(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第5495425号
(P5495425)

(45)発行日　平成26年5月21日(2014.5.21)　(24)登録日　平成26年3月14日(2014.3.14)

(51)Int.Cl.　F I

G06F 17/21	(2006.01)	G06F	17/21	550A
G06F 17/24	(2006.01)	G06F	17/24	554M
G06F 17/27	(2006.01)	G06F	17/27	Z
G06F 17/30	(2006.01)	G06F	17/30	170A
		G06F	17/30	210A

請求項の数 11　(全 16 頁)　最終頁に続く

(21)出願番号　特願2009-265784 (P2009-265784)
(22)出願日　平成21年11月21日 (2009.11.21)
(65)公開番号　特開2011-113097 (P2011-113097A)
(43)公開日　平成23年6月9日 (2011.6.9)
審査請求日　平成24年9月4日 (2012.9.4)

特許法第３０条第１項適用　平成２１年８月２０日　発行の「第８回情報科学技術フォーラム　講演論文集（ＤＶＤ）」に発表

（出願人による申告）平成２１年度、独立行政法人情報通信研究機構「高度通信・放送研究開発委託研究／インターネット上の違法・有害情報の検出技術の研究開発」、産業技術力強化法第１９条の適用を受ける特許出願

(73)特許権者　599108264
株式会社ＫＤＤＩ研究所
埼玉県ふじみ野市大原二丁目１番１５号
(74)代理人　100135068
弁理士　早原　茂樹
(72)発明者　池田　和史
埼玉県ふじみ野市大原二丁目１番１５号
株式会社ＫＤＤＩ研究所内
(72)発明者　柳原　正
埼玉県ふじみ野市大原二丁目１番１５号
株式会社ＫＤＤＩ研究所内
(72)発明者　松本　一則
埼玉県ふじみ野市大原二丁目１番１５号
株式会社ＫＤＤＩ研究所内

最終頁に続く

(54)【発明の名称】未知語を含む文章を修正するための文章修正プログラム、方法及び文章解析サーバ

(57)【特許請求の範囲】
【請求項１】
未知語を含む解析対象文章情報に対して、該未知語を修正するようにコンピュータを機能させる文章修正プログラムであって、
修正基準文章情報の集合を記憶した基準文章記憶手段と、
前記解析対象文章情報を形態素に分割し、未知語を抽出する未知語抽出手段と、
前記未知語に対応するワイルドカードと、当該未知語に対する前方及び後方の少なくとも一方の隣接形態素とからなる検索キーを生成する検索キー生成手段と、
前記基準文章記憶手段を用いて、前記解析対象文章情報の内容に類似する１つ以上の修正基準文章情報を検索する基準文章検索手段と、
前記修正基準文章情報の中から、前記検索キーを用いて、前記ワイルドカードで検索された１つ以上の修正ルールを検索する修正ルール検索手段と、
前記修正ルールの中から、文章構成指標に基づいて１つの修正ルールを選択する修正ルール選択手段と
してコンピュータを機能させることを特徴とする文章修正プログラム。
【請求項２】
前記基準文章記憶手段は、複数の修正基準文章情報を、その内容に基づいて複数のカテゴリに分類しており、
前記基準文章検索手段は、前記解析対象文章情報の内容に最も類似する前記カテゴリを検索し、当該カテゴリに含まれる前記修正基準文章情報を出力する

ようにコンピュータを機能させることを特徴とする請求項１に記載の文章修正プログラム。

【請求項３】

前記基準文章検索手段は、前記解析対象文章情報における特徴語を抽出し、該特徴語をキーとして１つ以上の修正基準文章情報を検索するようにコンピュータを機能させることを特徴とする請求項１に記載の文章修正プログラム。

【請求項４】

前記解析対象文章情報は、ネットワークを介して公開されているブログ(Weblog)、掲示板及び／又はクチコミコメントにおける不特定多数のユーザによって記述された文章情報であるようにコンピュータを機能させることを特徴とする請求項２又は３に記載の文章修正プログラム。

【請求項５】

前記基準文章記憶手段は、複数の修正基準文章情報を、その日時情報に基づいて複数のカテゴリに分類しており、

前記基準文章検索手段は、前記解析対象文章情報に記述された日時情報、又は、前記解析対象文章情報が作成された日時情報に最も近い前記カテゴリを検索し、当該カテゴリに含まれる前記修正基準文章情報を出力する

ようにコンピュータを機能させることを特徴とする請求項１に記載の文章修正プログラム。

【請求項６】

前記解析対象文章情報は、ネットワークを介して公開されているブログ、掲示板及び／又はクチコミコメントにおける不特定多数のユーザによって記述された文章情報であり、

前記解析対象文章情報が前記ブログである場合、前記日時情報は、当該ブログのＵＲＬ(Uniform Resource Locator)アドレスに含まれたものである

ようにコンピュータを機能させることを特徴とする請求項５に記載の文章修正プログラム。

【請求項７】

前記基準文章記憶手段は、日時情報に基づいて分類された前記カテゴリ毎に、更に、修正基準文章情報の内容に基づいて複数のカテゴリに分類しており、

前記基準文章検索手段は、日時情報によって検索された前記カテゴリの中から、前記解析対象文章情報の内容に最も類似する前記カテゴリを検索し、当該カテゴリに含まれる前記修正基準文章情報を出力する

ようにコンピュータを機能させることを特徴とする請求項５又は６に記載の文章修正プログラム。

【請求項８】

前記基準文章検索手段は、日時情報によって検索された前記カテゴリの中から、前記解析対象文章情報における特徴語を抽出し、該特徴語をキーとして１つ以上の修正基準文章情報を検索するようにコンピュータを機能させることを特徴とする請求項５又は６に記載の文章修正プログラム。

【請求項９】

前記修正ルール選択手段は、前記文章構成指標として、（１）当該修正ルールにおける出現頻度、（２）前記未知語と前記修正ルールに基づく修正形態素との間の編集距離、及び／又は、（３）修正前と修正後との形態素解析コスト値の差分、に基づいて１つの修正ルールを選択するようにコンピュータを機能させることを特徴とする請求項１から８のいずれか１項に記載の文章修正プログラム。

