

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号
特許第4200645号
(P4200645)

(45) 発行日 平成20年12月24日 (2008.12.24)

(24) 登録日 平成20年10月17日 (2008.10.17)

(51) Int.Cl.

F I

G O 6 F 17/28 (2006.01)

G O 6 F 17/21 (2006.01)

G O 6 F 17/28 Z

G O 6 F 17/21 5 5 O A

請求項の数 24 (全 19 頁)

(21) 出願番号	特願2000-278691 (P2000-278691)	(73) 特許権者	000004237
(22) 出願日	平成12年9月8日 (2000.9.8)		日本電気株式会社
(65) 公開番号	特開2002-82943 (P2002-82943A)		東京都港区芝五丁目7番1号
(43) 公開日	平成14年3月22日 (2002.3.22)	(74) 代理人	100084250
審査請求日	平成18年5月16日 (2006.5.16)		弁理士 丸山 隆夫
		(72) 発明者	福島 俊一
			東京都港区芝五丁目7番1号 日本電気株式会社社内
		審査官	成瀬 博之
		(56) 参考文献	特開平03-263260 (JP, A)
			特開平09-325960 (JP, A)
			特開平10-171806 (JP, A)
			特開平03-098165 (JP, A)
			最終頁に続く

(54) 【発明の名称】 情報処理装置、情報処理方法および記録媒体

(57) 【特許請求の範囲】

【請求項1】

テキストに含まれる固有表現の解釈を判別する情報処理装置であって、
ハイパーテキストデータベースから読み出された、テキスト及び当該テキストの有する他のテキストとのリンク関係を記憶するテキスト記憶手段と、
固有表現の候補毎にその一つ以上の解釈と当該解釈の場合に該固有表現の候補に共起する共起語とを対応付けて記憶する共起語記憶手段と、
記憶された固有表現の候補をテキストから抽出する単一テキスト解析手段と、
第1のテキストから第2のテキストを求めるためにリンク先又はリンク元をいずれか一方向に辿る回数が整数nである場合に前記第1のテキストと前記第2のテキストとがn階層のリンク関係を有すると定義した場合において、前記テキスト記憶手段で記憶されたリンク関係を参照して前記候補を抽出したテキストと少なくとも1階層のリンク関係を有する一つ以上の他のテキストを求め、当該求めた他のテキストにおいて前記共起語記憶手段で前記候補の各解釈と対応付けて記憶された共起語が検出される頻度に応じて当該候補の解釈を判別する複合テキスト解析手段と、
を備える装置。

【請求項2】

請求項1に記載の装置であって、
前記複合テキスト解析手段は、前記他のテキストにおいて前記共起語記憶手段で前記候補の各解釈と対応付けて記憶された共起語がリンク元キー文字列に含まれる頻度に応じて

10

20

当該候補の解釈を判別する、装置。

【請求項 3】

請求項 1 又は 2 に記載の装置であって、

前記複合テキスト解析手段は、前記他のテキストにおいて前記共起語記憶手段で前記候補の各解釈と対応付けて記憶された共起語がリンク元キー文字列の前後所定の範囲内で検出される頻度に応じて当該候補の解釈を判別する、装置。

【請求項 4】

請求項 1 乃至 3 のいずれか 1 項に記載の装置であって、

前記複合テキスト解析手段は、前記他のテキストにおいて前記共起語記憶手段で前記候補の各解釈と対応付けて記憶された共起語が検出される頻度から当該候補の解釈を判別できない場合に、前記他のテキストからのリンクをさらに辿ることにより、前記他のテキスト及び前記テキストではない新たな他のテキストを求め、前記求めた新たな他のテキストにおいて前記共起語記憶手段で前記候補の各解釈と対応付けて記憶された共起語が検出される頻度に応じて当該候補の解釈を判別する、装置。

10

【請求項 5】

請求項 1 乃至 4 のいずれか 1 項に記載の装置であって、

前記複合テキスト解析手段は、前記他のテキストを求める際に前記候補を抽出したテキストと予め定められた階層のリンク関係を有するテキストを前記他のテキストとする、装置。

20

【請求項 6】

請求項 1 乃至 5 のいずれか 1 項に記載の装置であって、

前記複合テキスト解析手段は、前記他のテキストを求める際に前記候補を抽出したテキストと1 階層のリンク関係を有するテキストを前記他のテキストとする、装置。

【請求項 7】

請求項 1 乃至 6 のいずれか 1 項に記載の装置であって、

前記複合テキスト解析手段は、前記共起語の検出されたテキストの数を前記頻度とする、装置。

【請求項 8】

請求項 1 乃至 6 のいずれか 1 項に記載の装置であって、

前記複合テキスト解析手段は、前記共起語の検出される位置に応じた重みを予め記憶し、
前記他のテキストにおいて前記共起語記憶手段で前記候補の各解釈と対応付けて記憶された共起語が検出される頻度を求める際に前記共起語が検出された位置に応じた重みによる重み付けを行う、装置。

30

【請求項 9】

テキストに含まれる固有表現の解釈を判別する情報処理方法であって、

コンピュータが、ハイパーテキストデータベースから読み出された、テキスト及び当該テキストの有する他のテキストとのリンク関係を記憶するテキスト記憶ステップと、

コンピュータが、固有表現の候補毎にその一つ以上の解釈と当該解釈の場合に該固有表現の候補に共起する共起語とを対応付けて記憶する共起語記憶ステップと、

40

コンピュータが、記憶された固有表現の候補をテキストから抽出する単一テキスト解析ステップと、

コンピュータが、第 1 のテキストから第 2 のテキストを求めるためにリンク先又はリンク元をいずれか一方に辿る回数が整数 n である場合に前記第 1 のテキストと前記第 2 のテキストとが n 階層のリンク関係を有すると定義した場合において、前記テキスト記憶ステップで記憶されたリンク関係を参照して前記候補を抽出したテキストと少なくとも 1 階層のリンク関係を有する一つ以上の他のテキストを求め、当該求めた他のテキストにおいて前記共起語記憶ステップで前記候補の各解釈と対応付けて記憶された共起語が検出される頻度に応じて当該候補の解釈を判別する複合テキスト解析ステップと、

を含む方法。

50

【請求項 10】

請求項 9 に記載の方法であって、

前記複合テキスト解析ステップは、前記他のテキストにおいて前記共起語記憶ステップで前記候補の各解釈と対応付けて記憶された共起語がリンク元キー文字列に含まれる頻度に応じて当該候補の解釈を判別する、方法。

【請求項 11】

請求項 9 又は 10 に記載の方法であって、

前記複合テキスト解析ステップは、前記他のテキストにおいて前記共起語記憶ステップで前記候補の各解釈と対応付けて記憶された共起語がリンク元キー文字列の前後所定の範囲内で検出される頻度に応じて当該候補の解釈を判別する、方法。

10

【請求項 12】

請求項 9 乃至 11 のいずれか 1 項に記載の方法であって、

前記複合テキスト解析ステップは、前記他のテキストにおいて前記共起語記憶ステップで前記候補の各解釈と対応付けて記憶された共起語が検出される頻度から当該候補の解釈を判別できない場合に、前記他のテキストからのリンクをさらに辿ることにより、前記他のテキスト及び前記テキストではない新たな他のテキストを求め、前記求めた新たな他のテキストにおいて前記共起語記憶ステップで前記候補の各解釈と対応付けて記憶された共起語が検出される頻度に応じて当該候補の解釈を判別する、方法。

【請求項 13】

請求項 9 乃至 12 のいずれか 1 項に記載の方法であって、

前記複合テキスト解析ステップは、前記他のテキストを求める際に前記候補を抽出したテキストと予め定められた階層のリンク関係を有するテキストを前記他のテキストとする、方法。

