(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3843320号
(P3843320)
(45)発行日　平成18年11月8日(2006.11.8)　(24)登録日　平成18年8月25日(2006.8.25)

(51)Int.Cl.　　F I
G06F 17/28 (2006.01)　　G06F 17/28　U
G06F 17/30 (2006.01)　　G06F 17/30　220Z

請求項の数 9　(全 32 頁)

(21)出願番号　特願2003-55193(P2003-55193)
(22)出願日　平成15年3月3日(2003.3.3)
(65)公開番号　特開2004-265169(P2004-265169A)
(43)公開日　平成16年9月24日(2004.9.24)
審査請求日　平成15年3月3日(2003.3.3)

特許法第３０条第１項適用　情報処理学会研究報告第２００２巻第１０４号（平成１４年１１月１２日発行）第１０１－１０６ページに発表

特許法第３０条第１項適用　２００２年情報科学技術フォーラム一般講演論文集第２分冊（平成１４年９月１３日発行）第１６３－１６４ページに発表

(73)特許権者　301022471
独立行政法人情報通信研究機構
東京都小金井市貫井北町４－２－１
(74)代理人　100085419
弁理士　大垣　孝
(72)発明者　山本　英子
東京都小金井市貫井北町４丁目２番１号
独立行政法人通信総合研究所内
(72)発明者　井佐原　均
東京都小金井市貫井北町４丁目２番１号
独立行政法人通信総合研究所内
(72)発明者　内山　将夫
東京都小金井市貫井北町４丁目２番１号
独立行政法人通信総合研究所内

最終頁に続く

(54)【発明の名称】特定データの抽出方法、抽出装置、およびプログラム

(57)【特許請求の範囲】
【請求項１】
用語データ抽出手段と重み付け手段と特定データ抽出手段とを有する特定データの抽出装置を用いて、文書データの中から関連性の高い用語データの組み合わせを特定データとして抽出する特定データの抽出方法において、
前記用語データ抽出手段が、複数の文書データの中から各用語データを抽出する用語データ抽出工程と、
前記重み付け手段が、前記用語データ抽出工程で抽出された各用語データに対し、各文書データ中に出現する用語データの数に応じて、用語データ毎に、用語データと各文書データとの相関度を示す値を付与する相関度付与工程と、
前記特定データ抽出手段が、前記相関度付与工程で用語データ毎に付与された前記用語データと各文書データとの相関度を示す値を用いて、以下の式（４）と式（５）に基づいて、２つの多値ベクトルF_gとT_gを算出し、算出した２つの多値ベクトルF_gとT_gを用いて、以下の式（６）に基づいて、重み付き補完類似度を算出することにより、用語データ毎に、他の用語データと組み合わせた場合の重み付き補完類似度を算出する重み付き補完類似度算出工程と、
前記特定データ抽出手段が、前記重み付き補完類似度算出工程で算出した前記重み付き補完類似度が高い順に所定数の用語データの組み合わせを特定データとして抽出する、または、前記重み付き補完類似度算出工程で算出した前記重み付き補完類似度が所定の閾値を超える用語データの組み合わせを特定データとして抽出する特定データ抽出工程と、

を含むことを特徴とする特定データの抽出方法。

【数１】

$$\vec{F_g} = \left\{(f_g)_1, (f_g)_2, \ldots, (f_g)_n\right\} \qquad \left((f_g)_i = 0.0 \ through \ 1.0\right) \quad \ldots (4)$$

$$\vec{T_g} = \left\{(t_g)_1, (t_g)_2, \ldots, (t_g)_n\right\} \qquad \left((t_g)_i = 0.0 \ through \ 1.0\right) \quad \ldots (5)$$

$$Sg\left(\vec{F_g}, \vec{T_g}\right) = \frac{a_g \times d_g - b_g \times c_g}{\sqrt{n \times (T_g)_2 - (T_g)^2}} = \frac{n \times a_g - F_g \times T_g}{\sqrt{n \times (T_g)_2 - (T_g)^2}} \quad \ldots (6)$$

ただし、

$$a_g = \sum_{i=1}^{n} (f_g)_i \times (t_g)_i, \qquad b_g = \sum_{i=1}^{n} \left(1-(f_g)_i\right) \times (t_g)_i,$$

$$c_g = \sum_{i=1}^{n} (f_g)_i \times \left(1-(t_g)_i\right), \qquad d_g = \sum_{i=1}^{n} \left(1-(f_g)_i\right) \times \left(1-(t_g)_i\right),$$

$$F_g = \sum_{i=1}^{n} (f_g)_i, \qquad T_g = \sum_{i=1}^{n} (t_g)_i, \qquad (T_g)_2 = \sum_{i=1}^{n} (t_g)_i^2$$

である。

ただし、ｎは、特定データの抽出の対象である文書データの総数とする。また、ｉは、ベクトルの次元数とする。また、a_g，b_g，c_g，d_gは、文書データｉ中に出現する２つの用語データの数に応じて０～１の間で設定される、用語データと文書データとの相関度とする。

【請求項２】

前記特定データの抽出装置は、前記用語データ抽出手段と前記重み付け手段と前記特定データ抽出手段とＩＤ付与手段とを含む本処理部を有しており、

前記本処理部が、前記用語データ抽出工程で、各用語データに固有のＩＤを付与するとともに、前記ＩＤに前記各文書データとの相関度を示す値を関連付けることを特徴とする請求項１に記載の特定データの抽出方法。

【請求項３】

前記特定データの抽出装置は、さらに、前処理部を有しており、

前記前処理部が、

前記用語データ抽出工程の前工程として、

複数の文書データを取得する文書データ取得工程と、

前記文書データ取得工程で取得された複数の文書データの中から特定データの抽出対象とならない領域のデータを除外する不要データ除外工程と、

前記不要データ除外工程で残された複数の文書データの各々を形態素解析して品詞毎に分類する品詞分類工程と、

を実行することを特徴とする請求項１に記載の特定データの抽出方法。

【請求項４】

前記特定データ抽出手段は、前記特定データ抽出工程で、特定データとして抽出する数、または、前記重み付き補完類似度の閾値を変更することによって、特定データとして抽

出する用語データの組み合わせの数を適宜変更できることを特徴とする請求項１に記載の特定データの抽出方法。

【請求項５】

前記用語データ抽出工程と、前記相関度付与工程と、前記重み付き補完類似度算出工程と、前記特定データ抽出工程とを、

２つの異なる言語によって作成された同じ内容の文書データを対象にして行い、言語毎に抽出された特定データを２つの言語間で比較することによって訳語の関係にある用語データの組み合わせを抽出することを特徴とする請求項１に記載の特定データの抽出方法。

【請求項６】

文書データの中から関連性の高いデータの組み合わせを特定データとして抽出する特定データの抽出装置において、

複数の文書データの中から各用語データを抽出する用語データ抽出手段と、

各文書データ中に出現する用語データの数をカウントするカウント手段と、

前記用語データ抽出手段によって抽出された各用語データに対し、各文書データ中に出現する用語データの数に応じて、用語データ毎に、用語データと各文書データとの相関度を示す値を付与する重み付け手段と、

前記重み付け手段によって用語データ毎に付与された前記用語データと各文書データとの相関度を示す値を用いて、以下の式（４）と式（５）に基づいて、２つの多値ベクトルF_gとT_gを算出し、算出した２つの多値ベクトルF_gとT_gを用いて、以下の式（６）に基づいて、重み付き補完類似度を算出することにより、用語データ毎に、他の用語データとの重み付き補完類似度を算出し、算出した前記重み付き補完類似度が高い順に所定数の用語データの組み合わせを特定データとして抽出する、または、算出した前記重み付き補完類似度が所定の閾値を超える用語データの組み合わせを特定データとして抽出する特定データ抽出手段と、

を有することを特徴とする特定データの抽出装置。

【数２】

$$\vec{F_g} = \left\{(f_g)_1, (f_g)_2, \ldots, (f_g)_n\right\} \qquad \left((f_g)_i = 0.0 \ through \ 1.0\right) \quad \ldots (4)$$

$$\vec{T_g} = \left\{(t_g)_1, (t_g)_2, \ldots, (t_g)_n\right\} \qquad \left((t_g)_i = 0.0 \ through \ 1.0\right) \quad \ldots (5)$$

$$Sg\left(\vec{F_g}, \vec{T_g}\right) = \frac{a_g \times d_g - b_g \times c_g}{\sqrt{n \times (T_g)_2 - (T_g)^2}} = \frac{n \times a_g - F_g \times T_g}{\sqrt{n \times (T_g)_2 - (T_g)^2}} \quad \ldots (6)$$

ただし、

$$a_g = \sum_{i=1}^{n} (f_g)_i \times (t_g)_i, \qquad b_g = \sum_{i=1}^{n} \left(1-(f_g)_i\right) \times (t_g)_i,$$

$$c_g = \sum_{i=1}^{n} (f_g)_i \times \left(1-(t_g)_i\right), \qquad d_g = \sum_{i=1}^{n} \left(1-(f_g)_i\right) \times \left(1-(t_g)_i\right),$$

$$F_g = \sum_{i=1}^{n} (f_g)_i, \qquad T_g = \sum_{i=1}^{n} (t_g)_i, \qquad (T_g)_2 = \sum_{i=1}^{n} (t_g)_i^2$$

である。

ただし、ｎは、特定データの抽出の対象である文書データの総数とする。また、ｉは、ベクトルの次元数とする。また、a_g，b_g，c_g，d_gは、文書データｉ中に出現する２つの用語データの数に応じて０～１の間で設定される、用語データと文書データとの相関度とする。

【請求項７】

前記特定データ抽出手段は、２つの異なる言語によって作成された同じ内容の文書データを対象にして特定データの抽出を行った場合に、言語毎に抽出された特定データを２つの言語間で比較することによって訳語の関係にある用語データの組み合わせを抽出することを特徴とする請求項６に記載の特定データの抽出装置。

【請求項８】

コンピュータを、

複数の文書データの中から各用語データを抽出する用語データ抽出手段と、

各文書データ中に出現する用語データの数をカウントするカウント手段と、

前記用語データ抽出手段によって抽出された各用語データに対し、各文書データ中に出現する用語データの数に応じて、用語データ毎に、用語データと各文書データとの相関度を示す値を付与する重み付け手段と、

前記重み付け手段によって用語データ毎に付与された前記用語データと各文書データとの相関度を示す値を用いて、以下の式（４）と式（５）に基づいて、２つの多値ベクトルF_gとT_gを算出し、算出した２つの多値ベクトルF_gとT_gを用いて、以下の式（６）に基づいて、重み付き補完類似度を算出することにより、用語データ毎に、他の用語データとの重み付き補完類似度を算出し、算出した前記重み付き補完類似度が高い順に所定数の用語データの組み合わせを特定データとして抽出する、または、算出した前記重み付き補完類似度が所定の閾値を超える用語データの組み合わせを特定データとして抽出する特定データ抽出手段として、

機能させるためのプログラム。

【数３】

$$\vec{F}_g = \left\{(f_g)_1, (f_g)_2, \ldots, (f_g)_n\right\} \qquad \left((f_g)_i = 0.0 \ through \ 1.0\right) \quad \ldots (4)$$

$$\vec{T}_g = \left\{(t_g)_1, (t_g)_2, \ldots, (t_g)_n\right\} \qquad \left((t_g)_i = 0.0 \ through \ 1.0\right) \quad \ldots (5)$$

$$Sg\left(\vec{F}_g, \vec{T}_g\right) = \frac{a_g \times d_g - b_g \times c_g}{\sqrt{n \times (T_g)_2 - (T_g)^2}} = \frac{n \times a_g - F_g \times T_g}{\sqrt{n \times (T_g)_2 - (T_g)^2}} \quad \ldots (6)$$