【請求項１０】

未知語を含む解析対象文章情報に対して、該未知語を修正する装置の文章修正方法であって、

修正基準文章情報の集合を、基準文章記憶部に記憶しており、

前記解析対象文章情報を形態素に分割し、未知語を抽出する第１のステップと、

前記未知語に対応するワイルドカードと、当該未知語に対する前方及び後方の少なくとも一方の隣接形態素とからなる検索キーを生成する第２のステップと、
前記基準文章記憶部を用いて、前記解析対象文章情報の内容に類似する１つ以上の修正基準文章情報を検索する第３のステップと、
前記修正基準文章情報の中から、前記検索キーを用いて、前記ワイルドカードで検索された１つ以上の修正ルールを検索する第４のステップと、
前記修正ルールの中から、文章構成指標に基づいて１つの修正ルールを選択する第５のステップと
を有することを特徴とする装置の文章修正方法。
【請求項１１】
未知語を含む解析対象文章情報を他の公開サーバからネットワークを介して取得し、該未知語を修正する文章解析サーバであって、
修正基準文章情報の集合を記憶した基準文章記憶手段と、
前記解析対象文章情報を形態素に分割し、未知語を抽出する未知語抽出手段と、
前記未知語に対応するワイルドカードと、当該未知語に対する前方及び後方の少なくとも一方の隣接形態素とからなる検索キーを生成する検索キー生成手段と、
前記基準文章記憶手段を用いて、前記解析対象文章情報の内容に類似する１つ以上の修正基準文章情報を検索する基準文章検索手段と、
前記修正基準文章情報の中から、前記検索キーを用いて、前記ワイルドカードで検索された１つ以上の修正ルールを検索する修正ルール検索手段と、
前記修正ルールの中から、文章構成指標に基づいて１つの修正ルールを選択する修正ルール選択手段と
を有することを特徴とする文章解析サーバ。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、未知語を含む文章を修正するための文章修正プログラム、方法及び文章解析サーバに関する。
【背景技術】
【０００２】
インターネットの普及により、ブログ、掲示板又はクチコミコメントを公開するＷｅｂサイトに、様々なテキストが記述されている。「ブログ」(Weblog)とは、一般的に個人によって運営され、時事ニュースや専門的トピックスに関する自らの意見を表明するために、日記的に更新することができるサイトをいう。また、「掲示板」とは、様々なテーマについて、他人と議論を逐次に交換するためのサイトをいう。更に、「クチコミコメント」とは、人の噂のような、物事の評判などに関するコメントを記述することができるサイトをいう。これらサイトは、親しみやすさから、口語的な「くだけた表現」で記述されることが多い。
【０００３】
近年、このようなサイトによって公開されるＷｅｂ文書も、情報抽出、検索及び統計処理の対象とされるようになってきた。これらＷｅｂ文書の文章内容を解析するために、少なくとも形態素解析が必要となる。形態素解析プログラムは、解析対象となる文章を形態素に分割する。英語の文章は、“This is a pen.”のように単語ごとに区切られた「分かち書き」にされている。これに対して、日本語の文章は、分かち書きされていないため、構文解析等に先立って、形態素解析による処理が必要となる。
【０００４】
「形態素」とは、文章の構成要素のうち、意味を持つ最小の単位をいう。形態素解析プログラムは、「単語」毎に「品詞」「読み」が登録された辞書を有する。分割された形態素には、辞書を用いて「品詞」「読み」の情報が付与され、単語の配列を規定する文法規則を用いて文章を解析する。このように、形態素解析プログラムの解析精度は、辞書に依

存する。そのため、辞書に登録されていない単語を含む文章は、十分な解析精度を得ることができない。

【０００５】

これに対し、ブログ、掲示板又はクチコミコメントによって公開されるＷｅｂ文書には、一般的に、以下のような「くだけた表現」が多数含まれる。

（１）「うっそー」「すごーい」のような会話における発音の変化傾向に併せた表記

（２）「カッコイイ」のように本来ひらがなで表記される語を意図的にカタカナにした表記

（３）「ヵゎぃぃ」（「かわいい」と読む）、「ゎたＵゎ」（「わたしは」と読む）のような特有の表記

【０００６】

これら表現は、一般的に、形態素解析プログラムの辞書には登録されていないために、「未知語」として処理される。辞書に登録されていない単語は、形態素相当の単位に分割された上で、「品詞」「読み」の情報に代えて「未知語」という情報のみが付与される。

【０００７】

勿論、「未知語」として処理された語を全て、人手によって辞書に登録することができれば、形態素解析の精度を向上させることができる。しかし、「未知語」の登録には、品詞及び活用形の登録、既存の辞書との互換性の維持といった、専門的な人手のスキルが必要となる。

【０００８】

これに対して、未知語を含む文章を形態素解析に適した文章に修正する技術がある（例えば非特許文献１参照）。この技術によれば、解析に失敗して「未知語」として出力されることを前提として、未知語を含む文章である解析対象文章を形態素解析する。そして、形態素解析の結果に基づいて、未知語の部分を任意文字列に置き換えた検索キーを生成する。生成された検索キーによって、修正基準文章から、自動的に修正候補文字列が検索され、修正候補文字列から１以上の修正ルールが生成される。生成された修正ルールは、（１）同じような文脈で頻繁に使用される表現か、（２）解析対象文章の表現から変化し過ぎていないか、（３）修正後の文章が日本語として自然か、といった指標を用いて、スコアリングされ、最適な修正ルールが選択される。解析対象文章は、選択された修正ルールによって、形態素解析に適した文章に修正される。

【先行技術文献】

【非特許文献】

【０００９】

【非特許文献１】池田和史、柳原正、松本一則、滝嶋康弘、「くだけた表現を修正するための教師なし学習方式の提案と評価」、第８回情報科学技術フォーラム、２００９

【発明の概要】

【発明が解決しようとする課題】

【００１０】

非特許文献１に記載された技術によれば、修正基準文章から検索された修正候補文字列に基づいて修正ルールが生成されるため、解析対象文章は、修正基準文章に依存して修正される。その結果、解析対象文章は、正しく形態素解析される文章には修正されても、異なる意味内容の文章に修正される場合や、過剰修正される場合があった。

【００１１】

また、非特許文献１に記載された技術によれば、修正基準文章には、例えば、新聞記事の文章のような形態素解析の解析精度が高い文章が用いられている。一般に、新聞記事の文章は、「定型的な表現」が多い。「定型的な表現」は、形態素解析プログラムの辞書に登録されている確率が高いため、新聞記事の文章は、形態素解析の精度が高い。

【００１２】

例えば、解析対象文章の「えーゆーはかっこいい」という表現は、新聞記事の文章を修正基準文章とすると、「英雄はかっこいい」と修正される。この解析対象文章がＩＴやコ

ンピュータに関する文章であれば、この表現は、「ａｕはかっこいい」と修正されることが望ましい。

【００１３】

また、例えば、解析対象文章の「そんなの関係ねぇ」という表現は、「そんなの関係ない」と修正される。しかし、この解析対象文章が数年前の流行を反映した文章であれば、この表現は、「そんなの関係ねぇ」のまま修正されないことが望ましい。