20

【請求項 14】

請求項 9 乃至 13 のいずれか 1 項に記載の方法であって、

前記複合テキスト解析ステップは、前記他のテキストを求める際に前記候補を抽出したテキストと 1 階層のリンク関係を有するテキストを前記他のテキストとする、方法。

【請求項 15】

請求項 9 乃至 14 のいずれか 1 項に記載の方法であって、

前記複合テキスト解析ステップは、前記共起語の検出されたテキストの数を前記頻度とする、方法。

30

【請求項 16】

請求項 9 乃至 14 のいずれか 1 項に記載の方法であって、

前記複合テキスト解析ステップは、前記共起語の検出される位置に応じた重みを予め記憶し、前記他のテキストにおいて前記共起語記憶ステップで前記候補の各解釈と対応付けて記憶された共起語が検出される頻度を求める際に前記共起語が検出された位置に応じた重みによる重み付けを行う、方法。

【請求項 17】

テキストに含まれる固有表現の解釈を判別する処理をコンピュータに実行させるプログラムを記録したコンピュータで読み込み可能な記録媒体であって、

40

コンピュータに、

ハイパーテキストデータベースから読み出された、テキスト及び当該テキストの有する他のテキストとのリンク関係を記憶するテキスト記憶処理、

固有表現の候補毎にその一つ以上の解釈と当該解釈の場合に該固有表現の候補に共起する共起語とを対応付けて記憶する共起語記憶処理、

記憶された固有表現の候補をテキストから抽出する単一テキスト解析処理、

第 1 のテキストから第 2 のテキストを求めるためにリンク先又はリンク元をいずれか一方に辿る回数が整数 n である場合に前記第 1 のテキストと前記第 2 のテキストとが n 階層のリンク関係を有すると定義した場合において、前記テキスト記憶処理で記憶されたリンク関係を参照して前記候補を抽出したテキストと少なくとも 1 階層のリンク関係を有す

50

る一つ以上の他のテキストを求め、当該求めた他のテキストにおいて前記共起語記憶処理で前記候補の各解釈と対応付けて記憶された共起語が検出される頻度に応じて当該候補の解釈を判別する複合テキスト解析処理、

を実行させるプログラムを記録した媒体。

【請求項 18】

請求項 17 に記載の媒体であって、

前記複合テキスト解析処理は、前記他のテキストにおいて前記共起語記憶処理で前記候補の各解釈と対応付けて記憶された共起語がリンク元キー文字列に含まれる頻度に応じて当該候補の解釈を判別する、媒体。

【請求項 19】

請求項 17 又は 18 に記載の媒体であって、

前記複合テキスト解析処理は、前記他のテキストにおいて前記共起語記憶処理で前記候補の各解釈と対応付けて記憶された共起語がリンク元キー文字列の前後所定の範囲内で検出される頻度に応じて当該候補の解釈を判別する、媒体。

【請求項 20】

請求項 17 乃至 19 のいずれか 1 項に記載の媒体であって、

前記複合テキスト解析処理は、前記他のテキストにおいて前記共起語記憶処理で前記候補の各解釈と対応付けて記憶された共起語が検出される頻度から当該候補の解釈を判別できない場合に、前記他のテキストからのリンクをさらに辿ることにより、前記他のテキスト及び前記テキストではない新たな他のテキストを求め、前記求めた新たな他のテキストにおいて前記共起語記憶処理で前記候補の各解釈と対応付けて記憶された共起語が検出される頻度に応じて当該候補の解釈を判別する、媒体。

【請求項 21】

請求項 17 乃至 20 のいずれか 1 項に記載の媒体であって、

前記複合テキスト解析処理は、前記他のテキストを求める際に前記候補を抽出したテキストと予め定められた階層のリンク関係を有するテキストを前記他のテキストとする、媒体。

【請求項 22】

請求項 17 乃至 21 のいずれか 1 項に記載の媒体であって、

前記複合テキスト解析処理は、前記他のテキストを求める際に前記候補を抽出したテキストと1 階層のリンク関係を有するテキストを前記他のテキストとする、媒体。

【請求項 23】

請求項 17 乃至 22 のいずれか 1 項に記載の媒体であって、

前記複合テキスト解析処理は、前記共起語の検出されたテキストの数を前記頻度とする、媒体。

【請求項 24】

請求項 17 乃至 22 のいずれか 1 項に記載の媒体であって、

前記複合テキスト解析処理は、前記共起語の検出される位置に応じた重みを予め記憶し、

前記他のテキストにおいて前記共起語記憶処理で前記候補の各解釈と対応付けて記憶された共起語が検出される頻度を求める際に前記共起語が検出された位置に応じた重みによる重み付けを行う、媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、テキスト中に出現する地名・人名・組織名などの固有表現を検出する情報処理装置、情報処理方法および記録媒体に関する。ここで、固有表現は、Named Entityに対応する日本語であり、地名・人名・組織名などを指す。「言語処理学会第 5 回年次大会」論文集（1999 年 3 月）の pp.128～131 に掲載された論文「固有表現の定義の困難さ—I R E X における N E 定義の事例から—」（著者：関根聡・江里口善生）、あるいは、1

10

20

30

40

50

９９９年９月に開催された「IREXワークショップ」などにおいて当該分野の用語として定義されている。

【０００２】

【従来の技術】

固有表現を検出するための最も基本的な方式は、固有表現の辞書を用意し、テキストと辞書とを照合することで、テキスト中に出現した固有表現を検出するものである。例えば、辞書のなかに「横浜市」（地名）、「横浜ベイスターズ」（組織名）のように登録しておき、テキスト中に「横浜市」が出現すれば、それを地名として検出し、「横浜ベイスターズ」が出現すれば、それを組織名として検出する。

【０００３】

しかし、単純に辞書と照合するだけでは、固有表現を判別できないことがある。例えば、テキスト中に「千葉」という表記が出現した場合、これは人名かもしれないし、地名かもしれないという、複数通りの解釈（曖昧性）が生ずる。同様に、テキスト中に「谷」という表記が出現した場合、これは人名かもしれないし、一般名詞かもしれないという曖昧性を持つ。さらには、テキスト中の「中央区」という表記が地名として検出できたとしても、この「中央区」が、「東京都中央区」なのか、「大阪市中央区」なのか、という解釈の曖昧性は残る。

【０００４】

このような固有表現の判別における曖昧性を解消するための手法として、従来、以下のような２通りの方法が考えられている。これらの手法はいずれも、「IREXワークショップ」（１９９９年９月）の予稿集に掲載された論文群、特に、「固有表現抽出システムの開発とIREX-NEにおける評価」（著者：竹元義美・福島俊一・山田洋志・奥村明俊・池田崇博）などに記載されている。

【０００５】

第一の手法は、固有表現の候補の前後あるいは同一テキスト内に出現する共起語を参照して、固有表現候補の曖昧性を解消する方法である。例えば、「千葉」という固有表現候補の直後に「選手」という共起語が出現すれば、この「千葉」は人名と判定できる。あるいは、「中央区」という固有表現候補について、同じテキスト中に「東京都」という共起語が出現していれば、この「中央区」は「東京都中央区」を意味する可能性が高いと解釈できる。

【０００６】

第二の手法は、固有表現の候補の表記を包含するような表記が、同一テキスト内に出現しているかを調べて、固有表現候補の曖昧性を解消する方法である。例えば、「横浜」という表記には地名と組織名の曖昧性があるが、同一テキスト内に「横浜ベイスターズ」という表記が出現しているならば、「横浜」は「横浜ベイスターズ」の省略表記、すなわち、組織名である可能性が高いと判断できる。

【０００７】

本発明と技術分野が類似する従来例１として、特開平６－５２２２１号公報の「固有名詞の自動抽出方式」がある。

【０００８】

本従来例は、オンライン・データベースやエキスパート・システム、機械翻訳システム等の自然言語インターフェースにおいて、オンライン・テキストをアクセスするデータベース・アクセス手段と、前もって作成してある固有名詞パターンを参照し、データベース・アクセス手段がアクセスして得たテキストから固有名詞候補を抽出する固有名詞抽出手段と、固有名詞抽出手段が抽出した固有名詞候補が既に辞書に登録してあるか否かを判定し、未登録の固有名詞候補を抽出する固有名詞判定手段と、固有名詞判定手段が抽出した未登録の固有名詞候補を辞書に登録する固有名詞登録手段とを有することを特徴としている。