ただし、

$$a_g = \sum_{i=1}^{n} (f_g)_i \times (t_g)_i, \qquad b_g = \sum_{i=1}^{n} \left(1-(f_g)_i\right) \times (t_g)_i,$$

$$c_g = \sum_{i=1}^{n} (f_g)_i \times \left(1-(t_g)_i\right), \qquad d_g = \sum_{i=1}^{n} \left(1-(f_g)_i\right) \times \left(1-(t_g)_i\right),$$

$$F_g = \sum_{i=1}^{n} (f_g)_i, \qquad T_g = \sum_{i=1}^{n} (t_g)_i, \qquad (T_g)_2 = \sum_{i=1}^{n} (t_g)_i^2$$

である。

ただし、ｎは、特定データの抽出の対象である文書データの総数とする。また、ｉは、ベクトルの次元数とする。また、a_g，b_g，c_g，d_gは、文書データｉ中に出現する２つの用語データの数に応じて０～１の間で設定される、用語データと文書データとの相関度とする。

【請求項９】

前記特定データ抽出手段を、２つの異なる言語によって作成された同じ内容の文書データを対象にして特定データの抽出を行った場合に、言語毎に抽出された特定データを２つの言語間で比較することによって訳語の関係にある用語データの組み合わせを抽出する手段として、

機能させるための請求項８に記載のプログラム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

この発明は、文書データの中から関連性の高い用語データの組み合わせを特定データとして抽出する技術に関する。特に、既存の用語辞書からでは抽出することができない用語データの組み合わせ（例えば、ある組織における人名と役職の関係のように、時間的な経過によって変化するデータの組み合わせや、ある事件における場所と人物、その他の関係のように、突発的に発生する用語データによって変化するデータの組み合わせなど）を特定データとして抽出する技術に関する。

【０００２】

【従来の技術】

近年、電子化された各種のデータの中から、あるデータをキーにしてそのデータに関連しているデータを抽出する技術が各種の装置に適用されている。このような装置として、例

えば、文章データの中から関連性の高いデータの組み合わせ（以下、特定データという）を抽出して要約文を作成する要約文作成装置や、特定データの検索、収集、分析作業などを行う検索装置などがある。

【０００３】

これら従来の技術は、特定データを抽出する場合に、予め用意された既存の用語辞書に基づいて抽出していた（例えば、特許文献１参照）。

【０００４】

【特許文献１】

特開平２０００－１３７７２９号公報　（段落００１４～段落００２２、図１、図２）

【０００５】

【発明が解決しようとする課題】

しかしながら、従来の技術は、予め用意された既存の用語辞書に基づいて特定データを抽出するため、既存の用語辞書からでは抽出することができない特定データ（例えば、ある組織における人名と役職の関係のように、時間的な経過によって変化する用語データの組み合わせや、ある事件における場所と人物、その他の関係のように、突発的に発生する現象によって変化する用語データの組み合わせなど）があるという課題があった。

【０００６】

この発明は、前述の課題を解決することができる（すなわち、既存の用語辞書からでは抽出することができない特定データを抽出できる）特定データの抽出方法、抽出装置およびそのプログラムを提供することを目的とする。

【０００７】

【課題を解決するための手段】

前述の課題を解決するために、請求項１に記載の特定データの抽出方法は、用語データ抽出手段と重み付け手段と特定データ抽出手段とを有する特定データの抽出装置を用いて、文書データの中から関連性の高いデータの組み合わせを特定データとして抽出する特定データの抽出方法において、前記用語データ抽出手段が、複数の文書データの中から各用語データを抽出する用語データ抽出工程と、前記重み付け手段が、前記用語データ抽出工程で抽出された各用語データに対し、各文書データ中に出現する用語データの数に応じて、用語データ毎に、用語データと各文書データとの相関度を示す値を付与する相関度付与工程と、前記特定データ抽出手段が、前記相関度付与工程で用語データ毎に付与された前記用語データと各文書データとの相関度を示す値を用いて、以下の式（４）と式（５）に基づいて、２つの多値ベクトルF_gとT_gを算出し、算出した２つの多値ベクトルF_gとT_gを用いて、以下の式（６）に基づいて、重み付き補完類似度を算出することにより、用語データ毎に、他の用語データと組み合わせた場合の重み付き補完類似度を算出する重み付き補完類似度算出工程と、前記特定データ抽出手段が、前記重み付き補完類似度算出工程で算出した前記重み付き補完類似度が高い順に所定数の用語データの組み合わせを特定データとして抽出する、または、前記重み付き補完類似度算出工程で算出した前記重み付き補完類似度が所定の閾値を超える用語データの組み合わせを特定データとして抽出する特定データ抽出工程と、を含むことを特徴とする。

【０００８】

請求項１に記載の発明は、用語データ毎に付与された用語データと各文書データとの相関度を示す値を用いて、以下の式（４）と式（５）に基づいて、２つの多値ベクトルF_gとT_gを算出し、算出した２つの多値ベクトルF_gとT_gを用いて、以下の式（６）に基づいて、重み付き補完類似度を算出することにより、用語データ毎に、他の用語データと組み合わせた場合の重み付き補完類似度を算出する。そのため、請求項１に記載の発明は、キーとなる用語データとの相関度が高い文書データ同士から抽出された用語データの組み合わせ（すなわち、キーとなる用語データが多数出現する文書データ同士から抽出された用語データの組み合わせ）ほど、高い重み付き補完類似度を算出できるので、重み付き補完類似度が高い順に所定数の用語データの組み合わせを特定データとして抽出する、または、重み付き補完類似度が所定の閾値を越える用語データの組み合わせを特定データと

して抽出することにより、既存の用語辞書によらなくても特定データを高精度に抽出することができる。

【００１１】

【数２】

$$\vec{F}_g = \{(f_g)_1, (f_g)_2, \ldots, (f_g)_n\} \qquad ((f_g)_i = 0.0 \ through \ 1.0) \quad \ldots (4)$$

$$\vec{T}_g = \{(t_g)_1, (t_g)_2, \ldots, (t_g)_n\} \qquad ((t_g)_i = 0.0 \ through \ 1.0) \quad \ldots (5)$$

$$Sg(\vec{F}_g, \vec{T}_g) = \frac{a_g \times d_g - b_g \times c_g}{\sqrt{n \times (T_g)_2 - (T_g)^2}} = \frac{n \times a_g - F_g \times T_g}{\sqrt{n \times (T_g)_2 - (T_g)^2}} \quad \ldots (6)$$

ただし、

$$a_g = \sum_{i=1}^{n} (f_g)_i \times (t_g)_i, \qquad b_g = \sum_{i=1}^{n} (1-(f_g)_i) \times (t_g)_i,$$

$$c_g = \sum_{i=1}^{n} (f_g)_i \times (1-(t_g)_i), \qquad d_g = \sum_{i=1}^{n} (1-(f_g)_i) \times (1-(t_g)_i),$$

$$F_g = \sum_{i=1}^{n} (f_g)_i, \qquad T_g = \sum_{i=1}^{n} (t_g)_i, \qquad (T_g)_2 = \sum_{i=1}^{n} (t_g)_i^2$$

である。

【００１２】

ただし、ｎは、特定データの抽出の対象である文書データの総数とする。また、ｉは、ベクトルの次元数とする。また、a_g，b_g，c_g，d_gは、文書データｉ中に出現する２つの用語データの数に応じて０～１の間で設定される、用語データと文書データとの相関度とする。

【００１３】

補完類似度は、主に文字認識の分野で用いられ、劣化印刷文字（すなわち、かすれている文字や汚れている文字など）を高い精度で認識できるように提唱された類似度の尺度である。補完類似度は、例えば、文字を画像特徴（すなわち、特徴的な画像要素の集合体）として扱い、劣化印刷文字の画像パターンとテンプレート文字（すなわち、文字辞書に登録された比較の対象となる文字）の画像パターンとの間で、一致情報と不一致情報の差分をとるパラメータを含む式に基づいて算出される。

【００１４】

ただし、補完類似度は、文字認識の分野に限らず、２つのパターンの類似度を求める場合に、２つのパターンの一致している部分だけに注目して算出する類似度の尺度としても用いられる。この場合、補完類似度は、２つのパターンの間で、一致情報と不一致情報の差分をとるパラメータを含む式に基づいて算出される。

【００１５】

補完類似度は、一般的な類似度が対称性を持つ（すなわち、一般的な類似度が２つのパターンを入れ替えても同じ値になる）のに対して、非対称性を持つ（すなわち、２つのパターンを入れ替えると異なる値になる）。これは、補完類似度を算出するための式が、一致

情報と不一致情報の差分をとるパラメータを含む形式になっているからである。補完類似度は、このような特性を持つため、包含関係を持つパターンに対して高い値をとる傾向にある。そのため、補完類似度は、包含関係を持つパターンを特定する尺度として有効である。

【００１６】

特に、重み付き補完類似度（すなわち、多値による重み付きの値が施されたパラメータを用いて算出された補完類似度）は、各文書データ中に出現する各用語データの数をパラメータに含む式に基いて算出される。そのため、重み付き補完類似度は、重み付きのない補完類似度（すなわち、２値のパラメータを用いて算出された補完類似度）よりも、高精度に特定データを抽出することができる。

【００１７】

また、請求項２に記載の発明に係る特定データの抽出方法は、請求項１に記載の発明に係る特定データの抽出方法において、前記特定データの抽出装置は、前記用語データ抽出手段と前記重み付け手段と前記特定データ抽出手段とＩＤ付与手段とを含む本処理部を有しており、前記本処理部が、前記用語データ抽出工程で、各用語データに固有のＩＤを付与するとともに、前記ＩＤに前記各文書データとの相関度を示す値を関連付けることを特徴とする。

【００１８】

請求項２に記載の発明は、各用語データに固有のＩＤを付与するとともに、前記ＩＤに前記各文書データとの相関度を示す値を関連付けているので、前記重み付き補完類似度算出工程において、用語データ毎の、他の用語データとの重み付き補完類似度の算出を容易に行うことができる。

【００１９】

また、請求項３に記載の発明に係る特定データの抽出方法は、請求項１に記載の発明に係る特定データの抽出方法において、前記特定データの抽出装置は、さらに、前処理部を有しており、前記前処理部が、前記用語データ抽出工程の前工程として、複数の文書データを取得する文書データ取得工程と、前記文書データ取得工程で取得された複数の文書データの中から特定データの抽出対象とならない領域のデータを除外する不要データ除外工程と、前記不要データ除外工程で残された複数の文書データの各々を形態素解析して品詞毎に分類する品詞分類工程と、を実行することを特徴とする。

【００２０】

請求項３に記載の発明は、複数の文書データの中から特定データの抽出対象とならない領域のデータを除外してから各用語データを抽出するので、用語データの抽出を短時間で行うことができる。

【００２１】

また、請求項４に記載の発明に係る特定データの抽出方法は、請求項１に記載の発明に係る特定データの抽出方法において、前記特定データ抽出手段は、前記特定データ抽出工程で、特定データとして抽出する数、または、重み付き補完類似度の閾値を変更することによって、特定データとして抽出する用語データの組み合わせの数を適宜変更できることを特徴とする。

【００２２】

請求項４に記載の発明は、特定データとして抽出する数、または、重み付き補完類似度の閾値を変更することによって、特定データの抽出範囲を広げる、または、狭めることができる。例えば、特定データとして抽出する数を予め設定しておき、その数で一次抽出し、その結果を参照して、抽出範囲を広げるように数を増やしたり、逆に、狭めるように数を減らして、二次抽出することができる。または、重み付き補完類似度の閾値を予め設定しておき、その閾値で一次抽出し、その結果を参照して、抽出範囲を広げるように閾値を下げたり、逆に、狭めるように閾値を上げて、二次抽出することができる。

