(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5869948号
(P5869948)
(45)発行日　平成28年2月24日(2016.2.24)　(24)登録日　平成28年1月15日(2016.1.15)

(51)Int.Cl.　F I
G06F 17/27 (2006.01)　G06F 17/27 665

請求項の数 14 (全 14 頁)

(21)出願番号　特願2012-95344(P2012-95344)
(22)出願日　平成24年4月19日(2012.4.19)
(65)公開番号　特開2013-222418(P2013-222418A)
(43)公開日　平成25年10月28日(2013.10.28)
審査請求日　平成27年1月8日(2015.1.8)

(73)特許権者　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　110001689
青稜特許業務法人
(74)代理人　110000350
ポレール特許業務法人
(72)発明者　柿下　容弓
東京都国分寺市東恋ヶ窪一丁目２８０番地
株式会社日立製作所　中央研究所内
(72)発明者　服部　英春
東京都国分寺市東恋ヶ窪一丁目２８０番地
株式会社日立製作所　中央研究所内

最終頁に続く

(54)【発明の名称】パッセージ分割方法、装置、及びプログラム

(57)【特許請求の範囲】
【請求項１】
処理部により、ドキュメントをパッセージに分割するパッセージ分割方法であって、
前記処理部は、
前記ドキュメントを文単位に分割し、
分割した前記文をクエリとして、予め記憶されている複数のドキュメントから、関連するドキュメントを抽出して、特徴量を作成し、
作成した前記特徴量の内の二つの特徴量の類似度が所定の閾値以上である、当該二つの特徴量の共通要素を用いて特徴量を更新し、
前記特徴量の類似度が所定の閾値以上である、当該二つの特徴量に対応する前記文、あるいはパッセージ候補を連結して新たなパッセージ候補とし、
前記特徴量の更新は前記新たなパッセージ候補の特徴量を得るものである
ことを特徴とするパッセージ分割方法。
【請求項２】
請求項１に記載のパッセージ分割方法であって、
前記処理部は、
前記特徴量として、ドキュメントベクトルを用いる、
ことを特徴とするパッセージ分割方法。
【請求項３】
請求項２に記載のパッセージ分割方法であって、

前記処理部は、
前記二つの特徴量である、二つのドキュメントベクトルＶ$_{i}$，Ｖ$_{j}$の類似度が所定の閾値以上である場合、二つの前記ドキュメントベクトルＶ$_{i}$，Ｖ$_{j}$の共通要素Ｖ$_{ij}$を抜き出し、検索クエリを生成し、
生成した前記検索クエリを用いて、新たなドキュメントベクトルＶ’$_{ij}$を得る、
ことを特徴とするパッセージ分割方法。
【請求項４】
請求項３に記載のパッセージ分割方法であって、
前記処理部は、
前記新たなドキュメントベクトルＶ’$_{ij}$が、前記共通要素Ｖ$_{ij}$の要素を含む度合いに対応して、前記新たなドキュメントベクトルＶ’$_{ij}$のベクトルサイズを修正する、
ことを特徴とするパッセージ分割方法。
【請求項５】
請求項３に記載のパッセージ分割方法であって、
前記処理部は、
前記新たなドキュメントベクトルＶ’$_{ij}$に対応する前記文、あるいはパッセージ候補として、二つの前記ドキュメントベクトルＶ$_{i}$，Ｖ$_{j}$に対応する前記文、あるいはパッセージ候補を連結して、新たなパッセージ候補とする、
ことを特徴とするパッセージ分割方法。
【請求項６】
請求項１に記載のパッセージ分割方法であって、
前記処理部は、
前記特徴量として、単語ベクトルを用いる、
ことを特徴とするパッセージ分割方法。
【請求項７】
請求項６に記載のパッセージ分割方法であって、
前記二つの特徴量である、二つの単語ベクトルＶ$_{i}$，Ｖ$_{j}$の類似度が所定の閾値以上である場合、二つの前記単語ベクトルＶ$_{i}$，Ｖ$_{j}$の共通要素Ｖ$_{ij}$を抜き出し、検索クエリを生成し、
生成した前記検索クエリを用いて、新たな単語ベクトルＶ’$_{ij}$を得る、
ことを特徴とするパッセージ分割方法。
【請求項８】
請求項７に記載のパッセージ分割方法であって、
前記処理部は、
前記新たな単語ベクトルＶ’$_{ij}$が、前記共通要素Ｖ$_{ij}$の要素を含む度合いに対応して、前記新たな単語ベクトルＶ’$_{ij}$のベクトルサイズを修正する、
ことを特徴とするパッセージ分割方法。
【請求項９】
請求項８に記載のパッセージ分割方法であって、
前記処理部は、
前記新たな単語ベクトルＶ’$_{ij}$に対応する前記文、あるいはパッセージ候補として、二つの前記単語ベクトルＶ$_{i}$，Ｖ$_{j}$に対応する前記文、あるいはパッセージ候補を連結して、新たなパッセージ候補とする、
ことを特徴とするパッセージ分割方法。
【請求項１０】
入力されるドキュメントをパッセージに分割するパッセージ分割装置であって、
処理部と記憶部とを備え、
前記処理部は、
前記ドキュメントを文単位に分割し、
分割した前記文をクエリとして、予め前記記憶部に記憶されている複数のドキュメント