【０００９】

【発明が解決しようとする課題】

10

20

30

40

50

しかしながら、上述した固有表現判別方式、並びに従来例 1 の固有名詞の自動抽出方式は、1つのテキスト内の解析処理によるものである。このような従来技術では、WWW (World Wide Web) のようなハイパーテキストデータベースを構成する各ノードページのテキストを対象とした場合、そのテキスト内の情報のみを用いた解析処理では、十分な判別精度が得られない可能性があるという問題がある。

【0010】

本発明は上記事情に鑑みてなされたものであり、WWWのようなハイパーテキストデータベースを構成する各ノードページのテキスト中に出現する固有表現（地名・人名・組織名など）を高精度に判別可能な情報処理装置、情報処理方法および記録媒体を提供することを目的とする。

10

【0011】

【課題を解決するための手段】

係る目的を達成するために請求項 1 記載の発明は、テキストに含まれる固有表現の解釈を判別する情報処理装置であって、ハイパーテキストデータベースから読み出された、テキスト及び当該テキストの有する他のテキストとのリンク関係を記憶するテキスト記憶手段と、固有表現の候補毎にその一つ以上の解釈と当該解釈の場合に該固有表現の候補に共起する共起語とを対応付けて記憶する共起語記憶手段と、記憶された固有表現の候補をテキストから抽出する単一テキスト解析手段と、第 1 のテキストから第 2 のテキストを求めるためにリンク先又はリンク元をいずれか一方向に辿る回数が整数 n である場合に第 1 のテキストと第 2 のテキストとが n 階層のリンク関係を有すると定義した場合において、テキスト記憶手段で記憶されたリンク関係を参照して候補を抽出したテキストと少なくとも 1 階層のリンク関係を有する一つ以上の他のテキストを求め、当該求めた他のテキストにおいて共起語記憶手段で候補の各解釈と対応付けて記憶された共起語が検出される頻度に応じて当該候補の解釈を判別する複合テキスト解析手段とを備える装置である。

20

【0012】

請求項 9 記載の発明は、テキストに含まれる固有表現の解釈を判別する情報処理方法であって、コンピュータが、ハイパーテキストデータベースから読み出された、テキスト及び当該テキストの有する他のテキストとのリンク関係を記憶するテキスト記憶ステップと、コンピュータが、固有表現の候補毎にその一つ以上の解釈と当該解釈の場合に該固有表現の候補に共起する共起語とを対応付けて記憶する共起語記憶ステップと、コンピュータが、第 1 のテキストから第 2 のテキストを求めるためにリンク先又はリンク元をいずれか一方向に辿る回数が整数 n である場合に第 1 のテキストと第 2 のテキストとが n 階層のリンク関係を有すると定義した場合において、テキスト記憶ステップで記憶されたリンク関係を参照して候補を抽出したテキストと少なくとも 1 階層のリンク関係を有する一つ以上の他のテキストを求め、当該求めた他のテキストにおいて共起語記憶ステップで候補の各解釈と対応付けて記憶された共起語が検出される頻度に応じて当該候補の解釈を判別する複合テキスト解析ステップとを含む方法である。

30

【0013】

請求項 17 記載の発明は、テキストに含まれる固有表現の解釈を判別する処理をコンピュータに実行させるプログラムを記録したコンピュータで読み込み可能な記録媒体であって、コンピュータに、ハイパーテキストデータベースから読み出された、テキスト及び当該テキストの有する他のテキストとのリンク関係を記憶するテキスト記憶処理、固有表現の候補毎にその一つ以上の解釈と当該解釈の場合に該固有表現の候補に共起する共起語とを対応付けて記憶する共起語記憶処理、記憶された固有表現の候補をテキストから抽出する単一テキスト解析処理、第 1 のテキストから第 2 のテキストを求めるためにリンク先又はリンク元をいずれか一方向に辿る回数が整数 n である場合に第 1 のテキストと第 2 のテキストとが n 階層のリンク関係を有すると定義した場合において、テキスト記憶処理で記憶されたリンク関係を参照して候補を抽出したテキストと少なくとも 1 階層のリンク関係を有する一つ以上の他のテキストを求め、当該求めた他のテキストにおいて共起語記憶処

40

50

理で候補の各解釈と対応付けて記憶された共起語が検出される頻度に応じて当該候補の解釈を判別する複合テキスト解析処理を実行させるプログラムを記録した媒体である。

【 0 0 3 2 】

【 発明の実施の形態 】

次に、添付図面を参照しながら本発明に係る実施の形態を詳細に説明する。図 1 ~ 図 1 6 を参照すると本発明に係る実施の形態が示されている。

【 0 0 3 3 】

本発明に係る第 1 の実施形態は、図 1 に示されるように、読み込み手段 2、入力メモリ 3、単一テキスト解析手段 4、中間メモリ 5、複合テキスト解析手段 6、出力メモリ 7 を備え、ハイパーテキストデータベース 1 を構成するノードページのテキストに対して、そのテキスト中に出現する固有表現のリストを出力する。これらの各手段は、プログラム制御によって動作するコンピュータを用いて実現できる。入力メモリ 3、中間メモリ 5、出力メモリ 7 は、コンピュータの主記憶部を用いてもよいが、磁気ディスク装置や光磁気ディスク装置などの外部記憶装置を用いてもよい。

【 0 0 3 4 】

まず、本実施形態が処理対象とするハイパーテキストについて説明する。

【 0 0 3 5 】

図 2 は、固有表現判別装置の入力となるハイパーテキストデータベース 1 の一部を示す例である。ハイパーテキストは、ページ（あるいはノードと呼ばれる）をテキストの 1 単位として、それらの間にリンクが設けられた形式をしている。図 2 における 1 0 と 1 1 は各々、ハイパーテキストデータベースを構成する 1 ページであり、それらの間をつなぐ矢印 1 3 はリンクを表している。リンク 1 3 に着目するならば、ページ 1 0 はリンク元ページ（リンク元テキスト）、ページ 1 1 はリンク先ページ（リンク先テキスト）となる。このようなハイパーテキストは、独自のデータ構造をもつものもあるが、最近は S G M L（Standard Generalized Markup Language）、H T M L（Hyper Text Markup Language）、X M L（Extensible Markup Language）などのマークアップ言語による記述が普及している。特に、インターネット上には W W W と呼ばれる大規模ハイパーテキストデータベースが存在し、そのなかでは H T M L による記述がスタンダードになっている。

【 0 0 3 6 】

図 3 は、図 2 のページ 1 0 を H T M L で記述した一例である。図 3 において、< > で囲まれた部分はマークアップタグであり、H T M L では A タグで囲まれた文字列がリンク元を表す。すなわち、図 2 における文字列「ベイスターズファン」は、リンク元キー文字列になる。さらに、A タグのなかで H R E F の直後に書かれた H T M L テキスト名が、そのリンク先のページ（テキスト）を意味する。すなわち、ページ 1 0 のリンク元キー文字列「ベイスターズファン」から「Baystars Fan」という名前の H T M L テキスト（ページ 1 1 がこれに相当する）へジャンプできることを意味している。なお、ここでは H T M L で記述されたハイパーテキストデータベースを例にあげて説明したが、本発明では、対象とするハイパーテキストの記述形式を H T M L に限定するものではない。S G M L や X M L で記述されたものでもよいし、独自のデータ構造を用いたハイパーテキストであってもかまわない。