【００２３】

また、請求項５に記載の発明に係る特定データの抽出方法は、請求項１に記載の発明に

係る特定データの抽出方法において、前記用語データ抽出工程と、前記相関度付与工程と、前記重み付き補完類似度算出工程と、前記特定データ抽出工程とを、２つの異なる言語によって作成された同じ内容の文書データを対象にして行い、言語毎に抽出された特定データを２つの言語間で比較することによって訳語の関係にある用語データの組み合わせを抽出することを特徴とする。

【００２４】

請求項５に記載の発明は、特定データの抽出を、複数組の、２つの異なる言語によって作成された同じ内容の文書データを対象にして行い、言語毎に抽出された特定データを２つの言語間で比較するので、訳語の関係にある用語データの組み合わせを抽出することができる。

【００２５】

また、請求項６に記載の発明に係る特定データの抽出装置は、文書データの中から関連性の高いデータの組み合わせを特定データとして抽出する特定データの抽出装置において、複数の文書データの中から各用語データを抽出する用語データ抽出手段と、各文書データ中に出現する用語データの数をカウントするカウント手段と、前記用語データ抽出手段によって抽出された各用語データに対し、各文書データ中に出現する用語データの数に応じて、用語データ毎に、用語データと各文書データとの相関度を示す値を付与する重み付け手段と、前記重み付け手段によって用語データ毎に付与された前記用語データと各文書データとの相関度を示す値を用いて、上述の式（４）と式（５）に基づいて、２つの多値ベクトルF_gとT_gを算出し、算出した２つの多値ベクトルF_gとT_gを用いて、上述の式（６）に基づいて、重み付き補完類似度を算出することにより、用語データ毎に、他の用語データとの重み付き補完類似度を算出し、算出した前記重み付き補完類似度が高い順に所定数の用語データの組み合わせを特定データとして抽出する、または、算出した前記重み付き補完類似度が所定の閾値を超える用語データの組み合わせを特定データとして抽出する特定データ抽出手段と、を有することを特徴とする。請求項６に記載の発明は、請求項１に記載の発明を実施するための装置の構成を明示している。

【００２６】

また、請求項７に記載の発明に係る特定データの抽出装置は、請求項６に記載の発明に係る特定データの抽出装置において、前記特定データ抽出手段は、２つの異なる言語によって作成された同じ内容の文書データを対象にして特定データの抽出を行った場合に、言語毎に抽出された特定データを２つの言語間で比較することによって訳語の関係にある用語データの組み合わせを抽出することを特徴とする。請求項７に記載の発明は、請求項５に記載の発明を実施するための装置の構成を明示している。

【００２７】

また、請求項８に記載の発明に係るプログラムは、コンピュータにより、請求項６に記載の発明に係る特定データの抽出装置を実現することを特徴とする。請求項８に記載の発明は、プログラムをコンピュータにインストールすることによって、請求項６に記載の発明に係る装置を実現することを明示している。

また、請求項９に記載の発明に係るプログラムは、コンピュータにより、請求項７に記載の発明に係る特定データの抽出装置を実現することを特徴とする。請求項９に記載の発明は、プログラムをコンピュータにインストールすることによって、請求項７に記載の発明に係る装置を実現することを明示している。

【００２８】

【発明の実施の形態】

以下に、図や表を参照してこの発明の実施の形態を説明する。なお、各図および各表は、この発明を理解できる程度に概略的に示してあるに過ぎない。また、各図において、共通する要素については、同一の符号を付与し、説明を省略する。

【００２９】

＜第１の実施の形態＞

この実施の形態は、既存の用語辞書によらなくても特定データを抽出することができるよ

うに、キーとなる用語データとの相関度が高い文書データ同士から抽出された用語データの組み合わせ（すなわち、キーとなる用語データが多数出現する文書データ同士から抽出された用語データの組み合わせ）ほど、高い類似度を算出できるようにし、これによって、算出された類似度が高い順に所定数の用語データの組み合わせを特定データとして、または、算出された類似度が所定の閾値を越える用語データの組み合わせを特定データとして抽出することを技術思想とする。

【００３０】

（抽出装置の構成）

以下に、図１を用いて、この発明に係る特定データの抽出装置の構成を説明する。図１はこの発明に係る抽出装置の構成を示す図である。

【００３１】

図１に示されるように、抽出装置１０は、コンピュータ本体１００と外部機器２００とから構成され、インターネットやＬＡＮ、無線回線などの通信回線網３００を介して、外部のサーバ群４００と接続される。

【００３２】

抽出装置１０には、コンピュータ本体１００を特定データの抽出装置として機能させるためのプログラム（以下、特定データ抽出プログラムという）がインストールされる。これによって、抽出装置１０は、コンピュータ本体１００に内蔵された図示しないＲＡＭをデータ格納部１０１として機能させ、また、図示しないＣＰＵを前処理部１０３および本処理部１０５として機能させ、更に、図示しないハードディスク装置をデータ記憶部１０７として機能させるようになる。なお、データ格納部１０１とは、文書データを含む各種のデータを一時的に格納する部位である。また、前処理部１０３とは、特定データが抽出しやすくなるように文書データに前処理を施す部位である。また、本処理部１０５とは、前処理部１０３によって前処理が施された文書データから特定データを抽出する部位である。また、データ記憶部１０７とは、特定データ抽出プログラムを含む各種のプログラムや用語辞書を含む各種の辞書データなどを記憶するとともに、本処理部１０５によって抽出された特定データを恒久的に記憶する部位である。

【００３３】

係る構成において、前処理部１０３は、不要データ除外手段や、形態素解析手段、品詞付与手段などを有する。なお、不要データ除外手段とは、複数の文章データを対象として、各文章データから不要データ（すなわち、文書データ中の、特定データの抽出対象とならない領域のデータ）を除外する部位である。また、形態素解析手段とは、本文データ（すなわち、文書データ中の、特定データの抽出対象となる領域のデータであり、文書データから不要データを除外することによって残る領域のデータ）の形態素（すなわち、意味を有する最小の言語形態）を解析する部位である。また、品詞付与手段とは、形態素解析手段によって解析された本文データの形態素毎に品詞を付与する部位である。

【００３４】

また、本処理部１０５は、用語データ抽出手段や、ＩＤ付与手段、並べ替え手段、カウント手段、重み付け手段、特定データ抽出手段などを有する。なお、用語データ抽出手段とは、前処理部１０３によって品詞が付与された各用語データを抽出する部位である。また、ＩＤ付与手段とは、用語データを特定するためのＩＤを各用語データに付与する部位である。また、並べ替え手段とは、各用語データを所定の順序に並べ替える部位である。また、カウント手段とは、文章データの数や、各文章データ中に出現する用語データの数などをカウントする部位である。また、重み付け手段とは、各文章データ中に出現する用語データの数に応じて、用語データ毎に、用語データと各文書データとの相関度を示す値（以下、重み付きの値という）を付与する部位である。また、特定データ抽出手段とは、後述の重み付き補完類似度を算出し、重み付け補完類似度が高い順に所定数の用語データの組み合わせを特定データとして、または、重み付き補完類似度が所定の閾値を超える用語データの組み合わせを特定データとして抽出する部位である。

【００３５】

抽出装置１０の外部機器２００は、読取装置２０１や、スキャナ２０３、キーボード２０５、マウス２０７、ディスプレイ２０９、プリンタ２１１、外部記憶装置２１３などである。ここで、読取装置２０１は、図示しない磁気記録媒体（例えばフロッピーディスク（登録商標））や光記録媒体（例えばＣＤ－ＲＯＭやデジタルバーサタイルディスク（ＤＶＤ））などから文書データを読み取るための入力装置である。なお、外部機器２００の構成は、図１に示される構成に限らず、適宜変更することができる。

【００３６】

（抽出装置の動作）

以下に、まず、図２と図３を用いて抽出装置１０の動作の概要について説明し、その後に、図４を用いて抽出装置１０の動作の詳細について説明する。なお、図２と図３は各データの関係を示す図であり、図４は特定データの抽出工程を示すフローチャートである。

【００３７】

まず、抽出装置１０は、複数の文書データの中から各用語データを抽出する。この動作は、具体的には以下の通りである。なお、ここでは、抽出装置１０の動作が分かりやすくなるように、特定データとして名詞の用語データ（以下、名詞データという）の組み合わせを抽出するものとして説明する。

【００３８】

すなわち、例えば、図２（ａ）～図２（ｃ）に示されるように、３つの文書データＤＡ，ＤＢ，ＤＣがあるものとする。抽出装置１０は、３つの文書データＤＡ，ＤＢ，ＤＣの中からそれぞれに含まれる全ての名詞データを抽出する。図２に示される例では、抽出装置１０は、文書データＤＡから２つの名詞データＮＡと２つの名詞データＮＢと１つの名詞データＮＣを抽出し、文書データＤＢから２つの名詞データＮＡと１つの名詞データＮＤを抽出し、文書データＤＣから２つの名詞データＮＡと１つの名詞データＮＣと２つの名詞データＮＤを抽出している。この関係を図３（ａ）に示す。図３（ａ）は、行方向に各名詞データを配置し、列方向に各文書データを配置して、各文章データ中に出現する各名詞データの数をマトリクス状に配置した表である。

【００３９】

次に、抽出装置１０は、各文書データ中に出現する名詞データの数に応じて、名詞データ毎に、重み付きの値（すなわち、各文書データとの相関度を示す値）を付与する。この関係を図３（ｂ）に示す。図３（ｂ）は、行方向に各名詞データを配置し、列方向に各文書データを配置して、重み付きの値ｋ１～ｋ３をマトリクス状に配置した表である。なお、重み付きの値については、後述の「重み付け補完類似度による特定データの抽出」の章で詳述する（「重み付きの値ｗｅｉｇｈｔ（ｔｆ）」に関する説明参照）。

【００４０】

次に、抽出装置１０は、重み付きの値を用いて、名詞データ毎に、他の名詞データと組み合わせた場合の類似度（すなわち、後述の重み付き補完類似度S_g（F_g，T_g））を算出するものとする。このとき算出された類似度を図３（ｃ）に示す。図３（ｃ）は、行方向に各名詞データを配置し、列方向に各文書データを配置して、後述の重み付き補完類似度S_g（F_g，T_g）をマトリクス状に配置した表である。抽出装置１０は、このようにして類似度を算出する。

【００４１】

次に、抽出装置１０は、類似度が高い順に所定数の名詞データの組み合わせを特定データとして、または、類似度が所定の閾値を超える名詞データの組み合わせを特定データとして抽出する。以上が、抽出装置１０の動作の概要である。

【００４２】

以下に、抽出装置１０の動作の詳細について説明する。なお、ここでは、特定データとして、名詞データに限らず、用語データ全般の組み合わせを抽出するものとして説明する。

【００４３】

図４に示されるように、まず、ステップ（以下、Ｓという）１０１において、抽出装置１０の前処理部１０３は、磁気記録媒体や光記録媒体、紙媒体、または外部のサーバ４００

などから複数の文書データを取得し、文章データを特定するためのコード（以下、文章データコードという）を各文章データに付与してデータ格納部１０１に一時格納する。

【００４４】

次に、Ｓ１０３，Ｓ１０５において、抽出装置１０の前処理部１０３は、データ格納部１０１に一時格納された複数の文書データの各々から、前述の不要データを除外して、前述の本文データを抽出する。

