求項の精神と範囲によってのみ限定される。さらに、これらの請求項では後に名詞又は要素を伴って「第1の」、「第2の」等の表現を用いる場合がある。特定の数量が示されない限り、このような用語は専用語であると理解すべきであり、修飾された要素の数量が上記専用語により限定されると解釈してはならない。記載した効果や利点はいずれも本発明のすべての実施形態にあてはまるとは限らない。当業者であれば、以下の請求項により定義される本発明の範囲から逸脱せずに、記載した実施形態を変形できることが理解されよう。さらに、本開示の要素及び構成要素は、以下の請求項に明記されているか否かを問わず、いずれも公衆に捧げる意図はない。

20

30

【符号の説明】

40

【0122】

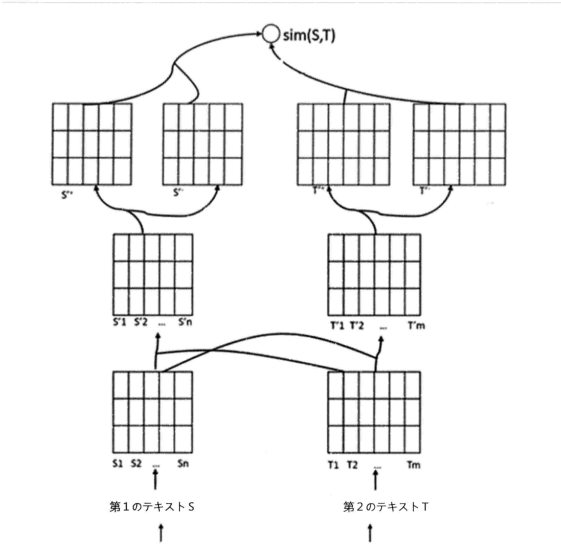
- 200 データ取得部
- 210 マッピングエンジン部
- 211 次元削減処理部
- 220 関連性算出部
- 221 セマンティックマッチング処理部
- 222 セマンティック分解処理部
- 223 算出部
- 401 ヒューマンコンピュータインタラクション・インターフェース
- 402 コンテンツマッチング処理部

50

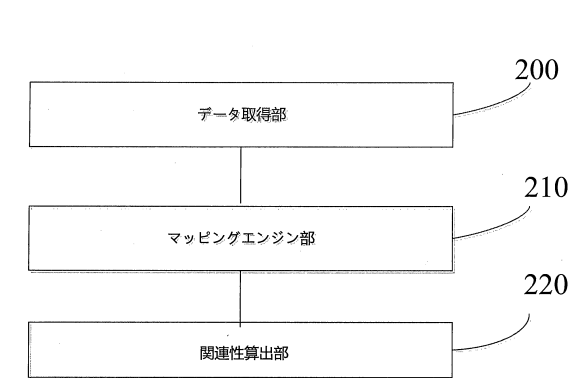
4 0 3 データベース
5 0 0 電子装置
5 0 1 メモリ
5 0 2 プロセッサ

【図面】

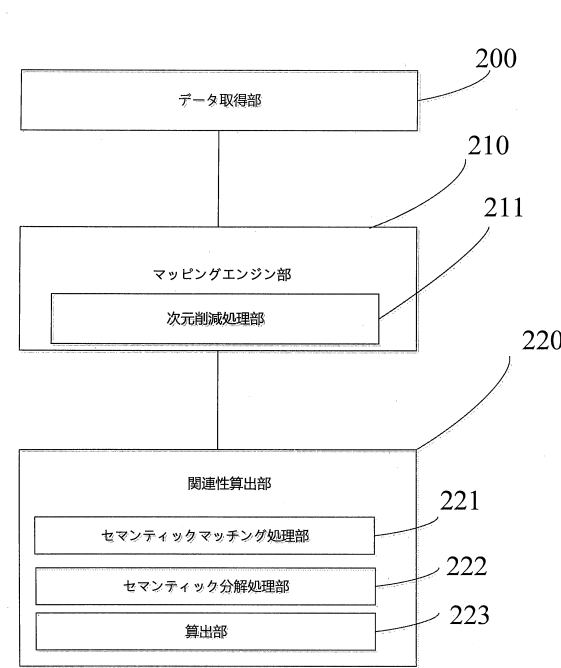
【図 1】



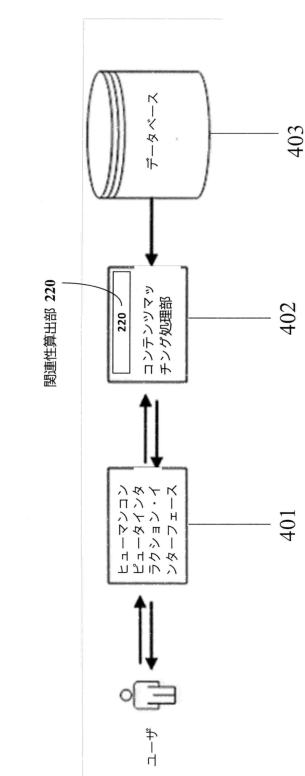
【図 2】



【図 3】



【図 4】



10

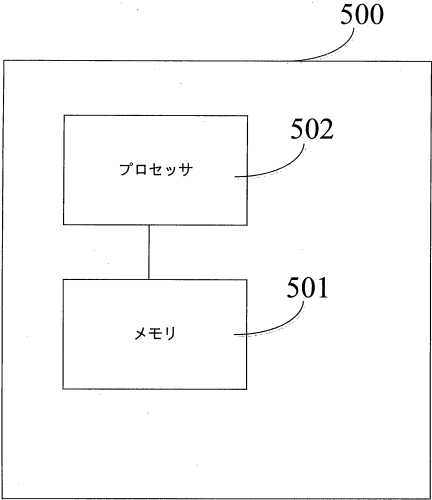
20

30

40

50

【図 5】



10

20

30

40

50

フロントページの続き

中国(CN)
中華人民共和国・１００１７６・ 베이ジン・ビーディーイー・ディゼ・ロード・ナンバー・９

合議体
審判長 林 毅
審判官 脇岡 剛
審判官 大塚 俊範

(56)参考文献 米国特許出願公開第２０１５／０１２０７２０（ＵＳ，Ａ１）
米国特許出願公開第２０１３／００５４６１２（ＵＳ，Ａ１）
特開２０１７－０２７１０６（ＪＰ，Ａ）
特開２０１７－０１０２４９（ＪＰ，Ａ）

(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １６／００－１６／９５８
Ｇ０６Ｆ ４０／２０－４０／５８