から、関連するドキュメントを抽出して、特徴量を作成し、
作成した前記特徴量の内の二つの類似度が所定の閾値以上である、当該特徴量の共通要素を用いて特徴量を更新し、
前記特徴量の類似度が所定の閾値以上である、当該二つの特徴量に対応する文またはパッセージ候補を連結して新たなパッセージ候補とし、
前記特徴量の更新は前記新たなパッセージ候補の特徴量を得るものである
ことを特徴とするパッセージ分割装置。
【請求項１１】
請求項１０に記載のパッセージ分割装置であって、
前記処理部は、
前記特徴量として、関連する前記ドキュメントに基づく、ドキュメントベクトルあるいは単語ベクトルを用いる、
ことを特徴とするパッセージ分割装置。
【請求項１２】
請求項１１に記載のパッセージ分割装置であって、
前記処理部は、
前記二つの特徴量である、二つのドキュメントベクトル、或いは単語ベクトルV_i，V_jの類似度が所定の閾値以上である場合、二つの前記ドキュメントベクトル、或いは単語ベクトルV_i，V_jの共通要素V_{ij}を抜き出し、検索クエリを生成し、
生成した前記検索クエリを用いて、新たなドキュメントベクトル、或いは単語ベクトルV'_{ij}を得、
前記新たなドキュメントベクトル、或いは単語ベクトルV'_{ij}が、前記共通要素V_{ij}の要素を含む度合いに対応して、前記新たなドキュメントベクトル、或いは単語ベクトルV'_{ij}のベクトルサイズを修正する、
ことを特徴とするパッセージ分割装置。
【請求項１３】
請求項１２に記載のパッセージ分割装置であって、
前記処理部は、
前記新たなドキュメントベクトルV'_{ij}に対応する前記文、あるいはパッセージ候補として、二つの前記ドキュメントベクトルV_i，V_jに対応する前記文、あるいはパッセージ候補を連結し、新たに連結されたパッセージ候補を前記記憶部に記憶する、
ことを特徴とするパッセージ分割装置。
【請求項１４】
処理部と記憶部とを備え、入力されるドキュメントをパッセージに分割するパッセージ分割装置の処理部で実行されるパッセージ分割プログラムであって、
前記処理部を、
前記ドキュメントを文単位に分割し、
分割した前記文をクエリとして、予め前記記憶部に記憶されている複数のドキュメントから、関連するドキュメントを抽出し、
抽出した前記関連するドキュメントを用いて特徴量を作成し、
作成した前記特徴量の内の二つの類似度が所定の閾値以上である、当該特徴量の共通要素を用いて特徴量を更新し、
前記特徴量の類似度が所定の閾値以上である、当該二つの特徴量に対応する文またはパッセージ候補を連結して新たなパッセージ候補とし、
前記特徴量の更新は前記新たなパッセージ候補の特徴量を得るものである
よう動作させる、
ことを特徴とするパッセージ分割プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】

本発明は、電子化された文書の処理に係り、特に電子化書類のパセージ分割技術に関する。

【背景技術】

【０００２】

近年、文書の電子化やデータベース化が進んだことで、自然言語処理技術も大きく発展し、例えば文書の自動要約や文書検索のための自動キーワード抽出などの研究が多くなされてきた。しかしこれらの技術の対象となる文書はパッセージ毎、すなわち、話題、あるいは内容的、意味的なまとまり単位毎に分割されている、または単一のパッセージしか含まない文書を想定していることが多い。そのため、複数のパッセージを含む文書に対しては、予めパッセージを分割することが有効である。従来、このようなパッセージ分割手法としては、特許文献１や特許文献２に記載のテキストセグメンテーション手法等が知られている。

【先行技術文献】

【特許文献】

【０００３】

【特許文献１】特開２００９－１５７９５号公報

【特許文献２】特開２００４－１４５７９０号公報

【発明の概要】

【発明が解決しようとする課題】

【０００４】

しかし、従来のパッセージ分割、テキストセグメンテーションに関する手法は意味の近い文、すなわちその特徴量が似た文を含む複数のパッセージが、一つの文書に含まれる場合、パッセージを正しく分割することが難しい。その結果、文書の自動要約や文書検索のための自動キーワード抽出などを効率的に進めることができない。

【０００５】

本発明の目的は、上記課題に鑑みてなされたものであり、複数のパッセージを含む文書を有効に分割するパッセージ分割方法、装置、及びプログラムを提供することにある。

【課題を解決するための手段】

【０００６】

上記の目的を達成するため、本発明においては、処理部により、ドキュメントをパッセージに分割するパッセージ分割方法であって、処理部は、ドキュメントを文単位に分割し、分割した文をクエリとして、予め記憶されている複数のドキュメントから、関連するドキュメントを抽出して、特徴量を作成し、作成した特徴量の内の二つの特徴量の類似度が所定の閾値以上である、当該二つの特徴量の共通要素を用いて特徴量を更新するパッセージ分割方法を提供する。

【０００７】

又、上記の目的を達成するため、本発明においては、入力されるドキュメントをパッセージに分割するパッセージ分割装置であって、処理部と記憶部とを備え、処理部は、ドキュメントを文単位に分割し、分割、記憶した文をクエリとして、予め記憶部に記憶されている複数のドキュメントから、関連するドキュメントを抽出して、特徴量を作成し、作成した特徴量の内の二つの類似度が所定の閾値以上である、当該特徴量の共通要素を用いて特徴量を更新する構成のパッセージ分割装置を提供する。