【００４５】

次に、Ｓ１０７，Ｓ１０９において、抽出装置１０の前処理部１０３は、複数の文書データの各々から抽出された各本文データの形態素を解析し、各本文データから切り出した各形態素に品詞を付与する。

【００４６】

この後、Ｓ１１１，Ｓ１１３において、抽出装置１０の本処理部１０５は、各本文データの中から各用語データを抽出し（図２参照）、各用語データにＩＤを付与する。なお、Ｓ１１１で抽出する用語データは品詞が名詞となっているものが好ましく、また、Ｓ１１３で付与されるＩＤは前述の文章データコードとリンク付けされるのが好ましい。

【００４７】

次に、Ｓ１１５において、抽出装置１０の本処理部１０５は、各用語データを、例えば英字、数字、その他に分類し、英字、数字、その他の並びの若い順に並べ替える。これによって同じ用語データ毎に集合が形成される。

【００４８】

次に、Ｓ１１７において、抽出装置１０の本処理部１０５は、各集合中の用語データの個数をカウントする（図３（ａ）参照）。

【００４９】

次に、Ｓ１１９において、抽出装置１０の本処理部１０５は、各文章データ中に出現する各用語データの数に応じて、各用語データに所定の重み付きの値（すなわち、各文書データとの相関度を示す値）を付与する（図３（ｂ）参照）。すなわち、前述の文章データコードとリンク付けされた各用語データのＩＤに、所定の重み付きの値（すなわち、各文書データとの相関度を示す値）をリンク付けする。

【００５０】

次に、Ｓ１２１，Ｓ１２３において、抽出装置１０の本処理部１０５は、重み付きの値を用いて、用語データ毎に、他の用語データと組み合わせた場合の類似度（すなわち、後述の重み付き補完類似度S_g（F_g，T_g））を算出する（図３（ｃ）参照）。そして、各用語データの組み合わせの関係を推定する。すなわち、組み合わせの一方の用語データをキーとなる用語データとし、他方の用語データをキーとなる用語データとの組み合わせ対象となる用語データとする場合に、キーとなる用語データから組み合わせ対象となる用語データへのベクトルに基づく重み付き補完類似度と、その逆のベクトルに基づく重み付き補完類似度とを比較する。これによって、キーとなる用語データが、組み合わせの対象となる用語データに対して、包含関係になっているのか、対等関係になっているのか、または被包含関係になっているのかを推定する。また、このとき、抽出装置１０の本処理部１０５は、算出した類似度が高い順に所定数の用語データの組み合わせを特定データとして抽出する。または、抽出装置１０の本処理部１０５は、算出した各用語データの組み合わせの類似度と所定の閾値とを比較し、所定の閾値を超える用語データの組み合わせを特定データとして抽出する。なお、このとき、特定データとして抽出する用語データの組み合わせの数は、特定データとして抽出する数、または、類似度の閾値を変更することによって、適宜変更できるものとする。

【００５１】

次に、Ｓ１２５において、抽出装置１０の本処理部１０５は、各特定データを所定の順番（例えば、類似度の高い順、または、英字、数字、その他の並びの若い順）に並べ替え、データ記憶部１０７に記憶する。ここで、本処理部１０５が各特定データを類似度の高い順に並べ替えた場合、抽出装置１０は関連性の高い用語データの組み合わせを優先的にオ

ペレータに提示することができるようになる。また、本処理部１０５が特定データを英字、数字、その他の並びの若い順に並べ替えた場合、抽出装置１０は先頭の文字の若い用語データの組み合わせを優先的にオペレータに提示することができるようになる。

【００５２】

次に、Ｓ１２７において、抽出装置１０の本処理部１０５は、抽出した特定データを外部機器２００に出力する。すなわち、例えばディスプレイ２０９に出力して表示させたり、外部記憶装置２１３に出力して記憶させたりする。以上が、抽出装置１０の動作の詳細である。

【００５３】

（類似度の算出）

以下に、類似度の算出について説明する。

【００５４】

従来の技術は、特定データとして抽出される用語データ間の関係（すなわち、一方の用語データと他方の用語データとの関係）が一対一の関係にあることを前提として、特定データを抽出していた。しかしながら、用語データ間の関係は一対多の関係にある場合があり、既存の用語辞書からでは抽出できない特定データがあった。このような特定データを抽出するためには、一対多の関係にある用語データ間の関係を検出する必要がある。一対多の関係にある用語データ間の関係は、一方の用語データが他方の用語データに包含されている場合に多く発生する。そこで、この発明では、一方の用語データが他方の用語データに包含されていることを検出できる類似尺度が必要となる。このような類似尺度としては、文字認識の分野で用いられている補完類似度が好適であり、特に、後述の重み付き補完類似度S_g（F_g，T_g）が好適である。

【００５５】

補完類似度（complementary similarity measure）は、主に文字認識の分野で用いられ、劣化印刷文字（すなわち、かすれている文字や汚れている文字など）を高い精度で認識できるように提唱された類似度の尺度である。補完類似度は、例えば、文字を画像特徴（すなわち、特徴的な画像要素の集合体）として扱い、劣化印刷文字の画像パターンとテンプレート文字（すなわち、文字辞書に登録された比較の対象となる文字）の画像パターンとの間で、一致情報と不一致情報の差分をとるパラメータを含む式に基づいて算出される。

【００５６】

補完類似度を算出するための式は具体的には以下の式（３）となる。すなわち、２つの用語データが、それぞれ以下の式（１）と式（２）に基づいて、算出された２つの２値ベクトルＦとＴとする場合における、以下の式（３）であるとき、補完類似度S_c（Ｆ，Ｔ）を算出するための式は以下の式（３）となる。なお、以下の式（１）～（３）に含まれる各パラメータについては、「重み付きのない補完類似度による特定データの抽出」の章で説明する。

【００５７】

【数３】

$$\vec{F} = \{ f_1, f_2, \ldots, f_n \} \quad (f_i = 0 \text{ or } 1) \qquad \ldots (1)$$

$$\vec{T} = \{ t_1, t_2, \ldots, t_n \} \quad (t_i = 0 \text{ or } 1) \qquad \ldots (2)$$

$$S_C(\vec{F}, \vec{T}) = \frac{a \times d - b \times c}{\sqrt{T \times (n-T)}} \qquad \ldots (3)$$

ただし、

$$a = \sum_{i=1}^{n} f_i \times t_i, \qquad b = \sum_{i=1}^{n} (1-f_i) \times t_i,$$

$$c = \sum_{i=1}^{n} f_i \times (1-t_i), \qquad d = \sum_{i=1}^{n} (1-f_i) \times (1-t_i),$$

$$a + b + c + d = n, \qquad T = \sum_{i=1}^{n} t_i$$

である。

【００５８】
補完類似度は、文字の汚れやかすれに強い特徴を持ち、かすれにおいては人の目による認識よりも高い精度を得ることができる。これは、劣化印刷文字の画像パターンがテンプレート文字の画像パターンに包含される形であれば、補完類似度が所定の閾値よりも高い値になり、装置が文字であると認識できるからである。

【００５９】
ただし、補完類似度は、文字認識の分野に限らず、２つのパターンの類似度を求める場合に、２つのパターンの一致している部分だけに注目して算出する類似度の尺度としても用いられる。この場合、補完類似度は、２つパターンの間で、一致情報と不一致情報の差分をとるパラメータを含む式に基づいて算出される。

【００６０】
補完類似度は、一般的な類似度が対称性を持つ（すなわち、一般的な類似度が２つのパターンを入れ替えても同じ値になる）のに対して、非対称性を持つ（すなわち、２つのパターンを入れ替えると異なる値になる）。これは、補完類似度を算出するための式が、一致情報と不一致情報の差分をとるパラメータを含む形式になっているからである。補完類似度は、このような特性を持つため、包含関係を持つパターンに対して高い値をとる傾向にある。そのため、補完類似度は、包含関係を持つパターンを特定する尺度として有効であり、高精度に特定データを抽出することができる。

【００６１】
特に、重み付き補完類似度（すなわち、多値による重み付きの値が施されたパラメータを用いて算出された補完類似度）は、各文書データ中に出現する各用語データの数をパラメータに含む式に基いて算出される。そのため、重み付き補完類似度は、重み付きのない補完類似度（すなわち、２値のパラメータを用いて算出された補完類似度）よりも高精度に特定データを抽出することができる。

【００６２】
重み付き補完類似度を算出するための式は具体的には以下の式（６）となる。すなわち、２つの用語データが、それぞれ以下の式（４）と式（５）に基づいて、前記用語データと各文書データとの相関度を示す値を用いて算出された２つの多値ベクトルF_gとT_gであるとき、補完類似度S_g（F_g，T_g）を算出するための式は以下の式（６）となる。なお、以下の式（４）～（６）に含まれる各パラメータについては、「重み付き補完類似度による特定データの抽出」の章で説明する。
【００６３】
【数４】

$$\vec{F}_g = \left\{(f_g)_1, (f_g)_2, \ldots, (f_g)_n\right\} \qquad \left((f_g)_i = 0.0 \ through \ 1.0\right) \quad \ldots (4)$$

$$\vec{T}_g = \left\{(t_g)_1, (t_g)_2, \ldots, (t_g)_n\right\} \qquad \left((t_g)_i = 0.0 \ through \ 1.0\right) \quad \ldots (5)$$

$$Sg\left(\vec{F}_g, \vec{T}_g\right) = \frac{a_g \times d_g - b_g \times c_g}{\sqrt{n \times (T_g)_2 - (T_g)^2}} = \frac{n \times a_g - F_g \times T_g}{\sqrt{n \times (T_g)_2 - (T_g)^2}} \quad \ldots (6)$$

ただし、

$$a_g = \sum_{i=1}^{n} (f_g)_i \times (t_g)_i, \qquad b_g = \sum_{i=1}^{n} \left(1-(f_g)_i\right) \times (t_g)_i,$$

$$c_g = \sum_{i=1}^{n} (f_g)_i \times \left(1-(t_g)_i\right), \qquad d_g = \sum_{i=1}^{n} \left(1-(f_g)_i\right) \times \left(1-(t_g)_i\right),$$

$$F_g = \sum_{i=1}^{n} (f_g)_i, \qquad T_g = \sum_{i=1}^{n} (t_g)_i, \qquad (T_g)_2 = \sum_{i=1}^{n} (t_g)_i^2$$

である。

【００６４】
以下に、重み付きのない補完類似度と重み付き補完類似度とを説明し、これによって、重み付き補完類似度が重み付きのない補完類似度よりも高精度に特定データを抽出できる理由を述べる。
【００６５】
（重み付きのない補完類似度による特定データの抽出）
以下に、図５を用いて、重み付きのない補完類似度について説明する。なお、図５は、文書データ中における用語データの出現パターンを示す図であり、複数の文書データ中における用語データの出現パターンを２値ベクトルで表わしたものである。なお、ここでは、主に文字認識の分野で用いられている補完類似度を、複数の文書データの中から特定データを抽出するための類似尺度として用いる。そのため、補完類似度を算出するための式において、画像パターンの２値ベクトルを、複数の文書データの中に出現する用語データの出現パターンのベクトルに置き換えるものとする。これにより、用語データ間の一対多の関係を推定できる形式（すなわち、特定データを抽出するのに適した形式）とした。また、以下の説明では、複数の文書データを単にコーパス（corpus）という場合がある。コーパスとは、電子化された大量の音声・言語データのことである。