【００１４】

そこで、本発明は、口語的な「くだけた表現」で記述された文章情報であっても、形態素解析によって未知語と判断されることのない、文章解析に適した文章に修正することができる文章修正プログラム、方法及び文章解析サーバを提供することを目的とする。

【課題を解決するための手段】

【００１５】

本発明によれば、未知語を含む解析対象文章情報に対して、該未知語を修正するようにコンピュータを機能させる文章修正プログラムであって、

修正基準文章情報の集合を記憶した基準文章記憶手段と、

解析対象文章情報を形態素に分割し、未知語を抽出する未知語抽出手段と、

未知語に対応するワイルドカードと、当該未知語に対する前方及び後方の少なくとも一方の隣接形態素とからなる検索キーを生成する検索キー生成手段と、

基準文章記憶手段を用いて、解析対象文章情報の内容に類似する１つ以上の修正基準文章情報を検索する基準文章検索手段と、

修正基準文章情報の中から、検索キーを用いて、ワイルドカードで検索された１つ以上の修正ルールを検索する修正ルール検索手段と、

修正ルールの中から、文章構成指標に基づいて１つの修正ルールを選択する修正ルール選択手段と

してコンピュータを機能させることを特徴とする。

【００１６】

本発明の文章修正プログラムにおける他の実施形態によれば、

基準文章記憶手段は、複数の修正基準文章情報を、その内容に基づいて複数のカテゴリに分類しており、

基準文章検索手段は、解析対象文章情報の内容に最も類似するカテゴリを検索し、当該カテゴリに含まれる修正基準文章情報を出力するようにコンピュータを機能させることも好ましい。

【００１７】

本発明の文章修正プログラムにおける他の実施形態によれば、基準文章検索手段は、解析対象文章情報における特徴語を抽出し、該特徴語をキーとして１つ以上の修正基準文章情報を検索するようにコンピュータを機能させることも好ましい。

【００１８】

本発明の文章修正プログラムにおける他の実施形態によれば、解析対象文章情報は、ネットワークを介して公開されているブログ(Weblog)、掲示板及び／又はクチコミコメントにおける不特定多数のユーザによって記述された文章情報であるようにコンピュータを機能させることも好ましい。

【００１９】

本発明の文章修正プログラムにおける他の実施形態によれば、基準文章記憶手段は、複数の修正基準文章情報を、その日時情報に基づいて複数のカテゴリに分類しており、

基準文章検索手段は、解析対象文章情報に記述された日時情報、又は、解析対象文章情報が作成された日時情報に最も近いカテゴリを検索し、当該カテゴリに含まれる修正基準文章情報を出力するようにコンピュータを機能させることも好ましい。

【００２０】

本発明の文章修正プログラムにおける他の実施形態によれば、

解析対象文章情報は、ネットワークを介して公開されているブログ、掲示板及び／又は

クチコミコメントにおける不特定多数のユーザによって記述された文章情報であり、
解析対象文章情報がブログである場合、日時情報は、当該ブログのＵＲＬ(Uniform Resource Locator)アドレスに含まれたものであるようにコンピュータを機能させることも好ましい。

【００２１】
本発明の文章修正プログラムにおける他の実施形態によれば、
基準文章記憶手段は、日時情報に基づいて分類されたカテゴリ毎に、更に、修正基準文章情報の内容に基づいて複数のカテゴリに分類しており、
基準文章検索手段は、日時情報によって検索されたカテゴリの中から、解析対象文章情報の内容に最も類似するカテゴリを検索し、当該カテゴリに含まれる修正基準文章情報を出力するようにコンピュータを機能させることも好ましい。

【００２２】
本発明の文章修正プログラムにおける他の実施形態によれば、基準文章検索手段は、日時情報によって検索されたカテゴリの中から、解析対象文章情報における特徴語を抽出し、該特徴語をキーとして１つ以上の修正基準文章情報を検索するようにコンピュータを機能させることも好ましい。

【００２３】
本発明の文章修正プログラムにおける他の実施形態によれば、修正ルール選択手段は、文章構成指標として、（１）当該修正ルールにおける出現頻度、（２）未知語と修正ルールに基づく修正形態素との間の編集距離、及び／又は、（３）修正前と修正後との形態素解析コスト値の差分、に基づいて１つの修正ルールを選択するようにコンピュータを機能させることも好ましい。

【００２４】
本発明によれば、未知語を含む解析対象文章情報に対して、該未知語を修正する装置の文章修正方法であって、
修正基準文章情報の集合を、基準文章記憶部に記憶しており、
解析対象文章情報を形態素に分割し、未知語を抽出する第１のステップと、
未知語に対応するワイルドカードと、当該未知語に対する前方及び後方の少なくとも一方の隣接形態素とからなる検索キーを生成する第２のステップと、
基準文章記憶部を用いて、解析対象文章情報の内容に類似する１つ以上の修正基準文章情報を検索する第３のステップと、
修正基準文章情報の中から、検索キーを用いて、ワイルドカードで検索された１つ以上の修正ルールを検索する第４のステップと、
修正ルールの中から、文章構成指標に基づいて１つの修正ルールを選択する第５のステップと
を有することを特徴とする。

【００２５】
本発明によれば、未知語を含む解析対象文章情報を他の公開サーバからネットワークを介して取得し、該未知語を修正する文章解析サーバであって、
修正基準文章情報の集合を記憶した基準文章記憶手段と、
解析対象文章情報を形態素に分割し、未知語を抽出する未知語抽出手段と、
未知語に対応するワイルドカードと、当該未知語に対する前方及び後方の少なくとも一方の隣接形態素とからなる検索キーを生成する検索キー生成手段と、
基準文章記憶手段を用いて、解析対象文章情報の内容に類似する１つ以上の修正基準文章情報を検索する基準文章検索手段と、
修正基準文章情報の中から、検索キーを用いて、ワイルドカードで検索された１つ以上の修正ルールを検索する修正ルール検索手段と、
修正ルールの中から、文章構成指標に基づいて１つの修正ルールを選択する修正ルール選択手段と
を有することを特徴とする。