【０００８】

更に、上記の目的を達成するため、本発明においては、処理部と記憶部とを備え、入力されるドキュメントをパッセージに分割するパッセージ分割装置の処理部で実行されるパッセージ分割プログラムであって、処理部を、ドキュメントを文単位に分割し、分割した文をクエリとして、予め記憶部に記憶されている複数のドキュメントから、関連するドキュメントを抽出し、抽出した関連するドキュメントを用いて特徴量を作成し、作成した特徴量の内の二つの類似度が所定の閾値以上である、当該特徴量の共通要素を用いて特徴量を更新するよう動作させるパッセージ分割プログラムを提供する。

【発明の効果】
【０００９】
本発明によれば、意味の近い文、すなわち特徴量が似た文を含む、複数のパッセージが一つの文書に含まれる場合でも、パッセージを正しく分割することが可能となる。
【図面の簡単な説明】
【００１０】
【図１Ａ】第１の実施例のパッセージ分割装置の一機能構成を示す図である。
【図１Ｂ】第１の実施例のパッセージ分割装置の一ハードウェア構成を示す図である。
【図２】第１の実施例に係る、パッセージ分割プログラムの動作の一例を示す図である。
【図３】第１の実施例に係る、ドキュメントベクトルの類似度に応じて文が連結される様子を示す図である。
【図４】第２の実施例のパッセージ分割装置の一機能構成を示す図である。
【図５】第２の実施例に係る、パッセージ分割プログラムの動作の一例を示す図である。
【図６】各実施例に係る、ドキュメントベクトルの一例を説明するための図である。
【図７】各実施例に係る、単語ベクトルの一例を説明するための図である。
【発明を実施するための形態】
【００１１】
以下、本発明の実施例を図面に従い説明するが、本発明は以下に説明する実施例に限定されるものではない。本明細書において、「文書」と「ドキュメント」とは、同義であることとする。また、「パッセージ」とは、話題、あるいは内容的、意味的なまとまりのある単位を意味する。更に、ドキュメントベクトルとは、蓄積されたドキュメントを次元とするベクトルを意味し、単語ベクトルとは、全ドキュメント中に出現する全ての単語を次元とするベクトルを意味するものとする。そして、本明細書において、文の「特徴量」とは、文の意味を定量的に示すものであり、例えば、ドキュメントベクトル、あるいは単語ベクトルはその一例として説明する。
【実施例１】
【００１２】
第１の実施例は、類似度計算にドキュメントベクトルを、類似文書検索に単語ベクトルを用いるパッセージ分割方法、装置、及びプログラムの実施例である。本実施例において、ドキュメントベクトルとは、分割装置のコーパス部に含まれる全てのドキュメントを次元とするベクトルである。
【００１３】
本実施例の詳細を説明するに先立ち、ドキュメントベクトルと単語ベクトルの一例を説明する。
図６にドキュメントベクトルの一例を示す。図６において、コーパス部に含まれるドキュメントの総数を１０として例示した。そして、検索の結果得られるドキュメントが、１、３、４、８である場合、ドキュメントベクトルは、同図の（ａ）に示すドキュメントベクトル６０１ように表わすことができる。同様に、検索の結果、検索スコアが得られる場合、得られた検索スコアを用いて、同図の（ｂ）に示すようなドキュメントベクトル６０２として表わすことができる。
【００１４】
図７に単語ベクトルの一例を示した。単語ベクトルとは、全文書中に出現する全ての単語を次元とするベクトルであり、図７の単語ベクトルでは、全てのドキュメントに出現する単語の種類を１０として例示した。そして、あるドキュメントに含まれる単語が、３、６、７、８であり、出願頻度がそれぞれ、１、５、３、９である場合、該当する要素に出現頻度を代入することで、同図に示す単語ベクトル７０１を得る。
【００１５】
図１Ａは、実施例１に係るパッセージ分割装置の機能ブロックの一例を示す図である。図１Ｂは、実施例１のパッセージ分割装置を実現するハードウェア構成の一例を示す図である。図１Ｂのハードウェア構成は、通常の処理部である中央処理部（Ｃｅｎｔｒａｌ

Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ：ＣＰＵ）１１、メモリ、ＲＡＭ、ＲＯＭ、ハードディスクドライブ（ＨＤＤ）、記憶装置等の記憶部１２、入出力部１３、ネットワークインタフェースである通信部１４からなり、これらの各ブロックは、内部バス１５によって相互に接続されているコンピュータを示している。

【００１６】

図１Ａにおいて、パッセージ分割装置１００は、制御部１０１と、入力部１０２と、文分割部１０３と、特徴量算出部１０４と、類似度計算部１０５と、検索クエリ生成部１０６と、特徴量更新部１０７と、パッセージ更新部１０８と、出力部１０９と、文記憶部１１０と、コーパス部１１１と、特徴量記憶部１１２と、パッセージ記憶部１１３と、形態素解析部１１４とを有する。前提として、コーパス部１１１には、例えば新聞記事のような文書、ドキュメントがS_D個記憶されているものとする。

【００１７】

この内、入力部１０２、出力部１０９が入出力部１３や通信部１４に対応し、文記憶部１１０と、コーパス部１１１と、特徴量記憶部１１２と、パッセージ記憶部１１３が記憶部１２のメモリや記憶装置に対応している。その余の制御部１０１、文分割部１０３と、特徴量算出部１０４と、類似度計算部１０５と、検索クエリ生成部１０６と、特徴量更新部１０７と、パッセージ更新部１０８と、形態素解析部１１４は、ＣＰＵ１１における、オペレーティングシステム（ＯＳ）や、ＲＯＭ等の記憶部に記憶された各種のプログラムの処理で実現できる。