【００６６】
図５は、「東京」，「大阪」，「名古屋」などの用語データが含まれた複数の文書データ１～ｎを対象として、各文書データの中に該当の用語データが存在する場合に、用語データと文書データとの相関度を示す値を「１」とし、文書データの中に該当の用語データが存在しない場合に、用語データと文書データとの相関度を示す値を「０」とすることによって形成された、各用語データ「東京」，「大阪」，「名古屋」の出現パターンを示している。
【００６７】
抽出装置１０の本処理部１０５は、図５に示される出現パターンに基づいて、以下のようにして重み付きのない補完類似度を算出する。
【００６８】
まず、抽出装置１０の本処理部１０５は、用語データにＩＤを付与する。このとき、例えば用語データが「東京」であれば、「東京」を特定するコード（例えば「０００１ａ」）と図５に示される用語データ「東京」の出現パターン「１０１１…１」とを合成した形式でＩＤを作成する。また、同様に、例えば用語データが「大阪」であれば、「大阪」を特定するコード（例えば「００１０ｂ」）と図５に示される用語データ「大阪」の出現パターン「０１０１…１」とを合成した形式でＩＤを作成する。また、同様に、例えば用語データが「名古屋」であれば、「名古屋」を特定するコード（例えば「１００１ｘ」）と図５に示される用語データ「名古屋」の出現パターン「００１０…１」とを合成した形式でＩＤを作成する。
【００６９】
次に、抽出装置１０の本処理部１０５は、各用語データを、分類して所定の順番に並べ替え、各文章データ中の用語データの有無を検出する。
【００７０】
次に、抽出装置１０の本処理部１０５は、用語データが文章データ中に出現する場合は用語データと文章データとの相関度を示す値を「１」とし、用語データが文章データ中に出現しない場合は用語データと文章データとの相関度を示す値を「０」とする。
【００７１】
次に、抽出装置１０の本処理部１０５は、２つの用語データが、それぞれ以下の式（１）と式（２）に基づいて、算出された２つの２値ベクトルＦとＴであるものとし、以下の式（３）に基づいて重み付きのない補完類似度S_c（Ｆ，Ｔ）を算出して、関係を推定し、特定データを抽出する。
【００７２】
【数５】

$$\vec{F} = \{ f_1, f_2, \ldots, f_n \} \quad (f_i = 0 \text{ or } 1) \qquad \ldots (1)$$

$$\vec{T} = \{ t_1, t_2, \ldots, t_n \} \quad (t_i = 0 \text{ or } 1) \qquad \ldots (2)$$

$$S_c(\vec{F}, \vec{T}) = \frac{a \times d - b \times c}{\sqrt{T \times (n-T)}} \qquad \ldots (3)$$

ただし、

$$a = \sum_{i=1}^{n} f_i \times t_i, \qquad b = \sum_{i=1}^{n} (1-f_i) \times t_i,$$

$$c = \sum_{i=1}^{n} f_i \times (1-t_i), \qquad d = \sum_{i=1}^{n} (1-f_i) \times (1-t_i),$$

$$a + b + c + d = n, \qquad T = \sum_{i=1}^{n} t_i$$

である。

【００７３】
なお、ここでは、２つの用語データをそれぞれＴｈｇ１，Ｔｈｇ２とする。また、式（３）におけるベクトルの次元数ｎは、特定データの抽出の対象である文書データの総数とする。

【００７４】
また、式（３）におけるパラメータａ，ｂ，ｃ，ｄは、コーパスにおける以下のような用語データ間の関係を示す情報とする。すなわち、パラメータａは、用語データＴｈｇ１，Ｔｈｇ２がどちらも出現する文書データの数とする。また、パラメータｂは、用語データＴｈｇ１は出現しないが、用語データＴｈｇ２は出現する文書データの数とする。また、パラメータｃは、用語データＴｈｇ１は出現するが、用語データＴｈｇ２は出現しない文書データの数とする。また、パラメータｄは、用語データＴｈｇ１，Ｔｈｇ２がどちらも出現しない文書データの数とする。これは、用語データと文書データとの相関度を示す値は、用語データが文書データｉに出現する場合に「１」とし、用語データが文書ｉに出現しない場合に「０」とすることを意味する。

【００７５】
このような式（３）に基づいて算出された補完類似度S_c（Ｆ，Ｔ）は、一方の用語データが他方の用語データに包含される形であれば所定の閾値よりも高い値になる。そのため、２つの用語データＴｈｇ１，Ｔｈｇ２間の関係（すなわち、一致の関係や、包含の関係、無関係）は、補完類似度S_c（Ｆ，Ｔ）を用いることによって推定することができる。

【００７６】
したがって、抽出装置１０の本処理部１０５は、補完類似度S_c（Ｆ，Ｔ）を用いることにより、一対多の関係であると認識できるので、特定データを高精度に抽出することができる。

【００７７】
（重み付き補完類似度による特定データの抽出）
以下に、図６を用いて、重み付き補正類似度について説明する。なお、図６は、文書デー

タ中における用語データの出現パターンを示す図である。

【００７８】

図６は、例えば、「東京」，「大阪」，「名古屋」などの用語データが含まれた複数の文書データ１～ｎ中における、各用語データ「東京」，「大阪」，「名古屋」の多値ベクトルによる出現パターンを表わしている。なお、出現パターンは、各文書データ中に出現する個々の用語データの数に応じて、用語データ毎に、用語データと文書データとの相関度を示す値が０～１の間の所定の値に設定されることによって形成される。

【００７９】

抽出装置１０の本処理部１０５は、図６に示される出現パターンに基づいて、以下のようにして重み付き補完類似度を算出する。

【００８０】

まず、抽出装置１０の本処理部１０５は、用語データにＩＤを付与する（図４のＳ１１３参照）。このとき、例えば用語データが「東京」であれば、「東京」を特定するコード（例えば「０００１ａ」）と図６に示される用語データ「東京」の出現パターン「０．５＋０＋０．７＋０．５＋…＋０．５」とを合成した形式でＩＤを作成する。また、同様に、例えば用語データが「大阪」であれば、「大阪」を特定するコード（例えば「００１０ｂ」）と図５に示される用語データ「大阪」の出現パターン「０＋０．５＋０＋０．７＋…＋０．７」とを合成した形式でＩＤを作成する。また、同様に、例えば用語データが「名古屋」であれば、「名古屋」を特定するコード（例えば「１００１ｘ」）と図５に示される用語データ「名古屋」の出現パターン「０＋０＋０．５＋０＋…＋０．７」とを合成した形式でＩＤを作成する。

【００８１】

次に、抽出装置１０の本処理部１０５は、各用語データを、分類して所定の順番に並べ替え、各文章データ中の用語データの個数をカウントする（図４のＳ１１５，Ｓ１１７参照）。

【００８２】

次に、抽出装置１０の本処理部１０５は、各文章データ中に出現する各用語データの数に応じて、用語データと文章データとの相関度を示す値を、所定の重み付きの値（すなわち、各文書データとの相関度を示す値）とする（図４のＳ１１９参照）。

【００８３】

ここで、以下に、重み付きの値について説明する。

【００８４】

重み付きの値ｗｅｉｇｈｔ（ｔｆ）は、関係を推定する対象となる用語データの文書内頻度ＴＦに基づいて算出されることが好ましく、例えば、用語データが所定の回数出現する文書の数の割合とする。具体的には以下のようにして算出される。

【００８５】

ここでは、文書データに含まれる固有用語や一般用語を、対象の用語データＴｈｇｊとする。また、文書内頻度を、０回，１回，…，ｍ回以上の（ｍ＋１）段階でカウントするものとする。ここでは、例えば０回，１回，２回，３回以上の４段階でカウントするものとする。また、各用語データＴｈｇｊが出現する全ての文書数をｄｆ（Ｔｈｇｊ）とし、各用語データＴｈｇｊが１回だけ出現する文書数をｄｆ１（Ｔｈｇｊ）とし、各用語データＴｈｇｊが２回だけ出現する文書数をｄｆ２（Ｔｈｇｊ）とする。

【００８６】

このような定義において、全ての用語データｗｊ（ただし、１≦ｊ≦ｍ）を対象とする重み付きの値ｗｅｉｇｈｔ（ｔｆ）は、ｔｆ＝０の場合（すなわち、各用語データＴｈｇｊが出現しない場合）に以下の式（７）によって算出された値（すなわち０）となり、ｔｆ＝１の場合（すなわち、各用語データＴｈｇｊが１回だけ出現する場合）に以下の式（８）によって算出された値となり、ｔｆ＝２の場合（すなわち、各用語データＴｈｇｊが１～２回出現する場合）に以下の式（９）によって算出された値となり、ｔｆ≧３の場合（すなわち、各用語データＴｈｇｊが１回でも出現する場合）に以下の式（１０）によって

算出された値（すなわち１）となる。

【００８７】

【数６】

$$weight(tf) = 0 \quad \ldots (7)$$

$$weight(tf) = \frac{\sum_{j=1}^{m} df1(w_j)/df(w_j)}{m} \quad \ldots (8)$$

$$weight(tf) = \frac{\sum_{j=1}^{m} (df1(w_j) + df2(w_j))/df(w_j)}{m} \quad \ldots (9)$$

$$weight(tf) = \frac{\sum_{j=1}^{m} df(w_j)/df(w_j)}{m} = 1 \quad \ldots (10)$$

【００８８】

例えば、新聞会社Ａが１９９１－２００１年に発行した新聞記事を電子データとして記録したＣＤ－ＲＯＭから新聞記事の文書データをコーパスとして用いるものとする。

【００８９】

その結果、重み付きの値ｗｅｉｇｈｔ（ｔｆ）は、例えば、ｔｆ＝０の場合に０となり、ｔｆ＝１の場合に０．８４となり、ｔｆ＝２の場合に０．９５となり、ｔｆ≧３の場合に１となったものとする。

【００９０】

重み付きの値ｗｅｉｇｈｔ（ｔｆ）は用語データと文書データｉとの相関度を示す値である。そのため、各用語データのベクトル要素$(f_g)_i$，$(t_g)_i$には、重み付きの値が付与される。例えば、３つの文書データｘ，ｙ，ｚがあり、文書データｘに用語データ「大阪」は２回、用語データ「東京」は１回現れ、文書データｙに用語データ「大阪」は０回、用語データ「東京」は２回現れ、文書データｚに用語データ「大阪」は１回、用語データ「東京」は４回現れたとする。このとき、文書データｘ，ｙ，ｚにおいて、用語データ「大阪」と用語データ「東京」のベクトル要素$(f_g)_x$，$(f_g)_y$，$(f_g)_z$（または$(t_g)_x$，$(t_g)_y$，$(t_g)_z$）には以下の表１に示される重み付きの値が付与される。

【００９１】

【表１】

文書	X	y	z
大阪	0.95	0	0.84
東京	0.84	0.95	1

【００９２】

次に、抽出装置１０の本処理部１０５は、２つの用語データが、それぞれ以下の式（４）と式（５）に基づいて、前記用語データと各文書データとの相関度を示す値を用いて算出された２つの多値ベクトルF_gとT_gであるものとし、以下の式（６）に基づいて重み付き

補完類似度Ｓｇ（Ｆｇ，Ｔｇ）を算出して、関係を推定し、特定データを抽出する（図４のＳ１２１，Ｓ１２３参照）。

【００９３】

【数７】

$$\vec{F}_g = \{(f_g)_1, (f_g)_2, \ldots, (f_g)_n\} \qquad ((f_g)_i = 0.0 \ through \ 1.0) \quad \ldots (4)$$

$$\vec{T}_g = \{(t_g)_1, (t_g)_2, \ldots, (t_g)_n\} \qquad ((t_g)_i = 0.0 \ through \ 1.0) \quad \ldots (5)$$

$$Sg(\vec{F}_g, \vec{T}_g) = \frac{a_g \times d_g - b_g \times c_g}{\sqrt{n \times (T_g)_2 - (T_g)^2}} = \frac{n \times a_g - F_g \times T_g}{\sqrt{n \times (T_g)_2 - (T_g)^2}} \quad \ldots (6)$$