【 0 0 3 7 】

読み込み手段 2 は、ハイパーテキストデータベース 1 から入力メモリ 3 へページ（テキスト）を読み込む。この読み込み手段 2 は、ハイパーテキストデータベース 1 がどこに置かれているかに応じて、ネットワークを介して外部と通信するための機構、あるいは、外部記憶装置にアクセスするための機構なども含む。どのページを読み込むかについては、（ a ）その都度、読み込む対象ページを指定して個別に読み込む方法、（ b ）ハイパーテキストの一部分を読み込む対象として範囲指定する方法、（ c ）ハイパーテキストのリンクを自動的にたどりながら、すべてのページを読み込む方法、などが考えられる。（ a ）や（ b ）の場合、外部から対象ページあるいは対象範囲を指定するため、読み込み手段 2 は、キーボードやマウスなど外部からコンピュータに指示入力を与える装置も含む。このよ

うな読み込み手段2の実現方法は、既に公知である。例えば、HTML形式のハイパーテキストデータベースを対象とするならば、(a)はNetscape NavigatorやMicrosoft Internet Explorer など広く普及したWWWブラウザのもつ基本機能であり、(b)はオートパイロットツールと呼ばれるもの、(c)はWWWロボットやクローラと呼ばれるシステムとして実用化されている。

【0038】

単一テキスト解析手段4は、入力メモリ3に読み込まれた各テキストから、そのテキスト内の解析処理によって、そのテキスト内に出現する固有表現候補を検出する。この単一テキスト解析手段4は、従来の固有表現判別装置に相当する。すなわち、本明細書の従来技術の項で述べたような公知の技術によって実現できる。単一テキスト解析手段4は、入力メモリ3からテキストを1件読み込み、そのテキストに対する解析処理を実行し、そのテキストにおける固有表現候補を中間メモリ5へ書き込む。中間メモリ5に書き込まれる情報は、対象テキスト中に検出された固有表現候補のリストであり、個々の固有表現候補に関しては、その固有表現候補がテキスト中のどこに出現したかという位置情報、および、その固有表現候補の種類(地名、人名、組織名などのいずれか)などを記録する。その際、固有表現候補の種類に曖昧性がある場合には、地名または人名、人名または一般名詞、というように複数通りの可能性を記録しておく。なお、図4には、単一テキスト解析手段4によって検出した固有表現候補の情報を記録する中間メモリ5の管理テーブルの構成が示されている。

【0039】

複合テキスト解析手段6は、単一テキスト解析手段4によって検出された固有表現候補の確からしさを、その固有表現候補の出現したリンク元あるいはリンク先のテキストを参照した解析処理によって計算する。この解析処理の実現方法、すなわち、リンク元あるいはリンク先のテキストを参照することによる固有表現候補の確からしさの計算方法としては、例えば、以下のようなものが考えられる。(ア)固有表現候補の確からしさを、その固有表現候補の出現したテキストに対するリンク元あるいはリンク先の複数テキストにおける、その固有表現候補と共起する単語の出現頻度情報をもとにして計算する方法、(イ)固有表現候補の確からしさを、その固有表現候補の出現したテキストに対するリンク元テキストにおけるリンク元キー文字列と、その固有表現候補との共起関係をもとにして計算する方法、(ウ)固有表現候補の確からしさを、その固有表現候補の出現したテキストに対するリンク元テキストにおけるリンク元キー文字列の前後一定範囲のテキストにおける、その固有表現候補と共起する単語の出現頻度情報をもとにして計算する方法、などである。このような計算を実行するために、複合テキスト解析手段6は、必要に応じて、入力メモリ3や中間メモリ5の内容を参照する。そして、複合テキスト解析手段6の結果は、出力メモリ7へ書き出す。また、図1における複合テキスト解析手段6は、装置全体を制御するような役割を持たされた構成になっており、読み込み手段2や単一テキスト解析手段4の動作制御も行う。ただし、そのような装置全体の動作制御の機能は、複合テキスト解析手段6と分離して構成するようにしてもよい。

【0040】

上記構成からなる本実施形態は、WWWのようなハイパーテキストデータベースを構成するノードページのテキスト中に出現する固有表現を高精度に判別することを目的としている。

【0041】

例えば、図2に示されたテキストには、「横浜」という表記が出現する。「横浜」という表記には地名と組織名の曖昧性があり、そのテキストを検索しただけでは、この曖昧性を解消することができない場合がある。また、図5では、テキスト16中に出現する地名「中央区」が「大阪市中央区」なのか、「東京都中央区」なのかといった曖昧性が生じている。

【0042】

このような不具合を解決するために、本実施形態は、単一テキスト解析手段4によって検

10

20

30

40

50

出したテキスト内に出現する固有表現候補の確からしさを、その固有表現候補の出現したテキストに対するリンク元、あるいはリンク先のテキストを参照した解析によって計算する複合テキスト解析手段 6 を設けたことを特徴としている。この複合テキスト解析手段 6 により、例えば、図 2 に示された例では、固有表現候補「横浜」が現れたテキストのリンク元テキスト 10 を参照することで、「横浜」は組織名である方が確からしいと判定することができる。また、図 5 に示された例では、リンク元テキスト 15 を参照すると、「大阪府」という共起語が出現していることから、テキスト 16 の「中央区」は「大阪市中央区」の解釈の方が確からしいと判定できる。

【 0 0 4 3 】

図 6 を参照しながら、本発明の特徴部分である複合テキスト解析手段 6 の詳細な構成及び動作を説明する。図 6 に示されるように複合テキスト解析手段 6 は、入力メモリ読み込み手段 30、テキストバッファ 31、固有表現候補バッファ 32、固有表現辞書部 33、共起語情報読み出し部 34、共起語情報バッファ 35、共起語検出部 36、検出結果記憶部 37、尤度計算部 38 を有して構成される。

【 0 0 4 4 】

入力メモリ読み込み手段 30 は、入力メモリ 3 に記憶されたテキストを読み込む。この入力メモリ読み込み手段 30 の読み込み対象となるテキストの範囲は、上述した（ア）、（イ）、（ウ）のそれぞれの方法において異なるが、これらについては後に詳述する。テキストバッファ 31 は、入力メモリ読み込み部 30 により読み込まれたテキストを一時的に記憶する。

【 0 0 4 5 】

固有表現候補バッファ 32 は、単一テキスト解析手段 4 の解析結果である固有表現候補を中間メモリ 5 から読み出す。

【 0 0 4 6 】

固有表現辞書 33 には、固有表現候補を特定するための辞書が記憶されている。図 7 にこの固有表現辞書の構成を示す。図 7 に示されるように固有表現辞書では、固有表現の表記 40 に対して、地名、人名、組織名などのいずれかの固有表現として解釈されるかの種別 41（一般名詞との解釈の曖昧性がある場合など固有表現以外の種別を入れることもある。）及び各々の解釈時の共起語リスト 42 を格納している。共起語リスト 42 は、単に共起語のリストだけではなく、位置関係の条件（固有表現の直後に接続するなど）も併せて格納しておいてもよい。

【 0 0 4 7 】

共起語情報読み出し部 34 は、固有表現候補バッファ 32 から固有表現候補、その位置情報、種類を読み出すと共に、この固有表現候補の、種別 41、及びその共起語リスト 42 を固有表現辞書 33 から読み出す。共起語情報読み出し部 34 により読み出された情報は、共起語情報バッファ 35 に一時的に記憶される。

【 0 0 4 8 】

共起語検出部 36 は、テキストバッファ 31 からテキストを読み出すと共に、共起語情報バッファ 35 から固有表現候補の共起語リストを読み出し、テキストの中から固有表現候補の共起語リストに挙げられた共起語を検出する。共起語検出部 36 にて検出された検出結果は、検出結果記憶部 37 に記憶される。

【 0 0 4 9 】

尤度計算部 38 は、検出結果記憶部 37 に記憶された共起語から固有表現候補の種別（地名、人名、組織名）毎の確からしさを判定し、判定結果を出力メモリ 7 に出力する。

【 0 0 5 0 】

ここで、複合テキスト解析手段 6 による（ア）、（イ）、（ウ）の解析手法について図 2 に示されたリンク元、及びリンク先のテキストを参照しながら説明すると共に、上記各解析手法の場合の複合テキスト解析手段 6 の処理動作を説明する。