【００１８】

図１Ａに示した実施例１のパッセージ分割装置の各機能ブロックの動きを順次説明する。

まず、パッセージ分割の対象となるドキュメントが入力部１０２から装置に入力される。文分割部１０３は、処理部であるＣＰＵ１１の所定プログラムの実行により、入力されたドキュメントを文単位に分割し、文記憶部１１０に分割結果である複数の文を記憶する。

【００１９】

同様に、特徴量算出部１０４は、文記憶部１１０から読み込んだ文各々を用いて、コーパス部１１１から関連するドキュメントを取得し、得られた複数の関連ドキュメントを、ドキュメントベクトル化して特徴量記憶部１１２に記憶する。すなわち、特徴量算出部１０４は、取得した関連ドキュメントに対応する次元に値を代入することで、図６で例示したようなドキュメントベクトルを生成する。

【００２０】

検索クエリ生成部１０６は、検索クエリを生成し、制御部１０１に送る機能を持つ。

【００２１】

特徴量算出部１０４は、制御部１０１を介して、検索クエリが与えられた場合、当該検索クエリに関連するドキュメントを文記憶部１１０から取得し、得られた複数の関連ドキュメントをドキュメントベクトル化し、特徴量として、特徴量記憶部に１１２に記憶すると共に、制御部１０１を介して、特徴量更新部１０７に出力する。

【００２２】

類似度計算部１０５は、制御部１０１の指定に基づいて、二つのドキュメントベクトルを特徴量記憶部１１２から読み出し、二つのドキュメントベクトルの類似度を計算する機能を有する。本実施例における類似度の計算方法については後述する。更に、類似度計算部１０５は、計算して得られた類似度が所定の閾値以上か否かを判断する。

【００２３】

検索クエリ生成部１０６は、制御部１０１の指定に基づいて、二つのドキュメントベクトルを特徴量記憶部１１２から読み出し、二つのドキュメントベクトルに共通するドキュメント群をコーパス部１１１から抽出する。抽出された共通するドキュメント群から検索クエリを生成し、制御部１０１へ出力する。この検索クエリの生成方法については後述する。

【００２４】
特徴量更新部１０７は、制御部１０１の指定に基づいて二つのドキュメントベクトルV_i，V_jを特徴量記憶部１１２から読み出す。また制御部１０１から一つのドキュメントベクトルV_kが特徴量更新部１０７に入力される。入力された三つのドキュメントベクトルV_k，V_i，V_jから信頼度を計算し、信頼度に基づいてV_kを修正する。この信頼度については後述する。その後、V_i，V_jを特徴量記憶部１１２から削除し、V_kを特徴量記憶部１１２に記憶する。
【００２５】
パッセージ更新部１０８は、制御部１０１の指定に基づいて、文記憶部１１０またはパッセージ記憶部１１３の中から二つの文またはパッセージ候補を読み出す。読み出された文またはパッセージ候補を文記憶部１１０またはパッセージ記憶部１１３の中から削除し、読み出された文またはパッセージ候補を連結して、その連結結果を、パッセージ候補としてパッセージ記憶部１１３に記憶する。
【００２６】
出力部１０９は文記憶部１１０とパッセージ記憶部１１３からそれぞれ文、パッセージ候補を読み出し、不明パッセージか否かを判定した上で、その判定結果に基づき、パッセージにラベルを付与して出力する。ここで不明パッセージとは、どのパッセージと連結するか判定できなかった文またはパッセージ候補を指す。不明パッセージの判定方法については後述する。
【００２７】
図２は本実施例に係るパッセージ分割装置で実行されるパッセージ分割プログラムの動作を示すフロー図である。以下、図２を用いてパッセージ分割プログラムの動作の一例について説明する。
ここでは例として、二つのパッセージを含むドキュメントが入力された場合について述べるが、入力されるドキュメント中のパッセージ数は二つ以上であっても良く、以後の処理は同じであるので、二つのパッセージを含むドキュメントを例にして説明する。
【００２８】
第一のパッセージに含まれる文をa_1，a_2，…，a_N、第二のパッセージに含まれる文をb_1，b_2，…，b_Mと定義する。ここでＮは第一のパッセージに含まれる文の数（自然数）、Ｍは第二のパッセージに含まれる文の数（自然数）である。
【００２９】
まず、ステップ２０１で入力部１０２からドキュメントが入力される。
ステップ２０２では入力されたドキュメントが、文分割部１０３により文単位に分割され、文記憶部１１０に記憶される。
【００３０】
ステップ２０３では文記憶部１１０に記憶された全ての文a_1，a_2，…，a_N、b_1，b_2，…，b_Mを特徴量算出部１０４に入力し、先に説明した通り、ドキュメントベクトルを得る。ドキュメントベクトルの算出方法としては、例えば、コサイン尺度を用いる方法が挙げられる。コサイン尺度とは二つのベクトルの類似度を計る手法の一つとして用いられるものである。二つのベクトルＱ、Ｐのコサイン尺度は以下の式１で計算される。
【００３１】
【数１】

$$\frac{\sum_i q_i p_i}{\sqrt{\sum_i q_i^2}\sqrt{\sum_i p_i^2}} \qquad (q_i \in Q, p_i \in P) \qquad \text{-----（式１）}$$

本実施例においては、上述の通り、類似するドキュメントの検索に単語ベクトルを用いる。そこで、例えば、コーパス部１１１に記憶された各ドキュメントに対して、含まれる単語の出現頻度を要素とする単語ベクトルW_i（$0 \leqq i < S_D$）を作成しておく。入力さ