ただし、

$$a_g = \sum_{i=1}^{n} (f_g)_i \times (t_g)_i, \qquad b_g = \sum_{i=1}^{n} (1-(f_g)_i) \times (t_g)_i,$$

$$c_g = \sum_{i=1}^{n} (f_g)_i \times (1-(t_g)_i), \qquad d_g = \sum_{i=1}^{n} (1-(f_g)_i) \times (1-(t_g)_i),$$

$$F_g = \sum_{i=1}^{n} (f_g)_i, \qquad T_g = \sum_{i=1}^{n} (t_g)_i, \qquad (T_g)_2 = \sum_{i=1}^{n} (t_g)_i^2$$

である。

【００９４】

なお、ここでは、２つの用語データをそれぞれＴｈｇ１，Ｔｈｇ２とする。また、式（６）におけるベクトルの次元数ｎは、特定データの抽出の対象である文書データの総数とする。

【００９５】

また、式（６）におけるパラメータａｇ，ｂｇ，ｃｇ，ｄｇは、コーパスにおける以下のような用語データ間の関係を示す情報とする。すなわち、パラメータａｇ，ｂｇ，ｃｇ，ｄｇは、文書データｉ中に出現する用語データＴｈｇ１，Ｔｈｇ２の数に応じて０～１の間で設定される、所定の重み付きの値（すなわち、用語データと文書データとの相関度）とする。

【００９６】

このような式（６）に基づいて算出された重み付き補完類似度Ｓｇ（Ｆｇ，Ｔｇ）は、各文書データ中に出現する各用語データの数をパラメータに含む式に基いて算出される。そのため、重み付き補完類似度Ｓｇ（Ｆｇ，Ｔｇ）は、重み付きのない補完類似度Ｓｃ（Ｆ，Ｔ）よりも、高精度に特定データを抽出することができる。

【００９７】

例えば、重み付きのない補完類似度Ｓｃ（Ｆ，Ｔ）と重み付き補完類似度Ｓｇ（Ｆｇ，Ｔｇ）の適合率は、以下の表２に示す通りとなる。

【００９８】

【表２】

コーパス	重み付きのない補完類似度		重み付き補完類似度	コーパス	重み付きのない補完類似度		重み付き補完類似度
1991	44.4	<	44.5	1998a	47.1	<	47.3
1992	48.3	<	48.4	1998b	45.7	<	45.8
1993	50.9	>	50.7	1999a	47.3	>	46.8
1994	47.8	<	48.1	1999b	48.8	<	49.1
1995	40.3	<	40.5	2000a	46.2	>	46.1
1996	47.3	=	47.3	2000b	46.7	<	46.9
1997	43.0	<	43.6	2001	44.6	<	45.2

【００９９】
なお、表２は、前述の新聞会社Ａが１９９１－２００１年に発行した新聞記事を電子データとして記録したＣＤ－ＲＯＭから新聞記事の文書データをコーパスとして用いた場合の適合率を示している。このコーパスにおいて、重み付きの値ｗｅｉｇｈｔ（ｔｆ）は、例えば、ｔｆ＝０の場合に０となり、ｔｆ＝１の場合に０．８４となり、ｔｆ＝２の場合に０．９５となり、ｔｆ≧３の場合に１となっている。

【０１００】
表２は、重み付きのない補完類似度S_c（Ｆ，Ｔ）を用いて抽出された特定データと重み付き補完類似度S_g（F_g，T_g）を用いて抽出された特定データとを対象にして適合率の評価した結果を示している。なお、適合率の評価は、例えば、類似度の値が高い上位１０００対の特定データを対象にして、特定データが既存の用語辞書に収録されている場合を正解（すなわち、用語データ間に何らかの関係がある）と見なすことによって行った。

【０１０１】
表２に示されるように、適合率は、１４個のコーパスのうち、１０個のコーパスで、重み付き補完類似度S_g（F_g，T_g）の方が重み付きのない補完類似度S_c（Ｆ，Ｔ）よりも高くなった。このように、適合率は、重み付き補完類似度S_g（F_g，T_g）の方が重み付きのない補完類似度S_c（Ｆ，Ｔ）よりも高くなる傾向がある。

【０１０２】
なお、適合率は、１４個のコーパスのうち、１個のコーパスで、重み付きのない補完類似度S_c（Ｆ，Ｔ）と重み付き補完類似度S_g（F_g，T_g）が同じになり、３個のコーパスで、重み付きのない補完類似度S_c（Ｆ，Ｔ）の方が重み付き補完類似度S_g（F_g，T_g）よりも高くなった。しかしながら、このような現象は、主語を省略する傾向にあるコーパスや、「これ」や「あれ」などの関係代用語を利用する傾向にあるコーパスで発生したものである。このようなコーパスに対して、省略された主語や関係代用語を考慮して適合率を算出すれば（例えば、省略された主語を推定する、関係代用語を除外する、または、関係代用語を他の用語に置き換えて適合率を算出すれば）、重み付き補完類似度S_g（F_g，T_g）の方が重み付きのない補完類似度S_c（Ｆ，Ｔ）よりも高くなる。

【０１０３】
また、例えば学術論文や特許出願明細書などのように、主語を省略しない傾向にあるコーパスや関係代用語を利用しない傾向にあるコーパスを対象にすれば、適合率は、重み付き補完類似度S_g（F_g，T_g）の方が重み付きのない補完類似度S_c（Ｆ，Ｔ）よりも高くなる。

【０１０４】
このように、適合率は、重み付き補完類似度S_g（F_g，T_g）の方が重み付きのない補完類似度S_c（Ｆ，Ｔ）よりも高くなる。以下に、その理由を説明する。

【０１０５】

重み付きのない補完類似度S_c（Ｆ，Ｔ）は、２値ベクトルで表わされる用語データの出現パターンに基づいて算出されるものであり、コーパスにおける用語データが出現した文書データの数（すなわち、文書頻度ＤＦ（Document Frequency））を用いて算出される。

【０１０６】

これに対して、重み付き補完類似度S_g（F_g，T_g）は、多値ベクトルで表わされる用語データの出現パターンに基づいて算出されるものであり、文書頻度に関する情報量を表わすＩＤＦ（Inverse Document Frequency）と文書内頻度（すなわち、各文書データ中において用語データが出現する数）を表わすＴＦ（Term Frequency）との内積ＩＤＦ・ＴＦを用いて算出される。

【０１０７】

ところで、一般に、例えば文書データの主題となるような、重要な用語データは、文書データ内に繰り返し現れる傾向がある。そのため、用語データの文書内頻度は、用語データの検索や用語データの抽出において重要な情報源となる。

【０１０８】

したがって、特定データの抽出は、用語データの文書内頻度を考慮した出現パターンを用いた重み付き補完類似度の方が重み付きのない補完類似度よりも、一対多の関係の関係を高精度に検出することができる。

【０１０９】

ところで、前述の新聞会社Ａが１９９１－２００１年に発行した新聞記事を電子データとして記録したＣＤ－ＲＯＭから新聞記事の文書データをコーパスとした場合において、例えば、重み付き補完類似度S_g（F_g，T_g）の値が高い上位２６対の特定データを以下の表３に示す。

【０１１０】

【表３】

類似度	事物1	事物2	抽出結果
11291.890	同時多発テロ	アフガン	*****
11124.555	小泉純一郎	小泉首相	
9310.221	選挙	参院選	
8587.430	官房長官	福田康夫	*****
7042.404	同時多発テロ	ウサマ・ビンラディン	*****
6733.389	選挙	選挙区	
6615.725	ファックス	Eメール	
6343.695	財務省	塩川正十郎	*****
5948.384	選挙	比例代表	
5869.183	選挙	投開票	
5769.533	株式市場	平均株価	
5726.595	訴訟	損害賠償	
5563.195	同時多発テロ	報復	*****
5559.984	ウサマ・ビンラディン	ビンラディン	
5455.257	米大リーグ	ア・リーグ	
5318.772	選挙	立候補	
5237.116	経済財政担当者	竹中平蔵	*****
5215.294	同時多発テロ	空爆	*****
5176.856	厚生労働省	厚労省	
5176.104	選挙	当選	
5155.707	TOPIX	東証株価指数	
5048.065	米大リーグ	ナ・リーグ	
4985.615	狂牛病	肉骨粉	
4787.469	扇千景	国土交通省	*****
4775.770	経済産業省	平沼赳夫	*****
4583.608	選挙	市長選	

【０１１１】

なお、ここでは、抽出装置１０の本処理部１０５は、重み付け補完類似度S_g（F_g，T_g）の値が高い上位２６対の用語データの組み合わせを特定データとして抽出しているが、以下のようにして、重み付け補完類似度S_g（F_g，T_g）が所定の閾値を超える用語データの組み合わせを特定データとして抽出することもできる。

【０１１２】

例えば、まず、抽出装置１０の本処理部１０５は、全用語データの重み付け類似度S_g（F_g，T_g）の総計を算出し、総計の数パーセントに相当する値を閾値とする。次に、抽出装置１０の本処理部１０５は、例えば表３に示されるように、重み付け補完類似度S_g（F_g，T_g）の値が高い順に用語データの組み合わせを並べる。次に、抽出装置１０の本処理部１０５は、上位から下位方向に、各用語データの組み合わせに対応する重み付け補完類似度S_g（F_g，T_g）の合計を順次算出し、算出した合計と閾値とを比較する。そして

、合計が閾値を超えた場合に、抽出装置１０の本処理部１０５は、そのときよりも上位にある用語データの組み合わせを特定データとして抽出する。

【０１１３】

表３中、マーク「＊＊＊＊＊」を付与した特定データは、既存の用語辞書からでは抽出できない、最新の用語データの組み合わせである。このような特定データは、例えば、ある組織における人名と役職の関係のように、時間的な経過によって変化するデータの組み合わせや、ある事件における場所と人物、その他の関係のように、突発的に発生する用語データによって変化するデータの組み合わせなどである。

【０１１４】

したがって、このように、重み付き補完類似度S_g（F_g，T_g）は、重み付きのない補完類似度S_c（Ｆ，Ｔ）よりも、既存の辞書にない関係を抽出するのに有効である。特に、独自の用語が存在する、医学や法学などの専門性の高い分野のコーパスや、著者の癖によって独特な言い回しや造語が存在する、論文や小説などのコーパスを対象とする場合に、有効である。

【０１１５】

＜第２の実施の形態＞

この実施の形態は、英語と日本語のように、異なる言語間における訳語の関係にある特定データを抽出するものである。特に、例えば、医学や法学など、専門性の高い分野では、その分野独自の用語が存在する。また、論文や小説では、著者の癖によって、独特な用語の言い回しや造語が存在する。この実施の形態によれば、分野や著者に応じて、どの用語がどのような形に訳される傾向にあるのかを抽出することができる。

【０１１６】

（抽出装置の構成）

抽出装置１０は、第１の実施の形態と同様の構成をしている。ただし、抽出装置１０は、用語辞書として原文に用いられている言語（以下、第１の言語という）用のものと翻訳文に用いられている言語（以下、第２の言語という）用のものとを有しており、これにより、原文に対して第１の言語用の用語辞書を用いて特定データを抽出することができるとともに、翻訳文に対して第２の言語用の用語辞書を用いて特定データを抽出することができる。また、抽出装置１０は、第１の言語と第２の言語との翻訳辞書を有しており、これにより、言語間で特定データ同士を比較することができるとともに、訳語の関係にある用語データを抽出することができる。