【 0 0 5 1 】

図 2 では、10 がリンク元テキスト、11 がリンク先テキストである。リンク先テキスト

10

20

30

40

50

11を対象として単一テキスト解析手段4が実行され、「横浜」が固有表現候補として検出されたものとする。図7に示された固有表現辞書を用いたとするならば、図2のリンク先テキスト11における「横浜」には、地名と組織名という2通りの解釈(曖昧性)が生ずる。この曖昧性は、「横浜」に関する共起語が、テキスト11内に出現していないことから、単一テキスト解析手段4では解消できない。このとき、複合テキスト解析手段6では、リンク元テキスト10を参照することで、テキスト11における固有表現候補「横浜」の確からしさを計算する。複合テキスト解析手段6の実現方法として前述の(ア)、(イ)、(ウ)では、次のように参照するテキスト範囲が異なる。

【0052】

まず(ア)では、リンク元テキスト10の全体を参照する。そして、そのなかに出現する「横浜」の共起語として、「プロ野球」、「球団」、「ベ이스ターズ」などを見つけることで、固有表現候補「横浜」は組織名としての解釈の方が確からしいことを判定できる。

【0053】

この解析方法の場合、複合テキスト解析手段6の入力メモリ読み込み部30は、入力メモリ3に記憶されたテキストの中からリンク元ページのテキスト全体を読み込み、テキストバッファ31に記憶する。

【0054】

共起語検出部36は、テキストバッファ31からリンク元ページのテキスト全体を参照して、共起語情報バッファ35から読み出した共起語リストに挙げられた固有表現候補の共起語を検出する。

【0055】

次に、(イ)の解析方法の場合、リンク元キー文字列のみを参照する。図2では、12の「ベ이스ターズファン」がリンク元キー文字列となる。この12のなかにも、「横浜」の共起語である「ベ이스ターズ」が出現していることから、固有表現候補「横浜」は組織名としての解釈の方が確からしいことを判定できる。

【0056】

この解析方法の場合、複合テキスト解析手段6の入力メモリ読み込み部30は、入力メモリ3に記憶されたリンク元ページのテキストの中から、リンク元キー文字列だけを読み込む。すなわち、リンク元ページの中から、固有表現候補の出現するテキストのテキスト名を、リンク先テキストに設定された文字列を読み込む。図2に示された例では、固有表現候補「横浜」の現れるテキスト名「Baystars Fan.html」が、アンカータグ(... ..)内に記載されたリンク元キー文字列を入力する。

【0057】

共起語検出部36は、テキストバッファ31から読み込んだこのリンク元キー文字列に、共起語情報バッファ35から読み込んだ共起語リストに挙げられた固有表現候補の共起語が現れるか否かを検出する。

【0058】

次に、(ウ)の解析方法の場合、図8における14のように、リンク元キー文字列の前後一定範囲のテキストを参照する。図8の14の範囲には、「球団」「ベ이스ターズ」などが「横浜」の共起語として出現しているので、固有表現候補「横浜」は組織名としての解釈の方が確からしいことを判定できる。なお、リンク元キー文字列の前後一定範囲の決め方は、前後に一定の文字数、前後に一定の行数、リンク元キー文字列を含む1段落(～3段落)などのように、いろいろな方法が考えられる。

【0059】

この解析方法の場合、複合テキスト解析手段6の入力メモリ読み込み部30は、入力メモリ3に記憶されたリンク元ページのテキストの中から、リンク元キー文字列、及びこのリンク元キー文字列の前後を一定の範囲を読み込む。

【0060】

共起語検出部36は、テキストバッファ31から読み込んだこのリンク元キー文字列、及びこの前後一定の範囲のテキスト内に、共起語情報バッファ35から読み込んだ共起語リ

10

20

30

40

50

ストに挙げられた固有表現候補の共起語が現れるか否かを検出する。

【0061】

このようにして、本実施形態は、固有表現候補の出現したテキスト内だけではなく、リンク元のテキストも参照して固有表現候補の種別を特定することで、より高精度な固有表現の判別を行うことができる。

【0062】

なお、複数テキストにおける共起語の出現頻度情報に着目する際の計算方法には、いろいろなバリエーションが考えられる。例えば、図9のテキスト17とテキスト19を考えて見ると、テキスト19における「中央区」の曖昧性に対して、リンク元テキストである17には「東京都」と「大阪府」の両方が出現していて、曖昧性を解消できない。そこで、
10
(ア)の方法では、リンク元テキスト1件だけでなく複数件を参照する。さらには、リンク先テキストも参照することまで行うようにしている。図9のテキスト19に対するリンク元テキスト17と18、さらにリンク先テキスト20を参照すると、「東京都」(1回)、「大阪府」(3回)、「近畿地方」(1回)、「京都府」(1回)などが出現しており、最も多く出現している共起語である「大阪府」に着目することで、「中央区」は「大阪市中央区」という解釈の方が確からしいと判断できる。

【0063】

また、上述した方法では、固有表現候補の曖昧性解消の際に、共起語のうちでリンク元・リンク先の複数テキストにおける出現頻度の総和が最も大きいものを優先した。それ以外にも、共起語のうちでリンク元・リンク先のなるべく多数件のテキストに出現するものを
20
優先する方法も考えられる。これを図9の例で説明すれば、「大阪府」の出現するテキストは17・18・20の3件、「近畿地方」の出現するテキストは18のみで1件、「京都府」の出現するテキストも18のみで1件とカウントし、「大阪府」が最も多数件のテキストに出現した共起語ということになり、これを曖昧性解消の手がかりに用いるという方法である。

【0064】

さらに、単語(共起語)の出現回数や出現テキスト件数の単純カウントではなく、リンク元テキストか、リンク先テキストかによって、重みを変えてカウントするという方法も考えられる。例えば、単語の出現回数をリンク元テキストについては2点、リンク先テキストについては1点というように異なる重みを与えると、図9の例に関して、「東京都」は
30
2点、「大阪府」は5点、「近畿地方」は2点、「京都府」は2点となる。また、共起語の出現位置によって重みを変える方法も考えられる。例えば、リンク元キー文字列に出現した場合は4点、リンク元キー文字列の前後一定範囲のテキストに出現した場合は3点、リンク元テキスト内に出現した場合は2点、リンク先テキスト内に出現した場合は1点というような重みの付け方である。

【0065】

また、以上で述べた複合テキスト解析手段6の実現方法・処理例では、固有表現候補の出現したテキストからリンクを1階層分たどった範囲で、曖昧性解消の手がかりとなる共起語を探した。しかし、1階層よりも広い範囲から共起語を探すようにしてもよい。図10の例では、テキスト23に出現した「横浜」が固有表現候補であり、地名と組織名という
40
曖昧性を持つ。このテキスト23のリンク元テキストの範囲(1階層分)では、「横浜」の曖昧性を解消する手がかりとなる共起語は出現していない。そこで、さらにもう1階層分、リンク元へ逆上ると、テキスト21を参照できる。テキスト21には「ベ이스ターズ」という共起語が存在するため、「横浜」は組織名としての解釈を優先することができる。このような複数階層逆上ったテキスト参照に関して、最初から1階層ではなく、N階層(Nは1より大きなある値)の範囲を参照範囲と決めておく方法もあるし、また、1階層の範囲で曖昧性解消ができなかった時に、参照範囲を1階層ずつ増やしていくという方法もある。例えば、リンク元及びリンク先の前後一階層分のテキストを参照して曖昧性を解消できなかった時に、リンク先のリンク、若しくはリンク元のリンクをたどり、共起語を検出していく方法である。

10

20

30

40

50

【 0 0 6 6 】

また、単純に階層を増やしていくのではなく、1階層目はそのテキスト全体を参照するが、2階層目はリンク元キー文字列（或いはリンク元キー文字列の前後一定範囲のテキスト）のみを参照範囲とするような方法も考えられる。図10の例で言えば、テキスト23に対して、1階層目のテキスト22はその全体を参照し、2階層目のテキスト21はリンク元キー文字列の「ベイスターズファン」の部分のみを参照するという方法である。