【発明の効果】
【００２６】
本発明の文章修正プログラム、方法及び文章解析サーバによれば、口語的な「くだけた表現」で記述された文章情報であっても、形態素解析によって未知語と判断されることのない、文章解析に適した文章に修正することができる。
【図面の簡単な説明】
【００２７】
【図１】本発明における文章修正プログラムの機能構成図である。
【図２】本発明におけるカテゴリに基づいて基準文章を検索する説明図である。
【図３】本発明における特徴語に基づいて基準文章を検索する説明図である。
【図４】本発明における日時情報に基づいて基準文章を検索する説明図である。
【図５】本発明における文章解析サーバのシステム構成図である。
【図６】本発明におけるシステムのシーケンス図である。
【発明を実施するための形態】
【００２８】
以下、本発明の実施の形態について、図面を用いて詳細に説明する。
【００２９】
図１は、本発明における文章修正プログラムの機能構成図である。
【００３０】
図１によれば、文章修正プログラム１は、基準文章記憶部１１と、基準文章検索部１２と、未知語抽出部１３と、検索キー生成部１４と、修正ルール検索部１５と、修正ルール選択部１６と、修正ルール適用部１７とを有する。基準文章記憶部１１を除くこれら機能部は、装置に搭載されたコンピュータを機能させるプログラムを実行することによって実現できる。尚、各機能部の処理の流れは、コンピュータを用いた文章修正方法として実行できる。
【００３１】
基準文章記憶部１１は、修正基準文章情報の集合を記憶する。修正基準文章は、例えば技術文書、ブログテキスト、雑誌記事及び新聞記事のような様々な分野の文章を含む。基準文章記憶部１１は、修正基準文章情報を、文章内容のカテゴリ又は特徴語によって分類していてもよいし、文章の内容的日時（又は作成日時）によって分類していてもよい。基準文章記憶部１１は、基準文章検索部１２によって参照される。
【００３２】
未知語抽出部１３は、解析対象文章情報を入力する。解析対象文章情報は、Ｗｅｂサイトに公開されているブログ、掲示板又はクチコミコメントのような不特定多数のユーザによって記述された文章情報であってもよい。未知語抽出部１３は、その解析対象文章情報を形態素解析によって形態素に分割する。ここで、「できるかどうゃゎ分かりません」というくだけた表現を含む解析対象文章を例に挙げて説明する。
　解析対象文章　：できるかどうゃゎ分かりません
　形態素解析結果：できる／か／どう／ゃゎ（未知語）／分かり／ませ／ん
くだけた表現は、形態素解析辞書に登録されていない場合が多い。そこで、形態素解析辞書に登録されていない表現「ゃゎ」は、未知語として処理される。
【００３３】
未知語抽出部１３は、解析対象文章から未知語を検出した場合、形態素解析によって抽出された未知語と、未知語に隣接する形態素とを合わせた文字列とを、検索キー生成部１４へ出力する。また、未知語抽出部１３は、未知語が検出された解析対象文章情報を、基準文章検索部１２へ出力する。更に、未知語抽出部１３は、未知語が検出された解析対象文章情報を修正ルール選択部１６と、修正ルール適用部１７とへ出力する。
【００３４】
基準文章検索部１２は、解析対象文章情報を入力する。また、基準文章検索部１２は、

基準文章記憶部１１を参照し、解析対象文章の内容に類似する１つ以上の修正基準文章情報を検索によって選択する。基準文章検索部１２は、基準文章記憶部１１の分類に基づいて、例えば以下の３つの条件で検索する。
（１）解析対象文章情報の内容に最も類似するカテゴリに対応した修正基準文章情報を検索する。
（２）解析対象文章情報における特徴語を抽出し、その特徴語をキーとして修正基準文章情報を検索する。
（３）解析対象文章情報に記述された日時情報、又は、解析対象文章情報が作成された日時情報に最も近い修正基準文章情報を検索する。
そして、基準文章検索部１２は、１つ以上の修正基準文章情報を、修正ルール検索部１５へ出力する。

【００３５】
検索キー生成部１４は、未知語と、当該未知語に対する前方及び後方の少なくとも一方の隣接形態素とからなる「検索キー」を生成する。ここで、未知語と、それに隣接する前方後方の各１形態素とが、検索キーの生成に利用されたものとして説明する。
入力文字列　：どう／ヵゎ（未知語）／分かり
検索キー　：どう＊分かり（ここで、「＊」は、１以上の任意文字列を示す。）

【００３６】
検索キー生成部１４は、未知語を任意文字列（例えばワイルドカード）とし、任意文字列と、未知語に隣接する文字列と合わせた検索キーを生成する。勿論、未知語は、２以上連続するものであってもよい。また、隣接する形態素は、未知語に対する前方及び後方の少なくとも一方があればよい。同様に、隣接する形態素も、２形態素以上連続するものであってもよい。検索キー生成部１４は、生成した検索キーを修正ルール検索部１５へ出力する。

【００３７】
修正ルール検索部１５は、検索キーと１つ以上の修正基準文章情報とを入力する。修正ルールとは、未知語（例えば「ヵゎぃぃ」）から、修正候補文字列（例えば、「かわいい」）へ文字列変換するためのルールをいう。修正ルール検索部１５は、修正基準文章情報の中から、検索キーを含む修正候補文字列を検索する。そして、修正ルール検索部１５は、抽出した修正候補文字列中の任意文字列に該当する部分を、未知語に近似する部分と判断し、修正ルールとして抽出する。

【００３８】
修正ルール検索部１５は、例えば、検索により、以下の修正候補文字列を得る。修正ルール検索部１５は、抽出した修正候補文字列中の任意文字列に該当する部分から、１以上の修正ルールを抽出する。
検索キー　：どう＊分かり（ここで、「＊」は、１以上の任意文字列を示す。）
修正候補文字列：これはどう／かは／分かりません
よくあるかどう／か／分かりません
どう／したらいいのか／分かりません
この先どう／かは／分かりません
本当かどう／か／分かりませんが
使うかどう／かは／分かりませんけどね
あるかどう／かは／分かりません
どう／なっているか／分かりませんよ
修正ルール　：ヵゎ⇒かは
ヵゎ⇒か
ヵゎ⇒したらいいのか
ヵゎ⇒なっているか

【００３９】
検索キーによる検索によって得られる修正ルールは、２以上であってもよい。修正ルー

ル検索部１５は、検索キーによる検索によって得た全ての修正ルールを、修正ルール選択部１６へ出力する。

【００４０】

修正ルール選択部１６は、入力された修正ルールが２以上ある場合は、文章構成指標に基づいて文脈に適した１つの修正ルールを選択する。

【００４１】

文章構成指標は、（１）修正ルールにおける出現頻度、（２）未知語と、修正ルールに基づく修正形態素との間の編集距離、及び／又は、（３）修正前と修正後との形態素解析コスト値の差分から算出される指標をいう。修正ルール選択部１６は、この文章構成指標に基づいて１つの修正ルールを選択する。