【０１１７】

（抽出装置の動作）

以下に、図７と図８を用いて抽出装置１０の動作の詳細について説明する。なお、図７と図８は訳語の関係にある品詞データの抽出工程を示すフローチャートである。

【０１１８】

まず、抽出装置１０は、第１の言語によって作成された原文データに対して、以下の処理を行う。なお、以下に説明するＳ６０１～Ｓ６２７において、Ｓ６０１～Ｓ６２５は、第１の実施の形態におけるＳ１０１～Ｓ１２５と同様である（ただし、Ｓ１１１に相当する工程はない）。

【０１１９】

すなわち、図７に示されるように、Ｓ６０１において、抽出装置１０の前処理部１０３は、磁気記録媒体や光記録媒体、紙媒体、または外部のサーバ４００などから第１の言語による複数の文書データを取得し、文章データ毎に文章データコードを付与してデータ格納部１０１に一時格納する。

【０１２０】

次に、Ｓ６０３，Ｓ６０５において、抽出装置１０の前処理部１０３は、データ格納部１０１に一時格納された複数の文書データの各々から、前述の不要データを除外して、前述の本文データを抽出する。

【０１２１】

次に、Ｓ６０７，Ｓ６０９において、抽出装置１０の前処理部１０３は、複数の文書データの各々から抽出された各本文データの形態素を解析し、各本文データから切り出した各形態素に品詞を付与する。

【０１２２】

この後、Ｓ６１３において、抽出装置１０の本処理部１０５は、各本文データの中から各用語データを抽出し（図２参照）、各用語データにＩＤを付与する。このＩＤは、前述の文章データコードとリンク付けされるのが好ましい。

【０１２３】

次に、Ｓ６１５において、抽出装置１０の本処理部１０５は、各用語データを、例えば英字、数字、その他に分類し、英字、数字、その他の並びの若い順に並べ替える。これによって同じ用語データ毎に集合が形成される。

【０１２４】

次に、Ｓ６１７において、抽出装置１０の本処理部１０５は、各集合中の用語データの個数をカウントする（図３（ａ）参照）。

【０１２５】

次に、Ｓ６１９において、抽出装置１０の本処理部１０５は、各文章データ中に出現する各用語データの数に応じて、各用語データに所定の重み付きの値（すなわち、各文書データとの相関度を示す値）を付与する（図３（ｂ）参照）。すなわち、前述の文章データコードとリンク付けされた各用語データのＩＤに、所定の重み付きの値（すなわち、各文書データとの相関度を示す値）をリンク付けする。

【０１２６】

次に、Ｓ６２１，Ｓ６２３において、抽出装置１０の本処理部１０５は、重み付きの値をパラメータとして持つ前述の式（６）に基づいて、用語データ毎に、他の用語データと組み合わせた場合の類似度S_g（F_g，T_g）を算出する（図３（ｃ）参照）。そして、各用語データの組み合わせの関係を推定する。また、このとき、抽出装置１０の本処理部１０５は、算出した類似度が高い順に所定数の用語データの組み合わせを特定データとして抽出する。または、抽出装置１０の本処理部１０５は、算出した各用語データの組み合わせの類似度と所定の閾値とを比較し、所定の閾値を超える用語データの組み合わせを第１の言語による特定データとして抽出する。

【０１２７】

次に、Ｓ６２５において、抽出装置１０の本処理部１０５は、各特定データを所定の順番（例えば、類似度の高い順、または、英字、数字、その他の並びの若い順）に並べ替え、データ記憶部１０７に記憶する。ここで、本処理部１０５が各特定データを類似度の高い順に並べ替えた場合、抽出装置１０は関連性の高い用語データの組み合わせを優先的にオペレータに提示することができるようになる。また、本処理部１０５が特定データを英字、数字、その他の並びの若い順に並べ替えた場合、抽出装置１０は先頭の文字の若い用語データの組み合わせを優先的にオペレータに提示することができるようになる。

【０１２８】

次に、Ｓ６２７において、抽出装置１０の本処理部１０５は、抽出した第１の言語による特定データをデータ格納部１０１に一時格納する。

【０１２９】

次に、抽出装置１０は、第２の言語によって作成された翻訳文データに対して、原文データに対して行ったのと同様の処理を行う。なお、以下に説明するＳ７０１～Ｓ７３３において、Ｓ７０１～Ｓ７２７は、第２の実施の形態におけるＳ６０１～Ｓ６２７と同様である。

【０１３０】

すなわち、図８に示されるように、Ｓ７０１において、抽出装置１０の前処理部１０３は、磁気記録媒体や光記録媒体、紙媒体、または外部のサーバ４００などから第２の言語による複数の文書データを取得し、文章データ毎に文章データコードを付与してデータ格納部１０１に一時格納する。

【０１３１】
次に、Ｓ７０３，Ｓ７０５において、抽出装置１０の前処理部１０３は、データ格納部１０１に一時格納された複数の文書データの各々から、前述の不要データを除外して、前述の本文データを抽出する。
【０１３２】
次に、Ｓ７０７，Ｓ７０９において、抽出装置１０の前処理部１０３は、複数の文書データの各々から抽出された各本文データの形態素を解析し、各本文データから切り出した各形態素に品詞を付与する。
【０１３３】
この後、Ｓ７１３において、抽出装置１０の本処理部１０５は、各本文データの中から各用語データを抽出し（図２参照）、各用語データにＩＤを付与する。
【０１３４】
次に、Ｓ７１５において、抽出装置１０の本処理部１０５は、各用語データを、例えば英字、数字、その他に分類し、英字、数字、その他の並びの若い順に並べ替える。これによって同じ用語データ毎に集合が形成される。
【０１３５】
次に、Ｓ７１７において、抽出装置１０の本処理部１０５は、各集合中の用語データの個数をカウントする（図３（ａ）参照）。
【０１３６】
次に、Ｓ７１９において、抽出装置１０の本処理部１０５は、各文章データ中に出現する各用語データの数に応じて、各用語データに所定の重み付きの値（すなわち、各文書データとの相関度を示す値）を付与する（図３（ｂ）参照）。すなわち、前述の文章データコードとリンク付けされた各用語データのＩＤに、所定の重み付きの値（すなわち、各文書データとの相関度を示す値）をリンク付けする。
【０１３７】
次に、Ｓ７２１，Ｓ７２３において、抽出装置１０の本処理部１０５は、重み付きの値をパラメータとして持つ前述の式（６）に基づいて、用語データ毎に、他の用語データと組み合わせた場合の類似度S_g（F_g，T_g）を算出する（図３（ｃ）参照）。そして、各用語データの組み合わせの関係を推定する。また、このとき、抽出装置１０の本処理部１０５は、算出した類似度が高い順に所定数の用語データの組み合わせを特定データとして抽出する。または、抽出装置１０の本処理部１０５は、算出した各用語データの組み合わせの類似度と所定の閾値とを比較し、所定の閾値を超える用語データの組み合わせを第２の言語による特定データとして抽出する。
【０１３８】
次に、Ｓ７２５において、抽出装置１０の本処理部１０５は、各特定データをＳ６２５と同様の順番に並べ替え、データ記憶部１０７に記憶する。
【０１３９】
次に、Ｓ７２７において、抽出装置１０の本処理部１０５は、抽出した第２の言語による特定データをデータ格納部１０１に一時格納する。
【０１４０】
次に、抽出装置１０の本処理部１０５は、原文データから抽出した特定データと翻訳文データから抽出した特定データとを対象にして、以下の処理を行う。
【０１４１】
すなわち、まず、Ｓ７２９において、抽出装置１０の本処理部１０５は、翻訳辞書を用いて、言語間で特定データ同士を比較する。すなわち、原文データから抽出されたキーとなる用語データに組み合わされた用語データと、翻訳文データから抽出されたキーとなる用語データに組み合わされた用語データとを比較する。なお、ここでは、キーとなる用語データは、第１の言語と第２の言語とで対応するように翻訳辞書に掲載されているものとする。
【０１４２】

次に、Ｓ７３１において、抽出装置１０の本処理部１０５は、キーとなる用語データに組み合わされた用語データの組み合わせを訳語の関係にある用語データの組み合わせとして抽出する。

【０１４３】

次に、Ｓ７３３において、抽出装置１０の本処理部１０５は、抽出した訳語の関係にある用語データを外部機器２００に出力する。すなわち、例えばディスプレイ２０９に出力して表示させたり、外部記憶装置２１３に出力して記憶させたりする。

【０１４４】

ここで、訳語の関係にあるものとして抽出した２４件の用語対を以下の表４に示す。

【０１４５】

【表４】

英単語	日本語	英単語	日本語
Yen	円	Had	た
And	や	economic	経済
percent	%	Not	ない
Was	た	economy	経済
And	など	To	に
Also	も	And	と
Should	べき	Or	や
Such	など	To	ため
No	ない	As	など
economy	景気	And	.
other	など	is	ある
As	として	To	こと

【０１４６】

このようにして、抽出装置１０は、訳語の関係にある特定データを抽出することができる。以上が、抽出装置１０の動作の詳細である。

【０１４７】

以上の通り、この実施の形態では、訳語の関係にある特定データを抽出することができる。特に、分野や著者に応じて、どの用語がどのような形に訳される傾向にあるのかを抽出することができる。そのため、分野や著者に応じて、好適な翻訳用の用語辞書を作成することができる。また、用語の翻訳は、分野や著者毎に作成した用語辞書を予め用意し、分野や著者を指定して、これに対応した用語辞書を用いることによって高精度に行うことができる。その結果、分野や著者に応じて、用語を独特の言い回しで翻訳することができる。

【０１４８】

＜付記＞

この発明を用いれば、例えば、新しい用語辞書の作成や、書籍に用いられた用語の目録の作成、特定の分野に用いられる用語の用法、特定の人物における用語の用法などを抽出することにも適用することができる。

【０１４９】

この発明は上記の実施の形態に限定されることなく、この発明の要旨を逸脱しない範囲で種々の応用及び変形が考えられる。例えば、用語データの抽出に限らず、なんらかの関係

（例えば、包含関係や、被包含関係など）にあるデータの組み合わせを抽出するのに用いてもよい。

【０１５０】

【発明の効果】

以上説明したこの発明には、次の効果がある。

【０１５１】

請求項１に記載の発明によれば、キーとなる用語データとの相関度が高い文書データ同士から抽出された用語データの組み合わせ（すなわち、キーとなる用語データが多数出現する文書データ同士から抽出された用語データの組み合わせ）ほど、高い重み付き補完類似度を算出できるので、重み付き補完類似度が高い順に所定数の用語データの組み合わせを特定データとして抽出する、または、重み付き補完類似度が所定の閾値を越える用語データの組み合わせを特定データとして抽出することにより、既存の用語辞書によらなくても特定データを高精度に抽出することができる。特に、分野や著者に応じて、どの用語がどのような形に訳される傾向にあるのかを抽出することができる。そのため、分野や著者に応じて、好適な翻訳用の用語辞書を作成することができる。また、用語の翻訳は、分野や著者毎に作成した用語辞書を予め用意し、分野や著者を指定して、これに対応した用語辞書を用いることによって高精度に行うことができる。その結果、分野や著者に応じて、用語を独特の言い回しで翻訳することができる。

【０１５２】

また、請求項１に記載の発明によれば、非対称性を持つ（すなわち、２つのパターンを入れ替えると異なる値になる）補完類似度を用いて特定データを抽出するので、包含関係を持つ特定データを高精度に抽出することができる。特に、重み付き補完類似度を用いて特定データを抽出するので、重み付きのない補完類似度よりも、高精度に特定データを抽出することができる。