【 0 0 6 7 】

また、リンクを逆上るだけでなく、リンク先の方向も含めて階層を増やしていく方法も考えられる。例えば、図10の例において、テキスト23に対して、リンク元のテキスト22へ1階層分逆上り、今度はそのリンク先の方向にたどって、テキスト24を参照することも可能である。テキスト23に対して、2階層分の範囲として、テキスト21、テキスト22、テキスト24の3つを参照する方法、リンク元方向のみにたどってテキスト21とテキスト22の2つを参照する方法、あるいは、兄弟関係のリンク参照を優先してテキスト22とテキスト24の2つを参照する方法などが考えられる。

【 0 0 6 8 】

次に、図11のフローチャートを用いて、本実施の形態の動作を説明する。

まず、図11のステップS201にあるように、読み込み手段2によって、ハイパーテキストデータベース1からテキストを読み込んで、入力メモリ3に書き込む。読み込み手段2の実現方法によっては、ステップS201と以降のステップSとを交互に進めるような処理手順もとる得るが、ここでは前述の読み取り手段2の実現方法（b）を用いて、ある範囲内のテキストをまとめて読み込むものとし、その結果、読み込まれたテキストの件数はN件であったとする。

【 0 0 6 9 】

次に、N件のテキストの各々について、ステップS204以降の手順を実行する。図11のフローチャートでは、ステップS202でkの値を1にセットした上で、ステップS209でkの値を1ずつ増やしなが、ループ処理を実行している部分が、これに該当する。k番目のテキスト（ $k = 1 \sim N$ ）に対する処理として、まず、ステップS204を実行する。ステップS204では、単一テキスト解析手段4によってテキストkを解析し、テキストk内に出現する固有表現の候補を検出して、中間メモリ5へ書き込む。ここで検出された固有表現候補の数を M_k とし、個々の固有表現候補を $c[k, j]$ （ $j = 1 \sim M_k$ ）で表すものとする。すなわち、 $c[k, j]$ は、テキストkにおいて検出されたj番目の固有表現候補である。

【 0 0 7 0 】

次に、テキストkにおいて検出された M_k 個の固有表現候補の各々について、ステップS207以降の手順を実行する。図11のフローチャートでは、ステップS205でjの値を1にセットした上で、ステップS208でjの値を1ずつ増やしなが、ループ処理を実行している部分が、これに該当する。j番目の固有表現候補 $c[k, j]$ に対する処理として、ステップS207を実行する。ステップS207では、複合テキスト解析手段6によって、テキストkに対するリンク元あるいはリンク先のテキストを参照して解析し、固有表現候補 $c[k, j]$ の確からしさを計算し、出力メモリ7へ書き込む。読み込み手段2の実現方法によっては、ステップS207の段階で、テキストkのリンク元テキストやリンク先テキストが入力メモリ3に含まれていないというケースもあり得る。その場合は、そのようなテキストkに関する固有表現候補 $c[k, j]$ については単一テキスト解析手段4の結果をそのまま出力することにしてもよいし、あるいは、ステップS207の段階で、読み込み手段2によってリンク元あるいはリンク先のテキストを改めて読み込むことにしてもよい。

【 0 0 7 1 】

テキストkに対する M_k 個の固有表現候補の確からしさの計算が終了したら、ステップS206を経て、次のテキスト（ $k + 1$ ）の処理へ進む。そして、N件のテキストに対する処理が完了したら、ステップS203を経て、フローチャート全体の処理が終了する。

【 0 0 7 2 】

次に、本発明に係る第 2 の実施形態について添付図面を参照しながら詳細に説明する。図 1 2 には、本発明の固有表現判別装置を利用した地図検索装置の構成を示すブロック図が示されている。

【 0 0 7 3 】

図 1 2 に示されるように固有表現判別装置を利用した地図情報検索装置は、ハイパーテキストデータベース 5 0、固有表現判別装置 5 1、位置依存コンテンツデータベース 5 2、地図データベース 5 3、データベース検索装置 5 4、表示装置 5 5、位置条件入力装置 5 6 を備える。

【 0 0 7 4 】

ハイパーテキストデータベース 5 0 は、ハイパーテキストが格納されている。例えば、インターネット上の WWW がこれに相当する。

【 0 0 7 5 】

固有表現判別装置 5 1 は、ハイパーテキストデータベース 5 0 内のテキストから地名表記を検出する。これまで説明した第 1 の実施の形態がこれに相当する。ただし、本実施形態は、固有表現のうち地名と判定されたもののみを使用する。

【 0 0 7 6 】

位置依存コンテンツデータベース 5 2 は地名表記と、そのノードページ番号が対応付けられて格納されている。例えば、「東京都港区」という地名表記に関して、ノードページ 3 1 が対応し、「群馬県前橋市」という地名表記に関して、ノードページ 4 0 が対応している。

【 0 0 7 7 】

地図データベース 5 3 は、地図の 2 次元座標データと、その上にマッピングされた地名表記を格納している。

【 0 0 7 8 】

位置条件入力装置 5 6 は、「東京都港区」というような地名表記を利用者が入力するための装置である。キーボードのような文字列入力手段、マウスなどのポインティングデバイス、さらには、GPS のような人や車の現在位置を自動的に取得するシステムなどが用いられる。

【 0 0 7 9 】

データベース検索装置 5 4 は、位置条件入力装置 5 6 で指定された条件で、地図データベース 5 3 と位置依存コンテンツデータベース 5 2 を検索して、その結果を表示装置 5 5 に表示する。

【 0 0 8 0 】

図 1 4 は、図 1 3 の位置依存コンテンツデータベースを用いて地図上にコンテンツを表示した例である。位置条件入力装置 5 6 からは関東地方の地名が指示されたものとする。

【 0 0 8 1 】

次に、本発明に係る第 3 の実施形態について図面を参照して詳細に説明する。図 1 5 を参照すると、本発明に係る第 3 の実施形態は、入力装置 1 0 0、データ処理装置 1 1 0、記憶装置 1 2 0、出力装置 1 4 0 を備え、さらに、上述した第 1 の実施形態の固有表現判別装置を実現するためのプログラムを記録した記録媒体 1 3 0 を備える。この記録媒体 1 3 0 は、磁気ディスク、半導体メモリ、CD-ROM その他の記録媒体であってよい。

【 0 0 8 2 】

入力装置 1 0 0 は、マウス、キーボード等、操作者からの指示を入力するための装置である。また、出力装置 1 4 0 は、表示画面、プリンタ等のデータ処理装置 1 1 0 による処理結果を出力する装置である。

【 0 0 8 3 】

固有表現判別装置を実現するためのプログラムは、記録媒体 1 3 0 からデータ処理装置 1 1 0 に読み込まれ、データ処理装置 1 1 0 の動作を制御し、記憶装置 1 2 0 に入力メモリ 3 と中間メモリ 5 と出力メモリ 7 とを生成する。データ処理装置 1 1 0 は、固有表現判別

10

20

30

40

50

装置を実現するためのプログラムの制御により第 1 の実施形態における読み込み手段 2、複合テキスト解析手段 6、及び単一テキスト解析手段 4 による処理と同一の処理を実行する。

【0084】

次に、本発明に係る第 4 の実施形態について図面を参照して詳細に説明する。図 16 を参照すると、本発明に係る第 4 の実施形態は、入力装置 200、データ処理装置 210、記憶装置 220、出力装置 240 を備え、さらに、上述した第 2 の実施形態の地図情報検索装置を実現するためのプログラムを記録した記録媒体 230 を備える。この記録媒体 230 は、磁気ディスク、半導体メモリ、CD-ROM その他の記録媒体であってよい。

【0085】

入力装置 200 は、マウス、キーボード等、操作者からの指示を入力するための装置である。また、出力装置 240 は、表示装置、プリンタ等のデータ処理装置 210 による処理結果を出力する装置である。