れた文についても同様に単語ベクトル化し、$W_{current}$とする。単語ベクトル$W_{current}$と、単語ベクトルW_i（$0 \leqq i < S_D$）のコサイン尺度を計算し、得られた類似度が高いドキュメントからＬ番目（Ｌは所定の自然数）までのドキュメントを得て、ドキュメントベクトル化し、特徴量記憶部１１２に蓄積する。

【００３２】

尚、ここでは類似度計算の例として、コサイン尺度を用いたが、その他の尺度を用いて、類似度を計算しても良い。ドキュメントベクトルの各要素の値としては、図６の（ａ）、（ｂ）で説明したように、選定されたドキュメントは１、その他のドキュメントは０としても良いし、算出された類似度を用いるなど、なんらかの重み付けを行っても良い。

【００３３】

次にステップ２０４では、特徴量記憶部１１２に蓄積されているドキュメントベクトルを二つ読み出し、類似度計算部１０５を用いて、最も類似度の高いドキュメントベクトルの組V_i，V_jを見つける。この場合における類似度の計算方法としては、上述したコサイン尺度等を用いても良いし、二つのドキュメントベクトルの両方に存在する要素、すなわち共通要素の数などを用いても良い。

【００３４】

ステップ２０５では、類似度計算部１０５が、ステップ２０４で算出した最大類似度が、予め設定した閾値以上か否かを判定する。閾値は予め設定した固定値でも良いし、ステップ２０４で類似度を計算した際に、計算した類似度の平均や分散を計算しておき、これを用いても良い。

【００３５】

ステップ２０６およびステップ２０７は検索クエリ生成部１０６にて行われる。ステップ２０６では、ステップ２０４で算出された最大類似度が閾値以上である場合、ドキュメントベクトルの組V_i，V_jの共通要素を抽出し、これをドキュメントベクトルの共通要素V_{ij}とする。

【００３６】

ステップ２０７では、ステップ２０６で得られた共通要素V_{ij}から検索クエリを生成する。検索クエリの生成方法としては、例えばＴＦＩＤＦを用いた方法が挙げられる。ＴＦＩＤＦとは単語に関する重みの一種である。ＴＦ（Ｔｅｒｍ　Ｆｒｅｑｕｅｎｃｙ）とＩＤＦ（ＩｎＶｅｒｓｅ　Ｄｏｃｕｍｅｎｔ　Ｆｒｅｑｕｅｎｃｙ）はそれぞれ次の式で表され、ＴＦＩＤＦはＴＦとＩＤＦの積で求められる。

【００３７】

【数２】

$$tf_i = \frac{n_i}{\sum_k n_k}, \quad idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \qquad \text{-----（式２）}$$

ここでn_iはドキュメントｄにおける単語ｉの出現回数、｜Ｄ｜は総ドキュメント数、｜｛ｄ：t_i∈ｄ｝｜は単語t_iを含むドキュメント数である。本実施例においては、総ドキュメント数Ｄはコーパス部１１１に記憶されている全ドキュメント数に相当する。

【００３８】

ドキュメントdに対して、形態素解析部１１４を用いて形態素解析を行い、ＴＦＩＤＦが大きい順にS_W個の単語を抽出し、これを検索クエリとする。ＴＦＩＤＦ以外でも、例えば出現頻度の多さで重要度を決めても良いし、ドキュメントのタイトルをクエリとしても良いし、その他の方法で検索クエリを生成しても良い。

【００３９】

ステップ２０８では、ステップ２０７で生成された検索クエリを、制御部１０１を介して特徴量算出部１０４に入力し、特徴量算出部１０４において、新たなドキュメントベクトルV'_{ij}を得る。

【００４０】

続いて、新たに得られたドキュメントベクトルＶ’$_{ij}$の信頼度の計算等を実行するステップ２０９およびステップ２１０を実行する。これらのステップ２０９およびステップ２１０は、図１に示した特徴量更新部１０７にて実行される。まず、ステップ２０９では、ステップ２０８で得られたドキュメントベクトルＶ’$_{ij}$の信頼度を計算し、その結果に応じて、ドキュメントベクトルのベクトルサイズを修正する。

【００４１】

本実施例において信頼度とは、ドキュメントベクトルＶ’$_{ij}$に共通要素Ｖ$_{ij}$の要素がどれだけ含まれているかを数値化した指標である。信頼度の算出としては、例えばドキュメントベクトルＶ’$_{ij}$がドキュメントベクトルの組Ｖ$_{i}$，Ｖ$_{j}$の共通要素Ｖ$_{ij}$の要素をいくつ含んでいるかを数え上げ、共通要素Ｖ$_{ij}$の要素数で割る方法が挙げられる。その他にも、共通要素Ｖ$_{ij}$の要素が重要度によって重み付けされている場合、重み付けされた重要度の高さに応じて信頼度を算出しても良い。何れにしろ、この信頼度が、所定の値より低い場合、得られたドキュメントベクトルＶ’$_{ij}$のベクトルサイズを増減する等の信頼度のフィードバックを行う。

【００４２】

ステップ２１０では、共通要素Ｖ$_{ij}$を生成した際のドキュメントベクトルＶ$_{i}$，Ｖ$_{j}$を、特徴量記憶部１１２から削除し、新たに得られたドキュメントベクトルＶ’$_{ij}$を特徴量記憶部１１２に記憶させる。

【００４３】

ステップ２１１では、本実施例のパッセージ分割方法のために、パッセージ更新部１０８にて、Ｖ$_{i}$，Ｖ$_{j}$に対応する二つの文またはパッセージ候補を連結する。一度も連結されていない文は文記憶部１１０に記憶されている。文が連結された場合、連結前の文を文記憶部１１０から削除する。パッセージ候補と文が連結された場合、あるいはパッセージ候補同士が連結された場合には、連結前の文の削除のみならず、連結前のパッセージ候補をパッセージ記憶部１１３から削除する。連結された文またはパッセージ候補は新たなパッセージ候補としてパッセージ記憶部１１３に記憶する。