【００４２】

修正ルール選択部１６は、例えば、以下の修正ルールを入力したとする。

修正ルール　：ヵゎ⇒かは
ヵゎ⇒か
ヵゎ⇒したらいいのか
ヵゎ⇒なっているか

【００４３】

（１）修正ルールにおける出現頻度

修正ルールにおける出現頻度は、検索された修正ルールに該当する検索結果文字列が出現した頻度をいう。以下の表では、検索結果文字列の出現頻度に基づくスコアリングの例を表す。

【表１】

修正候補文字列の出現頻度に基づくスコアリング例

修正ルール	出現頻度	出現頻度/検索件数
ｶわ⇒ かは	4	0.5
ｶわ⇒ か	2	0.25
ｶわ⇒ したらいいのか	1	0.125
ｶわ⇒ なっているか	1	0.125

【００４４】

出現頻度が高い文字列は、未知語が出現した文脈と類似した文脈の中で頻繁に利用される表現であると考えられ、修正候補文字列である可能性が高い。一方、類似した文脈の中であまり利用されていない表現は、修正候補文字列ではない可能性が高い。そこで、出現頻度の高い修正ルールは、スコアが高くなる。スコアは、出現頻度を検索件数で割り、正規化することにより、検索件数に依存しないものとしてもよい。

【００４５】

（２）未知語と、修正ルールに基づく修正形態素との間の編集距離

編集距離とは、二つの文字列がどの程度異なっているかを表す指標であり、一方の文字列を他方の文字列に変換するために必要な挿入、削除、置換の最小回数として与えられる。修正ルールに基づく修正形態素は、未知語に対して少数文字の挿入や削除、置換を実行したものであることが多い。例えば、「フォーラム」から「ファーム」への編集は、「ォ」を「ァ」に置換し、「ラ」を削除する方法が、最小の編集回数である２回となるため、編集距離は２である。以下の表は、編集距離に基づくスコアリングの例を表す。

【表２】

編集距離に基づくスコアリング例

修正ルール	編集手順	編集距離
ｶわ⇒かは	置換:2回	2
ｶわ⇒か	置換:1回、削除:1回	2
ｶわ⇒したらいいのか	置換:2回、挿入:5回	7
ｶわ⇒なっているか	置換:2回、挿入:4回	6

【００４６】

編集距離の小さい修正ルールは、スコアが高くなる。また、Ｗｅｂ文書では、「ヤバい」や「カッコイイ」のように本来ひらがなで表記されるべき語がカタカナで表記されている例が多い。そのため、例えば、カタカナをひらがなに置換する編集距離を小さくする重み付き編集距離を用いてもよい。

【００４７】

（３）形態素解析コスト値の差分

形態素解析コスト値とは、複数ある単語区切りの中で、その単語区切りがどのくらい確からしいかを表す指標である。形態素解析コスト値は、例えば、単語単体での出現確率（生起コスト)や複数単語が連続して出現する確率（連接コスト）から算出される。形態素解析コスト値は、修正ルールの文脈における適応度を評価する指標として用いられる。

【００４８】

文全体の形態素解析コスト値は、文頭から文末までの各形態素の連接コストと単語生起コストとの和を累積して算出する（累積コスト）。修正ルールの適用により、文脈における適応度が高い表現が生成された場合、その表現周辺の生起コストや連接コストは小さくなるため、文全体の形態素解析コスト値は小さくなる。一方、文脈における適応度が低い表現が生成された場合、その表現周辺の生起コストや連接コストは大きくなるため、文全体の形態素解析コスト値は大きくなる。

【００４９】

ここでは、修正ルール適用後の文全体の形態素解析コスト値と、修正前の文全体の形態素解析コスト値との差分から、形態素解析コスト値に基づくスコアとして算出する。修正ルールによって生成された表現が文脈に適応する場合、算出されるスコアは高くなる。

【００５０】

文章構成指標(score)は、（１）修正ルールにおける出現頻度(freq)、（２）未知語と、修正ルールに基づく修正形態素との間の編集距離（dist）、及び／又は、（３）形態素解析コスト値の差分(cost)から、例えば、以下の計算式により算出する。

$$score = \alpha \cdot freq + \beta \cdot dist + \gamma \cdot cost$$

ここで、α、β、γは、重み付け関数であり、修正ルールの適用と学習により、最適値を算出することができる。また、修正ルールは、適用する閾値を設定することができる。閾値を低く設定した場合、適用される修正ルールは増加するが、その中に含まれる修正ルールの誤適用も増加する。一方、閾値を高く設定した場合、適用される修正ルールは減少するが、その中に含まれる修正ルールの誤適用も減少させることができる。

【００５１】

修正ルール選択部１６は、文章構成指標に基づいて、１つの修正ルールを選択する。ここで、修正ルール選択部１６は、閾値以上且つ最大のスコアを持つ修正ルールを選択してもよい。修正ルール選択部１６は、選択した１つの修正ルールを、修正ルール適用部１７へ出力する。

【００５２】

修正ルール適用部１７は、解析対象文章に修正ルールを適用する。修正ルール適用部１７から出力された修正済みの文章情報は、様々な文章解析に適するものとなる。

【００５３】

本発明の特徴は、解析対象文章に対して、適切な修正基準文章を検索によって選択する

ことにある。従って、基準文章記憶部１１及び基準文章検索部１２における複数の実施形態を、以下の図２～図４によって説明する。

【００５４】

図２は、本発明におけるカテゴリに基づいて基準文章を検索する説明図である。

【００５５】

カテゴリに基づく基準文章検索では、予め複数の修正基準文章情報を、その内容に基づいて複数のカテゴリに分類している。基準文章のカテゴリは、文章中に出現する語の偏りに基づいて分類される。

例えば、「着信履歴」「通話」といった語が、他の語と比較して多く出現する文章は、「携帯電話」のカテゴリに分類する。

また、「ウィルス」「スパイウェア」「ファイアウォール」といった語が、他の語と比較して多く出現する文章は、「情報セキュリティ」のカテゴリに分類する。

【００５６】

カテゴリに基づく基準文章検索は、そのカテゴリに偏って多く出現した語をキーワードとして、解析対象文章をフィルタリングし、キーワードと一致する語数を計測する。ここで、解析対象文章は、キーワードと一致する語数が最も多いカテゴリに属する文章と推定される。基準文章検索部１２は、解析対象文章情報の内容に最も類似するカテゴリを検索し、カテゴリに含まれる修正基準文章情報を出力する。

【００５７】

図３は、本発明における特徴語に基づいて基準文章を検索する説明図である。

【００５８】

特徴語に基づく基準文章検索では、解析対象文章から特徴語を抽出する。抽出された特徴語をキーワードとして検索することによって、関連性の高い文章を収集する。基準文章検索部１２は、解析対象文章と関連性の高い１つ以上の修正基準文章情報を出力する。