【0086】

地図情報検索装置を実現するためのプログラムは、記録媒体 230 からデータ処理装置 210 に読み込まれ、データ処理装置 210 の動作を制御し、記憶装置 220 に入力メモリ 3、中間メモリ 5、出力メモリ 7、位置依存コンテンツデータベース 52、地図データベース 53 を生成する。データ処理装置 210 は、地図情報検索装置を実現するためのプログラムの制御により第 1 の実施形態における読み込み手段 2、複合テキスト解析手段 6、単一テキスト解析手段 4、第 2 の実施形態におけるデータベース検索装置 54 による処理と同一の処理を実行する。

【0087】

上述した実施形態は、本発明の好適な実施の形態である。但し、これに限定されるものではなく、本発明の要旨を逸脱しない範囲内において種々変形実施が可能である。

【0088】

【発明の効果】

以上の説明より明らかなように本発明は、固有表現の検出・判別に関して、その固有表現の出現したテキスト内だけではなく、リンク元やリンク先のテキストも参照して解釈することで、従来よりも曖昧性を解消することが可能になり、高精度な固有表現判別を実現できる。

【図面の簡単な説明】

【図 1】本発明の実施の形態を示すブロック図である。

【図 2】本発明の実施の形態における処理対象の例を示す図である。

【図 3】HTML で記述されたテキストの例を示す図である。

【図 4】中間メモリの管理テーブルの例を示す図である。

【図 5】本発明の実施の形態における処理対象の例を示す図である。

【図 6】複合テキスト解析手段の構成を表すブロック図である。

【図 7】固有表現辞書の内容を示す図である。

【図 8】本発明の実施の形態における処理対象の例を示す図である。

【図 9】本発明の実施の形態における処理対象の例を示す図である。

【図 10】本発明の実施の形態における処理対象の例を示す図である。

【図 11】本発明の実施形態の動作を示すフローチャートである。

【図 12】本発明に係る第 2 の実施の形態を示すブロック図である。

【図 13】位置依存コンテンツデータベースのデータ内容を示す図である。

【図 14】地図情報検索装置の表示結果を示す図である。

【図 15】本発明に係る第 3 の実施の形態を示すブロック図である。

【図 16】本発明に係る第 4 の実施の形態を示すブロック図である。

【符号の説明】

1 ハイパーテキストデータベース

2 読み込み手段

10

20

30

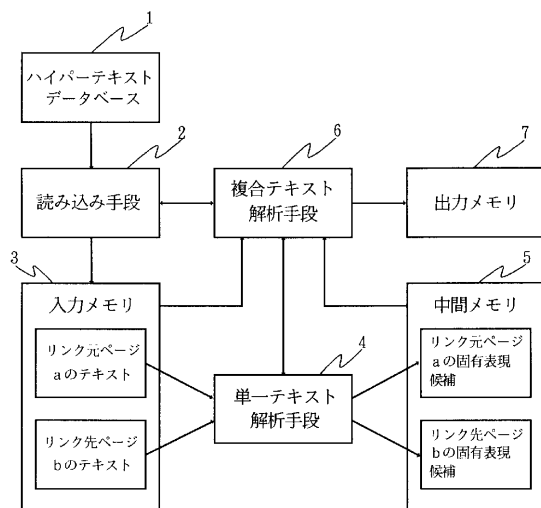
40

50

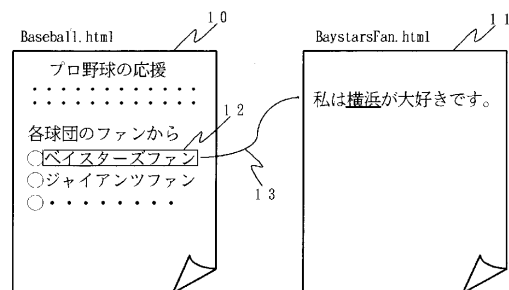
- 3 入力メモリ
- 4 単一テキスト解析手段
- 5 中間メモリ
- 6 複合テキスト解析手段
- 7 出力メモリ
- 10 リンク元ページ
- 11 リンク先ページ
- 12 リンク元キー文字列
- 13 リンク
- 14 リンク元キー文字列の前後一定範囲のテキスト
- 40 固有表現辞書における表記
- 41 固有表現辞書における種別
- 42 固有表現辞書における共起語