【００４４】

本実施例のパッセージ分割方法、装置においては、図２のフローにおいて、ステップ２０４からステップ２１１を繰り返すことで、目的とするパッセージを作成する。そして、ステップ２０５において、二つのドキュメントベクトルの最大類似度が所定の閾値未満の場合、パッセージの作成を終了するため、ステップ２１２を実行する。

【００４５】

ステップ２１２は、出力部１０９にて実行され、不明パッセージの判定とパッセージの出力を行うステップである。不明パッセージの判定方法の一例として、文またはパッセージ候補の中に含まれる形態素数を調べる方法がある。文またはパッセージ候補の中に含まれる形態素数が少ない場合、ドキュメントベクトルが適切に作成されず、連結が難しい場合がある。よって、スッテプ２１において、残された文またはパッセージ候補に含まれる形態素数がある閾値以下の場合、出力部４０９は、不明パッセージのラベルをつけて出力し、処理フローを終了する。

【００４６】

図３は本実施例において、ドキュメントベクトルの類似度に応じて、文が連結されていく様子を模式的に示した一例である。図２のステップ２０５における閾値は“１０”とする。

一度目の類似度算出結果が３０１である。結果３０１の中で最も類似度が高いのは、ａ$_{2}$とａ$_{3}$の組の類似度４０である。

【００４７】

よってこの組に対して図２のステップ２０５からステップ２１１の処理を行い、再度図２のステップ２０４に戻る。連結された結果をａ$_{2\ 3}$と表す。同様に結果３０２ではｂ$_{1}$とｂ$_{2}$、結果３０３ではａ$_{1}$とａ$_{2\ 3}$が類似度の最も高い組として選定され、図２のステップ２０５から図２のステップ２１１の処理が行われる。閾値を１０と設定したので、結

果３０４で選ばれる組はなく、パッセージの作成が完了する。

【００４８】

以上詳述した実施例１によれば、意味の近い文、すなわち、特徴量が似た文を含む複数のパッセージが、一つの文書に含まれる場合でも、複数のパッセージを正しく分割することが可能となり、更には、文書の自動要約や文書検索のための自動キーワード抽出など。

【実施例２】

【００４９】

実施例２は類似度計算に単語ベクトルを、類似文書検索にも単語ベクトルを用いたパッセージ分割方法、装置、及びプログラムの実施例である。

図４は実施例２に係るパッセージ分割装置の機能ブロック図である。同図のパッセージ分割装置のハードウェア構成も、実施例１の図１Ａの装置同様、図１Ｂに示したコンピュータ等で実現できることは言うまでもなく、ここではハードウェア構成の図示説明を省略する。

【００５０】

入力部４０２と、文分割部４０３と、パッセージ更新部４０８と、出力部４０９と、文記憶部４１０と、特徴量記憶部４１２と、パッセージ記憶部４１３と、形態素解析部４１４とは実施例１の対応するブロックと共通であるので、実施例１と異なる、コーパス部４１１と、特徴量算出部４０４と、類似度計算部４０５と、検索クエリ生成部４０６と、特徴量更新部４０７についてのみ説明する。なお、形態素解析部４１４は特徴量算出部４０４に接続される。

【００５１】

コーパス部４１１には、例えば新聞記事などのドキュメントの集合やシソーラス、あるいはその両方を用いる。

【００５２】

特徴量算出部４０４は、文記憶部４１０から読み込んだ文に対し、形態素解析部４１４を用いて形態素解析を行い、文を単語ベクトルへ変換する。単語ベクトルの要素数が十分でない場合にはコーパス部４１１を使用して要素数を増やす方法が有効である。例えばコーパスとしてシソーラスを用いた場合、入力文から得られた各単語をクエリとして類義語を検索し、結果として得られた類義語を単語ベクトルに追加する。またコーパスとしてドキュメントの集合を用いた場合、入力文から得られた単語ベクトルに、コーパス内の各ドキュメントから抽出した単語ベクトルを追加することができる。

【００５３】

単語ベクトルの要素を追加する方法の他の例として、上位数件のドキュメントからＴＦＩＤＦ等を用いて重要語を抜き出し、単語ベクトルに追加する方法が挙げられる。これに限らず、他の方法で文に関連する単語を得て追加して、単語ベクトルの要素数を十分にしてもよい。そして、得られた単語ベクトルを特徴量記憶部４１２に記憶する。また検索クエリ生成部４０６から、制御部４０１を介して単語ベクトルが特徴量算出部４０４に与えられた場合も、同様の方法で単語ベクトルの要素数を拡充し、特徴量記憶部１１２に記憶すると共に、制御部４０１を介して特徴量更新部４０７へ単語ベクトルを出力する。

【００５４】

本実施例の類似度計算部４０５は、制御部４０１の指定に基づいて、二つの単語ベクトルを特徴量記憶部４１２から読み出し、二つの単語ベクトルの類似度を計算する。類似度の計算方法としては、例えば、上述したコサイン尺度等が挙げられる。

【００５５】

本実施例の検索クエリ生成部４０６は、制御部４０１の指定に基づいて、二つの単語ベクトルを特徴量記憶部４１２から読み出し、二つの単語ベクトルに共通する単語群をコーパス４１１から抽出する。抽出された共通する単語群から単語ベクトルを作成し、制御部４０１を介して特徴量算出部４０４に出力する。