【００５９】

特徴語の抽出には、例えばtf・idf法を用いることができる。tf・idf法とは、文章中に出現する単語の重み（＝特徴度合い）を計算する方法である。tf・idf法は、文章中の特徴的な語を抽出するためのアルゴリズムとして用いられ、tf（Term Frequency、単語の出現頻度）とidf（Inverse Document Frequency、逆出現頻度）の２つの指標によって計算される。

【００６０】

この方法によれば、単語tの文書dにおける重みw(t,d)は、次のように計算される。

$$w(t,d)=tf(t,d)\cdot idf(t)$$

tfは、その単語が一つの文章中に出現する頻度である。tf(d,t)は、文書dにおける単語tが現れる頻度を、文書内の形態素数で割った値である。tfは、文章中で出現頻度が高い単語は、その文章において重要であると判断する指標である。

【００６１】

一方、多くの文章に出現する単語は、文章を特定する性質を持たないことが多い。idfは、多くの文章に出現する語の重要度を下げ、その文章にのみ出現する単語の重要度を上げる指標である。idfは、文書の数Ｎと、単語ｔが一回以上出現する文書の数によって、以下の式のように定義される。

$$idf(t)=\log(N/df(t))+1$$

【００６２】

図４は、本発明における日時情報に基づいて基準文章を検索する説明図である。

【００６３】

日時情報に基づく基準文章検索では、予め複数の修正基準文章情報を、その日時情報に基づいて複数のカテゴリに分類している。基準文章のカテゴリは、文章の日時情報に基づいて分類される。文章の日時情報には、文章に記述された日時情報と、文章が作成された日時情報と、文章が更新された日時情報とがある。

【００６４】

文章に記述された日時情報は、文章中に記載されている情報から判断される。例えば、「池田和史が、2003/12/09に、チーム松本に入団」と記載されている場合、文章に記述された日時情報は、2003年12月9日と判断される。

【００６５】

文章が作成された日時情報は、本文中に含まれている時間情報から判断される。例えば、「http://www.blog.jp/20040105/index.html」のようにＵＲＬ(Uniform Resource Locator)アドレスに含まれていることが多い。この場合、文章が作成された日時情報は、2004年1月5日と判断される。

【００６６】

文章が更新された日時情報は、Ｗｅｂページのヘッダに含まれている時間情報から判断される。例えば、「Last-Modified:Tue,19 Aug 2003 06:10:54 GMT」のような情報がＷｅｂページのヘッダに含まれている。この場合、文章が更新された日時情報は、2003年8月19日、6時10分54秒と判断される。

【００６７】

基準文章検索部１２は、解析対象文章情報に記述された日時情報、又は、解析対象文章情報が作成された日時情報に最も近いカテゴリを検索し、そのカテゴリに含まれる修正基準文章情報を出力する。

【００６８】

また、基準文章検索部１２は、解析対象文章情報が更新された日時情報に最も近いカテゴリを検索し、そのカテゴリに含まれる修正基準文章情報を出力してもよい。更に、日時情報として、文章に記述された日時情報、文章が作成された日時情報及び文章が更新された日時情報のうち複数が検索できる場合、例えば、優先度をつけて、「文章に記述された日時情報」＞「文章が作成された日時情報」＞「文章が更新された日時情報」の順に判断してもよい。

【００６９】

基準文章検索部１２は、予め複数の修正基準文章情報を、その日時情報に基づいて複数のカテゴリに分類したものを、更に、その修正基準文章の内容に基づいて複数のカテゴリに分類したものに対して、検索が実行されてもよい。また、基準文章検索部１２は、予め複数の修正基準文章情報を、その日時情報に基づいて複数のカテゴリに分類したものに対して、解析対象文章から抽出した特徴語をキーワードとして、検索が実行されてもよい。

【００７０】

図５は、本発明における文章解析サーバのシステム構成図である。

【００７１】

図５によれば、文章解析サーバ２は、通信インタフェース部２０と、基準文章入力部２１と、解析対象文章入力部２２と、文章修正機能部２３と、基準文章記憶部１１とを有する。文章解析サーバ２は、通信インタフェース部２０を介してインターネットに接続する。

【００７２】

また、図５によれば、文章解析サーバ２は、インターネットを介して、Ｗｅｂサーバ３と通信することができる。また、Ｗｅｂサーバ３は、投稿者用端末４から接続される。

【００７３】

ＷｅｂサーバＡは、ブログテキスト、雑誌記事、新聞記事又は技術文書のような様々な文章情報を公開している。文章解析サーバ２は、インターネットを介して、ＷｅｂサーバＡから、それら文章情報を修正基準文章として収集する。

【００７４】

ＷｅｂサーバＢは、投稿者用端末４から受信した、解析対象文章であるブログテキストを公開する。文章解析サーバ２は、インターネットを介して、ＷｅｂサーバＢから、そのブログテキストを解析対象文章として取得する。