10

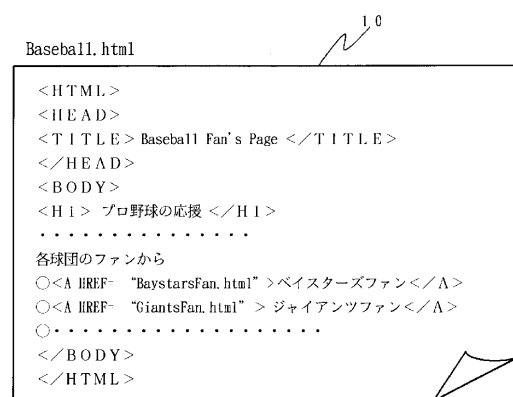
【 図 1 】



【圖 2】



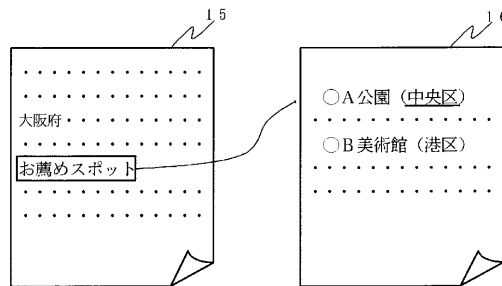
【図 3】



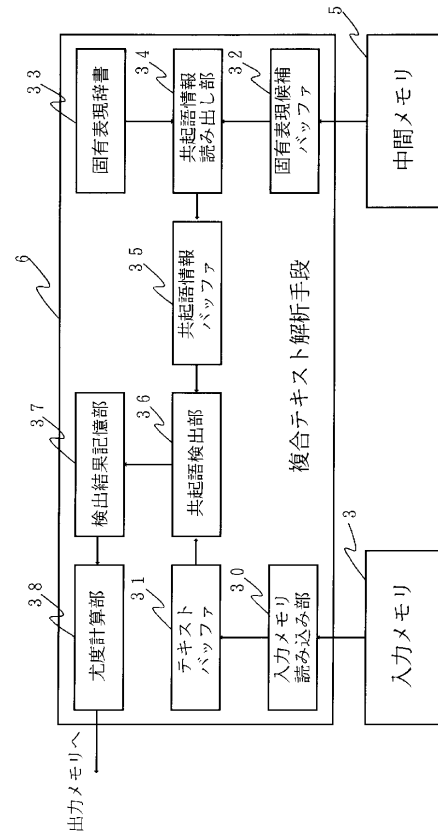
【図 4】

検出対象テキスト識別情報			
識別番号	固有表現候補	位置情報	種類

【図 5】



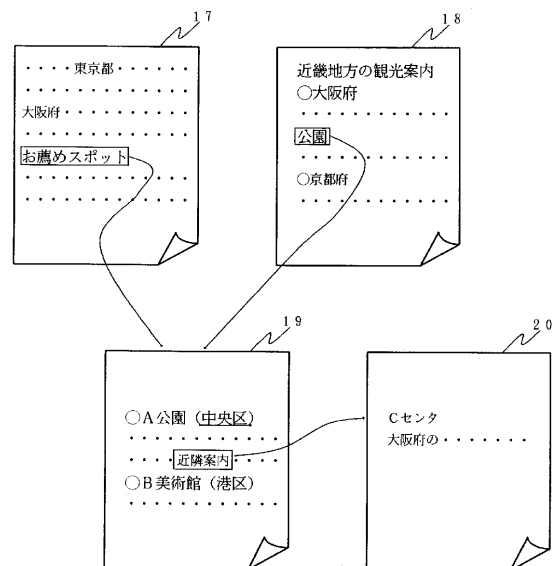
【図 6】



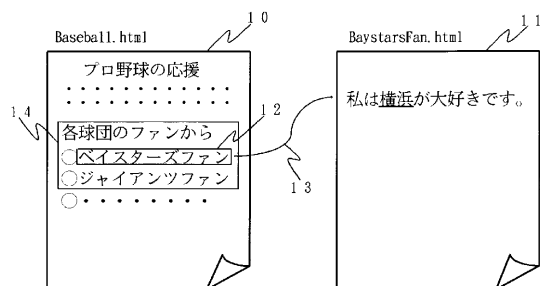
【図 7】

固有表現	種別	共起語リスト
横浜	地名	神奈川、関東、・・・
	組織名	ベイスターズ、プロ野球、球団・・・
⋮	⋮	⋮

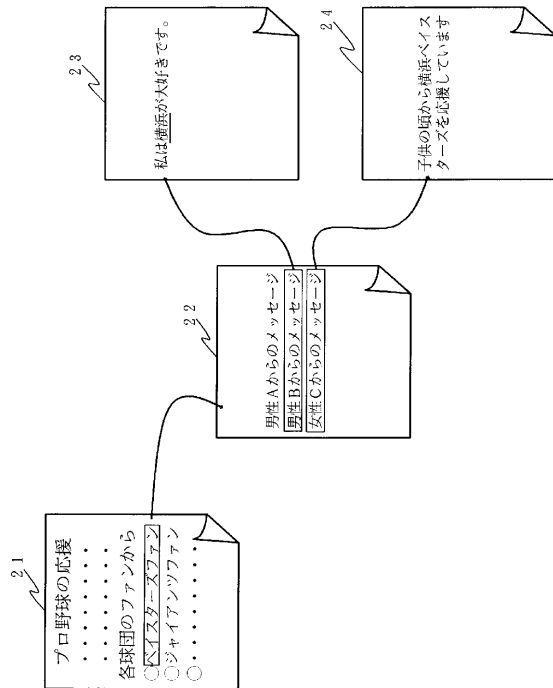
【図 9】



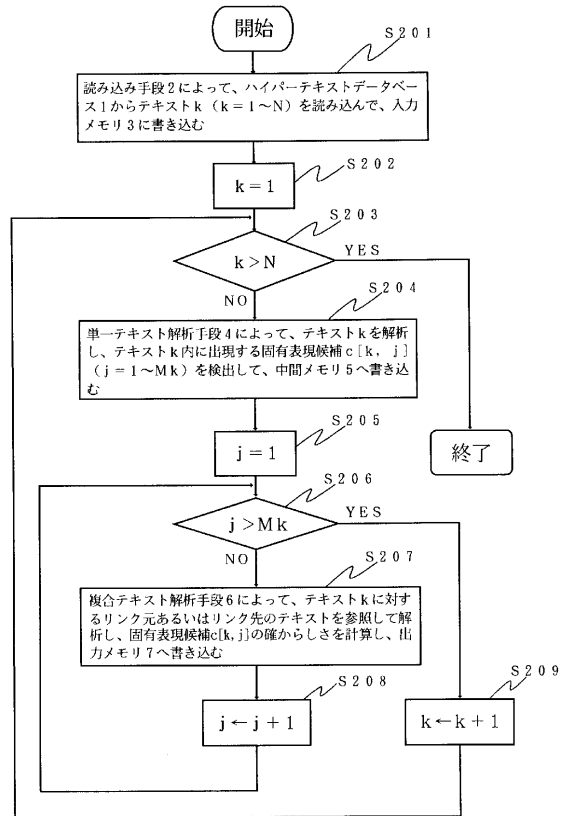
【図 8】



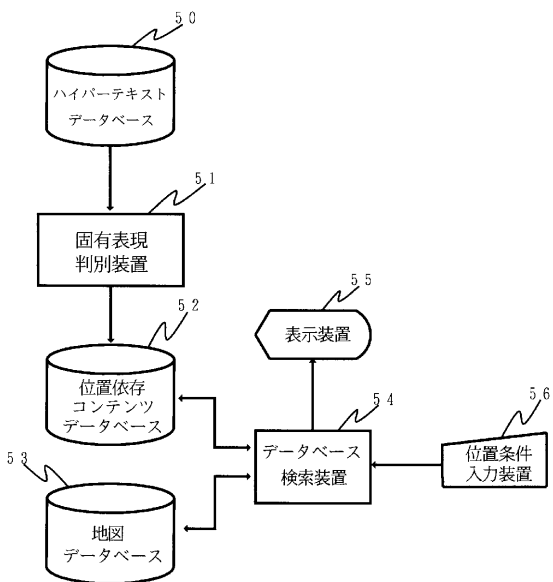
【図 10】



【図 11】



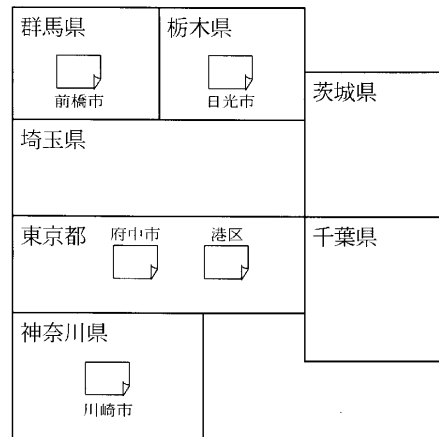
【図 12】



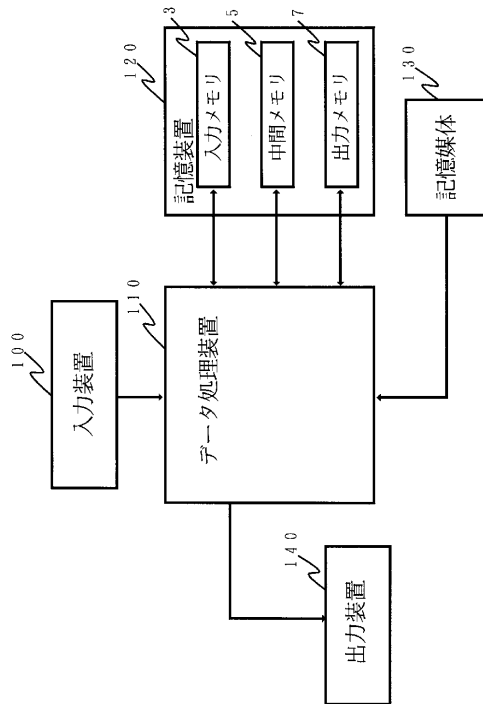
【図 13】

地名表記	ノードページ番号
東京都港区	3 1
東京都府中市	3 9
神奈川県川崎市	4 0
群馬県前橋市	4 0
栃木県日光市	5 2

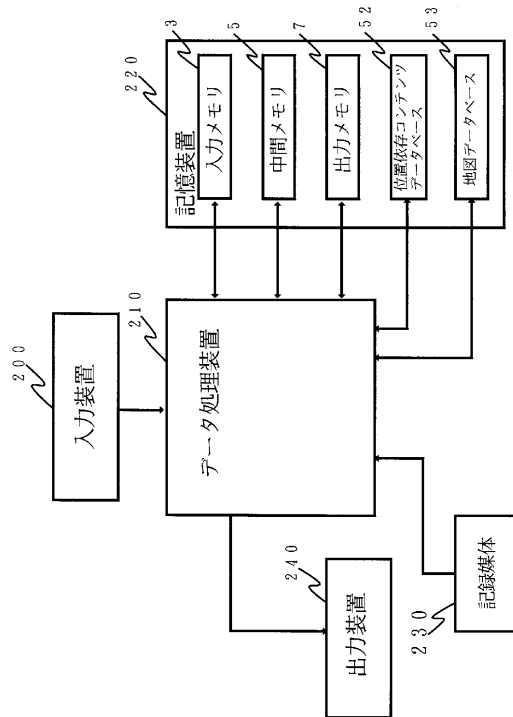
【図 14】



【図 15】



【図 16】



フロントページの続き

(58)調査した分野(Int.Cl. , D B 名)

G06F 17/20-17/28