【００５６】

特徴量更新部４０７は、制御部４０１の指定に基づいて二つの単語ベクトルV_i，V_j

を特徴量記憶部４１２から読み出す。また制御部４０１から一つの単語ベクトルV_kが入力される。入力された三つの単語ベクトルV_k，V_i，V_jから信頼度を計算し、信頼度に基づいてV_kのベクトルサイズを修正する。その後V_i，V_jを特徴量記憶部４１２から削除し、V_kを特徴量記憶部４１２に記憶する。

【００５７】

図５は実施例２に係るプログラムの動作を示した処理フロー図である。実施例１では、類似度計算としてドキュメントベクトルを用いているが、実施例２では上述の通り、単語ベクトルを用いており、その点が実施例１と異なるが、それ以外の動作は実施例１と同様である。

【００５８】

実施例２によれば、意味の近い文、即ち、特徴量が似た文を含む複数のパッセージが、一つの文書に含まれる場合でも、パッセージを正しく分割することが可能となる。

【００５９】

なお、本発明は上記した実施例に限定されるものではなく、様々な変形例が含まれる。例えば、上記した実施例は本発明を分かりやすく説明するために詳細に説明したのであり、必ずしも説明の全ての構成を備えるものに限定されものではない。また、ある実施例の構成に他の実施例の構成を加えることが可能である。また、各実施例の構成の一部について、他の構成の追加・削除・置換をすることが可能である。

【００６０】

上記の各構成、機能、処理部、処理手段等は、それらの一部又は全部を、例えば集積回路で設計する等によりハードウェアで実現してもよい。また、上記の各構成、機能等は、それぞれの機能を実現するプログラムを実行することによりソフトウェアで実現する場合を例示して説明したが、各機能を実現するプログラム、テーブル、ファイル等の情報はメモリのみならず、ハードディスク、ＳＳＤ（Solid State DriVe）等の記憶装置、または、ＩＣカード、ＳＤカード、ＤＶＤ等の記録媒体におくことができるし、必要に応じてネットワーク等を介してダウンロード、インストールすることも可能である。

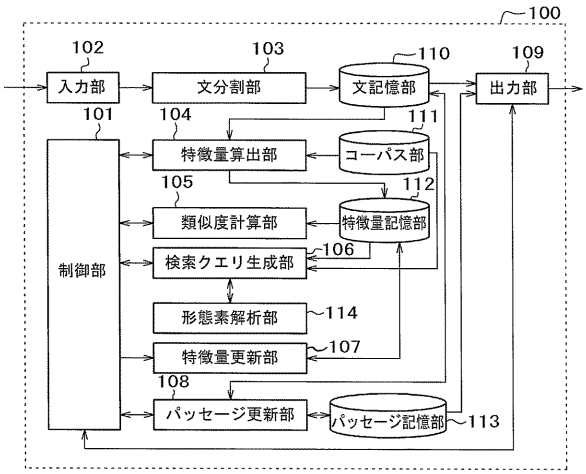
【符号の説明】

【００６１】

１１　ＣＰＵ
１２　記憶部
１３　入出力部
１４　通信部
１００、４００　パッセージ分割装置
１０１、４０１　制御部
１０２、４０２　入力部
１０３、４０３　文分割部
１０４、４０４　特徴量算出部
１０５、４０５　類似度計算部
１０６、４０６　検索クエリ生成部
１０７、４０７　特徴量更新部
１０８、４０８　パッセージ更新部
１０９、４０９　出力部
１１０、４１０　文記憶部
１１１、４１１　コーパス部
１１２、４１２　特徴量記憶部
１１３、４１３　パッセージ記憶部
１１４、４１４　形態素解析部

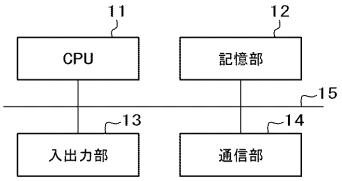
【図１Ａ】

図１Ａ



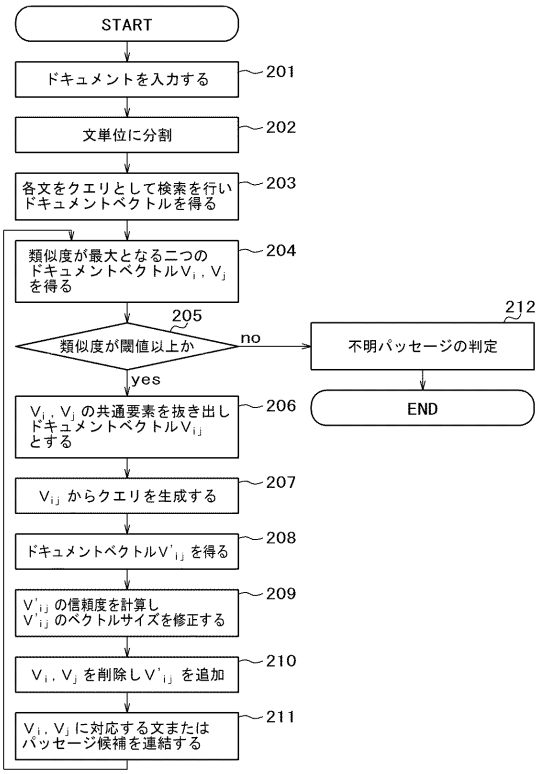
【図１Ｂ】

図１Ｂ



【図２】

図２



【図３】

図３

301

	a_1	a_2	a_3	b_1	b_2
a_1	−	25	15	5	15
a_2	−	−	40	10	7
a_3	−	−	−	5	6
b_1	−	−	−	−	35
b_2	−	−	−	−	−