【００７５】

基準文章入力部２１は、通信インタフェース部２０を介して、修正基準文章を受信する

。その修正基準文章は、基準文章記憶部１１へ出力される。
【００７６】
解析対象文章入力部２２は、通信インタフェース部２０を介して、解析対象文章を受信する。その解析対象文章を、文章修正機能部２３へ出力される。
【００７７】
文章修正機能部２３は、図１で前述した機能構成部と全く同様である。文章修正機能部２３は、解析対象文章入力部２２から解析対象文章を入力し、修正後文章を出力する。
【００７８】
図６は、本発明におけるシステムのシーケンス図である。
【００７９】
（Ｓ６１）文章解析サーバ２は、ＷｅｂサーバＡから修正基準文章を収集する。それら修正基準文章は、基準文章記憶部１１によって記憶される。
（Ｓ６２）投稿者用端末４は、解析対象文章であるブログテキストをＷｅｂサーバＢへ投稿する。
（Ｓ６３）文章解析サーバ２は、ＷｅｂサーバＢから解析対象文章（「えーゆーはかっこいい」）を受信する。その解析対象文章は、文章修正機能部２３へ出力される。
（Ｓ６４）文章修正機能部２３は、解析対象文章情報を形態素に分割する。
（Ｓ６５）文章修正機能部２３は、形態素に未知語が含まれていた場合、未知語を抽出する。また、文章修正機能部２３は、未知語が検出された解析対象文章情報を、基準文章検索部１２へ出力する。
（Ｓ６６）基準文章検索部１２は、解析対象文章から、文章のカテゴリ、特徴語又は日時情報を抽出する。基準文章検索部１２は、解析対象文章情報の内容に類似する１つ以上の修正基準文章情報を基準文章記憶部１１から検索する。基準文章検索部１２は、文章修正機能部２３へ、修正基準文章情報を出力する。
（Ｓ６７）文章修正機能部２３は、Ｓ６５で抽出した未知語と、その未知語に対する前方及び後方の少なくとも一方の隣接形態素とからなる検索キーを生成する。
（Ｓ６８）文章修正機能部２３は、修正基準文章情報の中から、検索キーを用いて、未知語に近似する１つ以上の修正ルールを検索する。
（Ｓ６９）文章修正機能部２３は、Ｓ６８で検索した修正ルールの中から、文章構成指標に基づいて１つの修正ルールを選択する。文章修正機能部２３は、選択した修正ルールを解析対象文章に適用し、修正後文章（「ａｕはかっこいい」）を出力する。
【００８０】
以上、詳細に説明したように、本発明の文章修正プログラム、方法及び文章解析サーバによれば、口語的な「くだけた表現」で記述された文章情報であっても、形態素解析によって未知語と判断されることのない、文章解析に適した文章に修正することができる。
【００８１】
解析対象文章に類似するカテゴリの文章集合を修正基準文章として利用することによって、解析対象文章を、関連性のある意味内容の文章に修正することができる。また、解析対象文章に記述された日時又は作成された日時に近い文章集合を修正基準文章として利用することによって、文章作成時の流行を反映した文章に修正することができる。これにより、異なる意味内容の文章に修正されたり、過剰修正されたりすることなく、文章を修正することができる。
【００８２】
また、修正基準文章が、形態素解析の精度が高い一定の文章（例えば新聞記事のみ）である場合と比較して、修正後の文章における未知語の割合を減少させることができ、文章修正の精度を上げることができる。
【００８３】
前述した本発明の種々の実施形態について、本発明の技術思想及び見地の範囲の種々の変更、修正及び省略は、当業者によれば容易に行うことができる。前述の説明はあくまで例であって、何ら制約しようとするものではない。本発明は、特許請求の範囲及びその均

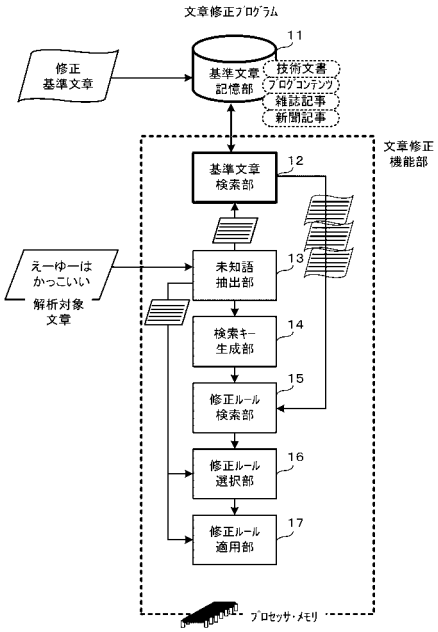
等物として限定するものにのみ制約される。

【符号の説明】

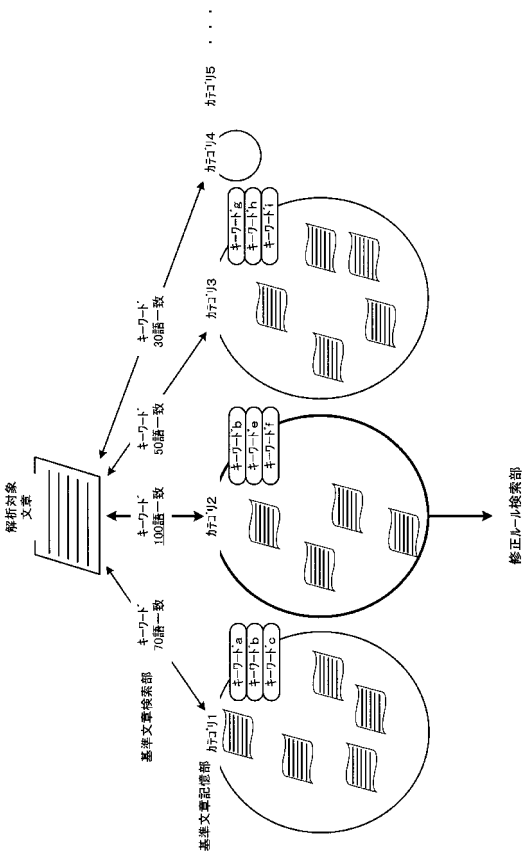
【００８４】

１　文章修正プログラム
１１　基準文章記憶部
１２　基準文章検索部
１３　未知語抽出部
１４　検索キー生成部
１５　修正ルール検索部
１６　修正ルール選択部
１７　修正ルール適用部
２　文章解析サーバ
２０　通信インタフェース部
２１　基準文章入力部
２２　解析対象文章入力部
２３　文章修正機能部
３　Ｗｅｂサーバ
４　投稿用端末