文a_2と文a_3との類似度が
最も高いため、文a_2と文a_3を
連結し、文a_{23}に置き換える

302

	a_1	a_{23}	b_1	b_2
a_1	−	30	5	15
a_{23}	−	−	7	12
b_1	−	−	−	35
b_2	−	−	−	−

文b_1と文b_2との類似度が
最も高いため、文b_1と文b_2を
連結し、文b_{12}に置き換える

303

	a_1	a_{23}	b_{12}
a_1	−	30	3
a_{23}	−	−	2
b_{12}	−	−	−

文a_1と文a_{23}との類似度が
最も高いため、文a_1と文a_{23}を
連結し、文a_{123}に置き換える

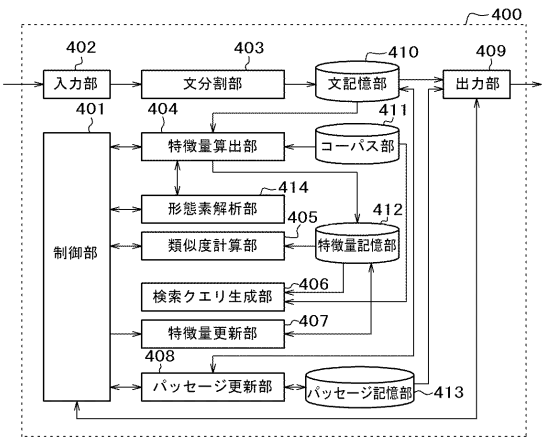
304

	a_{123}	b_{12}
a_{123}	−	5
b_{12}	−	−

類似度が閾値以下の
ため、連結を終了する

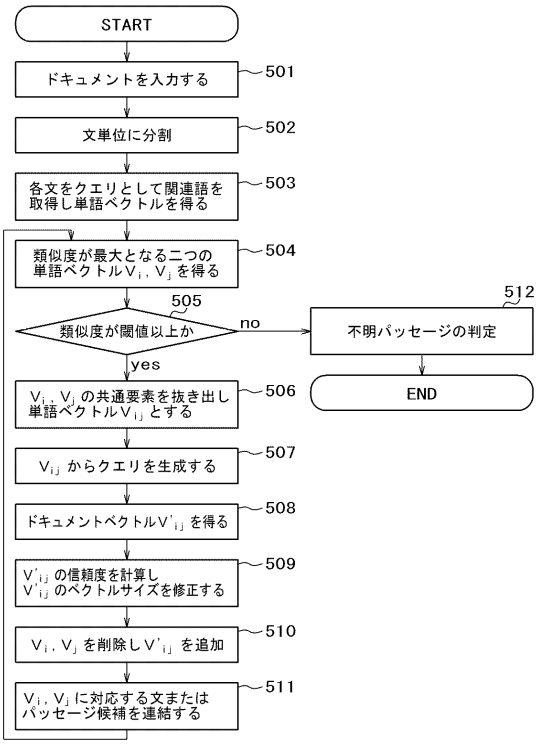
【図４】

図４



【図５】

図５



【図６】

図６

コーパスに含まれるドキュメントの総数=１０とする。
検索の結果得られたドキュメントがドキュメント１，３，４，８である場合（ａ）のように該当する要素に１を代入することでドキュメントベクトルを得る。
検索スコアが得られる場合（ｂ）のようにスコアを代入してもよい。
ここでは以下のようにスコアが算出されたとする。
ドキュメント１のスコア=０．１
ドキュメント３のスコア=０．５
ドキュメント４のスコア=０．４
ドキュメント８のスコア=０．９

601		602	
0	ドキュメント0	0	ドキュメント0
1	ドキュメント1	0.1	ドキュメント1
0	ドキュメント2	0	ドキュメント2
1	ドキュメント3	0.5	ドキュメント3
1	ドキュメント4	0.4	ドキュメント4
0	ドキュメント5	0	ドキュメント5
0	ドキュメント6	0	ドキュメント6
0	ドキュメント7	0	ドキュメント7
1	ドキュメント8	0.9	ドキュメント8
0	ドキュメント9	0	ドキュメント9

【図７】

図７

全てのドキュメントに出現する単語の種類を１０とする。
あるドキュメントに含まれる単語が単語３，６，７，８であり、以下のように出現頻度が求められたとする。
単語３の出現頻度=１
単語６の出現頻度=５
単語７の出現頻度=３
単語８の出現頻度=９
該当する要素に出現頻度を代入することで単語ベクトルを得る。

701	
0	単語0
0	単語1
0	単語2
1	単語3
0	単語4
0	単語5
5	単語6
3	単語7
9	単語8
0	単語9

フロントページの続き

(72)発明者　村上　智一
東京都国分寺市東恋ヶ窪一丁目２８０番地　株式会社日立製作所　中央研究所内
(72)発明者　今一　修
東京都国分寺市東恋ヶ窪一丁目２８０番地　株式会社日立製作所　中央研究所内

審査官　早川　学

(56)参考文献　特開２００４－１４５７９０（ＪＰ，Ａ）
望月源、外２名，語彙的連鎖に基づくパッセージ検索，自然言語処理，言語処理学会，１９９９年　4月１０日，第６巻，第３号，ｐｐ．１０１～１２６

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／２７
Ｇ０６Ｆ　１７／２１
Ｇ０６Ｆ　１７／３０