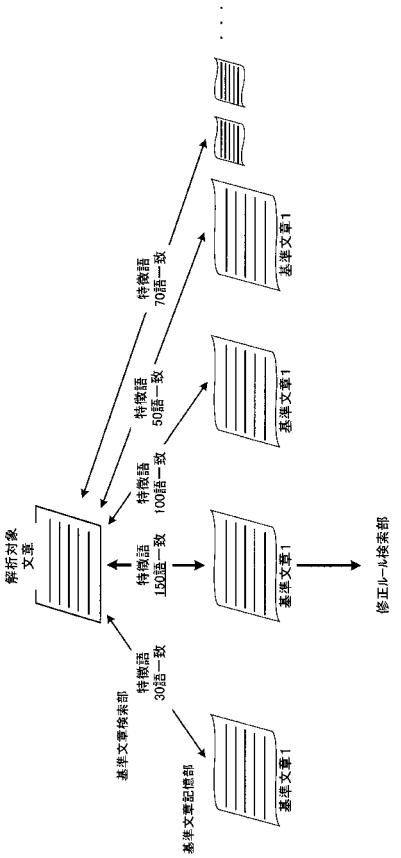
【図１】



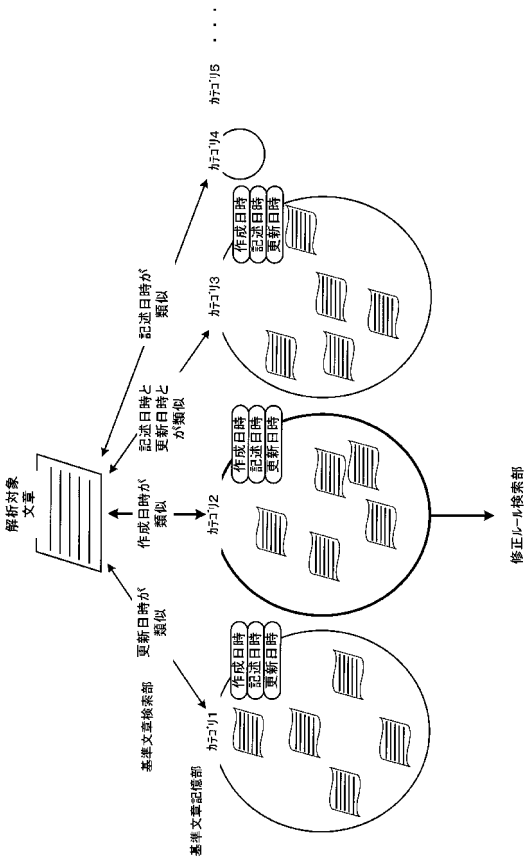
【図２】



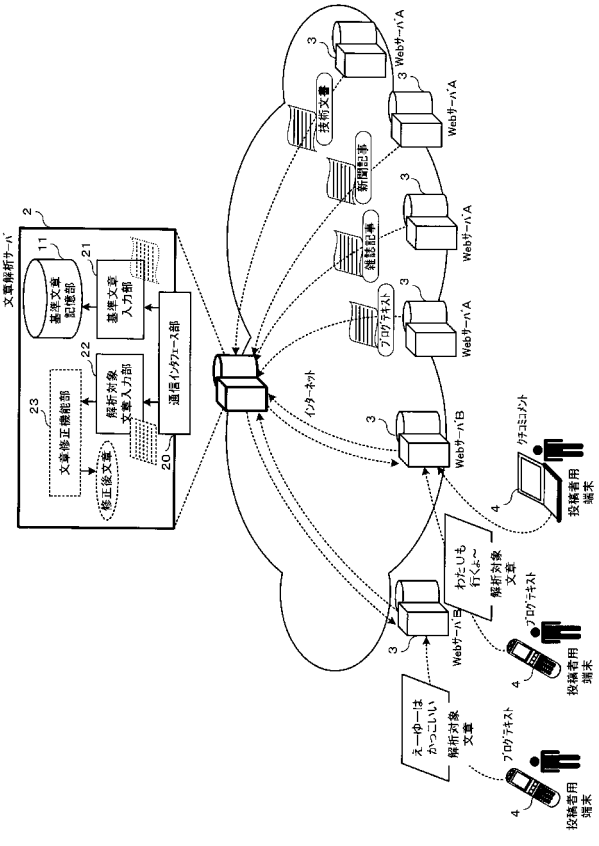
【図３】


解析対象
文章
基準文章検索部
基準文章記憶部
特徴語
150語一致
特徴語
100語一致
特徴語
50語一致
特徴語
70語一致
特徴語
30語一致
基準文章1
修正ルール検索部

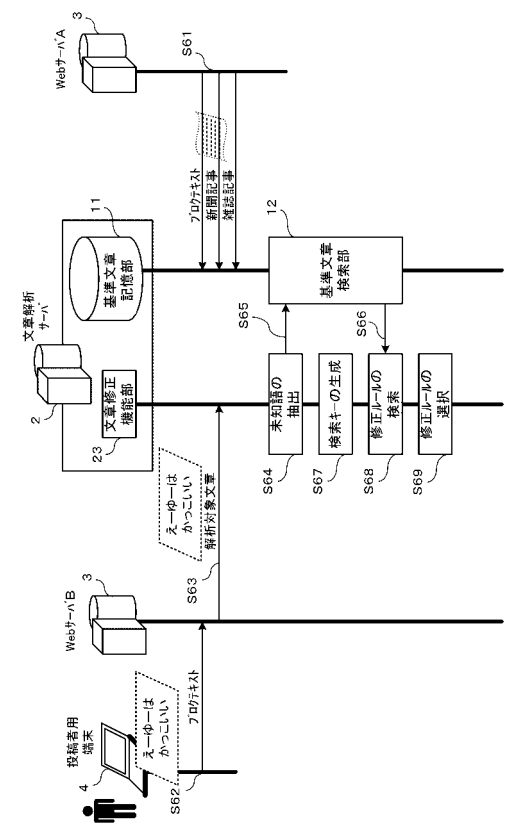
【図４】


解析対象
文章
基準文章検索部
基準文章記憶部
作成日時が
類似
記述日時と
更新日時と
が類似
更新日時が
類似
記述日時が
類似
カテゴリ1
カテゴリ2
カテゴリ3
カテゴリ4
カテゴリ5
作成日時
記述日時
更新日時
修正ルール検索部

【図５】


文章解析サーバ
基準文章
記憶部
基準文章
入力部
文章修正機能部
解析対象
文章入力部
通信インタフェース部
修正後文章
インターネット
Webサーバ A
Webサーバ B
技術文書
新聞記事
雑誌記事
ブログテキスト
クライアント
投稿者用
端末
えーゆーは
かっこいい
解析対象
文章
わたしも
行くよ～
ブログテキスト

【図６】


文章解析
サーバ
文章修正
機能部
基準文章
記憶部
基準文章
検索部
未知語の
抽出
検索キーの生成
修正ルールの
検索
修正ルールの
選択
Webサーバ A
Webサーバ B
投稿者用
端末
えーゆーは
かっこいい
解析対象文章
ブログテキスト
新聞記事
雑誌記事
S61
S62
S63
S64
S65
S66
S67
S68
S69

フロントページの続き

(51)Int.Cl. FI
G06F 17/30 350C

(72)発明者 滝嶋 康弘
埼玉県ふじみ野市大原二丁目１番１５号 株式会社ＫＤＤＩ研究所内

審査官 長 由紀子

(56)参考文献 特開平０９－１５３０３４（ＪＰ，Ａ）
特開平０５－３３４２９３（ＪＰ，Ａ）
大鹿 広憲 外3名，Ｇｏｏｇｌｅを活用した英作文支援システムの構築，ＤＥＷＳ２００５論文集 ［ｏｎｌｉｎｅ］，日本，（社）電子情報通信学会データ工学研究専門委員会，２００５年 5月 2日，DEWS2005 4B-i8
日野 浩平 外5名，ウェブからの関連語収集手法を用いた専門用語の訳語推定，言語処理学会第１１回年次大会発表論文集，日本，言語処理学会，２００５年 3月１５日，p.21-24
池田 和史 外3名，ブログにおける表記の揺れを修正するためのルール自動作成システムの提案，第７１回（平成２１年）全国大会講演論文集（２） 人工知能と認知科学，日本，社団法人情報処理学会，２００９年 3月１０日，p.2-79～2-80

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １７／２０－２８
Ｇ０６Ｆ １７／３０