【０１５３】

また、請求項２に記載の発明によれば、各用語データに固有のＩＤを付与するとともに、ＩＤに各文書データとの相関度を示す値を関連付けているので、重み付き補完類似度算出工程において、用語データ毎の、他の用語データとの重み付き補完類似度の算出を容易に行うことができる。

【０１５４】

また、請求項３に記載の発明によれば、複数の文書データの中から不要データ（すなわち、特定データの抽出対象とならない領域のデータ）を除外してから各用語データを抽出するので、用語データの抽出を短時間で行うことができる。

【０１５５】

また、請求項４に記載の発明によれば、特定データとして抽出する数、または、重み付き補完類似度の閾値を変更することによって、特定データの抽出範囲を広げる、または、狭めることができる。例えば、特定データとして抽出する数を予め設定しておき、その数で一次抽出し、その結果を参照して、抽出範囲を広げるように数を増やしたり、逆に、抽出範囲を狭めるように数を減らして、二次抽出することができる。または、重み付き補完類似度の閾値を予め設定しておき、その閾値で一次抽出し、その結果を参照して、抽出範囲を広げるように閾値を下げたり、逆に、狭めるように閾値を上げて、二次抽出することができる。

【０１５６】

また、請求項５に記載の発明によれば、特定データの抽出を、複数組の、２つの異なる言語によって作成された同じ内容の文書データを対象にして行い、言語毎に抽出された特定データを２つの言語間で比較するので、訳語の関係にある用語データの組み合わせを抽出することができる。特に、請求項１に記載の発明と同様に、分野や著者に応じて、どの用語がどのような形に訳される傾向にあるのかを抽出することができる。そのため、分野や著者に応じて、好適な翻訳用の用語辞書を作成することができる。また、用語の翻訳は、分野や著者毎に作成した用語辞書を予め用意し、分野や著者を指定して、これに対応し

た用語辞書を用いることによって高精度に行うことができる。その結果、分野や著者に応じて、用語を独特の言い回しで翻訳することができる。

【図面の簡単な説明】

【図１】この発明に係る抽出装置の構成を示す図である。

【図２】各データの関係を示す図である。

【図３】各データの関係を示す図である。

【図４】特定データの抽出工程を示すフローチャートである。

【図５】文書データ中における用語データの出現パターンを示す図である。

【図６】文書データ中における用語データの出現パターンを示す図である。

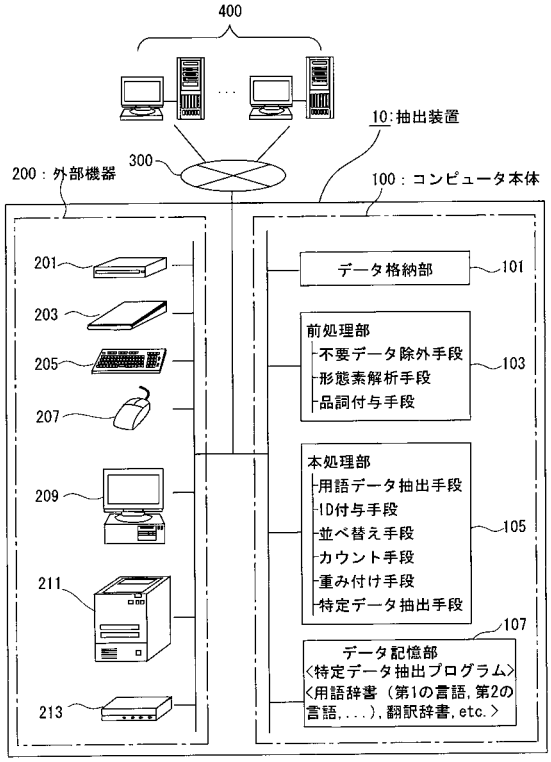
【図７】訳語の関係にある品詞データの抽出工程を示すフローチャートである。

【図８】訳語の関係にある品詞データの抽出工程を示すフローチャートである。

【符号の説明】

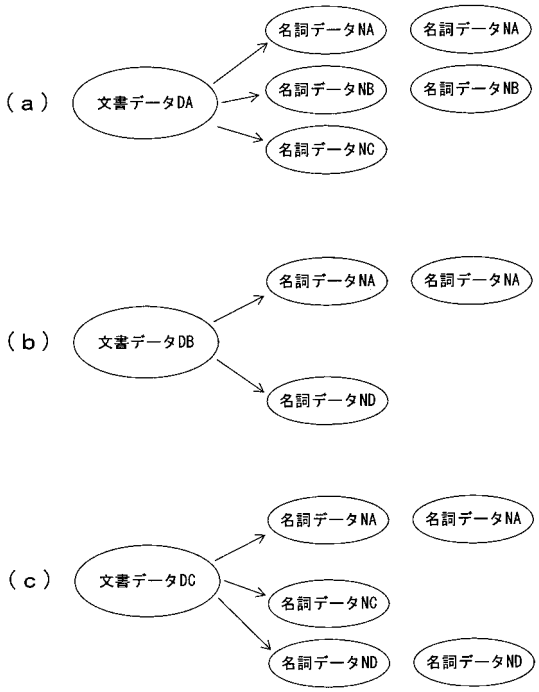
１０　抽出装置
１００　コンピュータ本体
１０１　データ格納部
１０３　前処理部
１０５　本処理部
１０７　データ記憶部
２００　外部機器
２０１　読取装置
２０３　スキャナ
２０５　キーボード
２０７　マウス
２０９　ディスプレイ
２１１　プリンタ
２１３　外部記憶装置
３００　通信回線
４００　サーバ群

【図１】



この発明に係る抽出装置の構成を示す図

【図２】



各データの関係を示す図

【図３】

（ａ）

	名詞データNA	名詞データNB	名詞データNC	名詞データND
文書データDA	2	2	1	0
文書データDB	2	0	0	1
文書データDC	2	0	1	2

（ｂ）

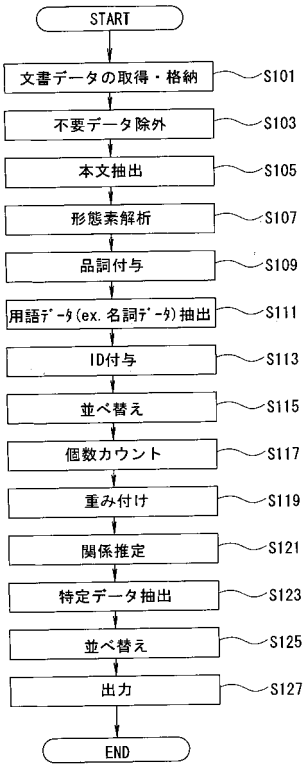
	名詞データNA	名詞データNB	名詞データNC	名詞データND
文書データDA	k3	k3	k2	k1
文書データDB	k3	k1	k1	k2
文書データDC	k3	k1	k2	k3

（ｃ）

	名詞データNA	名詞データNB	名詞データNC	名詞データND
名詞データNA	—	$S_{NANB}(F_{NA}, T_{NB})$	$S_{NANC}(F_{NA}, T_{NC})$	$S_{NAND}(F_{NA}, T_{ND})$
名詞データNB	$S_{NBNA}(F_{NB}, T_{NA})$	—	$S_{NBNC}(F_{NB}, T_{NC})$	$S_{NBND}(F_{NB}, T_{ND})$
名詞データNC	$S_{NCNA}(F_{NC}, T_{NA})$	$S_{NCNB}(F_{NC}, T_{NB})$	—	$S_{NCND}(F_{NC}, T_{ND})$
名詞データND	$S_{NDNA}(F_{ND}, T_{NA})$	$S_{NDNB}(F_{ND}, T_{NB})$	$S_{NDNC}(F_{ND}, T_{NC})$	—

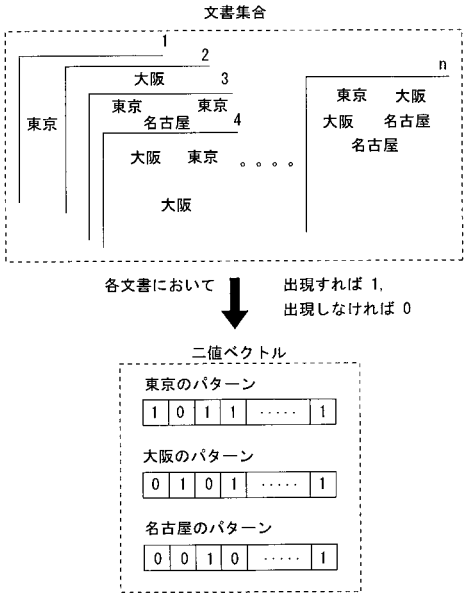
各データの関係を示す図

【図４】



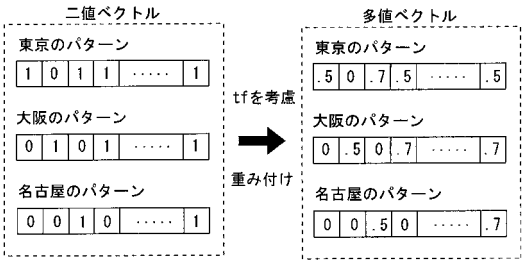
特定データの抽出工程を示すフローチャート

【図５】



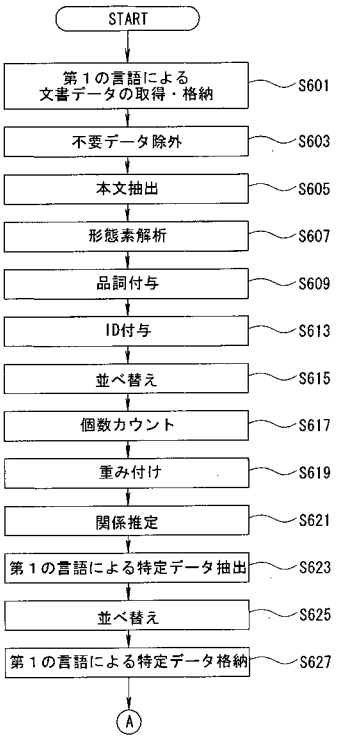
文書データ中における用語データの出現パターンを示す図

【図６】



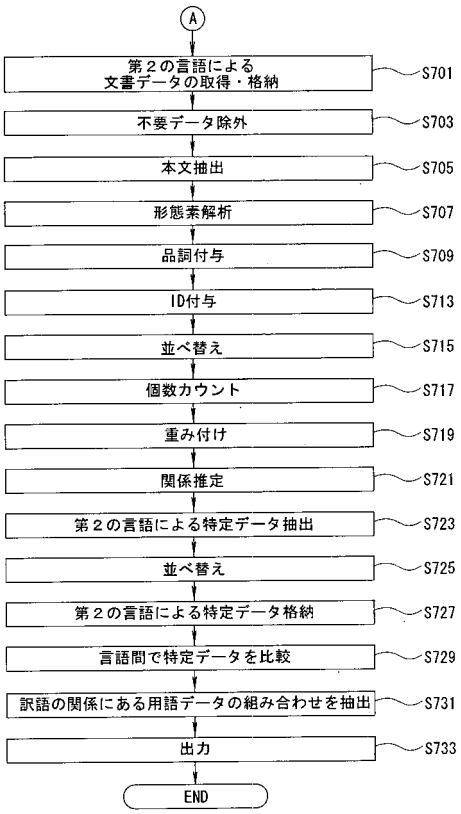
文書データ中における用語データの出現パターンを示す図

【図７】



訳語の関係にある品詞データの抽出工程を示すフローチャート

【図８】



訳語の関係にある品詞データの抽出工程を示すフローチャート

フロントページの続き

審査官　深津　始

(56)参考文献　松本兼一、梅村恭司，補間類似度を用いた固有名詞のグルーピングの試み，電子情報通信学会技術研究報告，１９９６年　7月１８日，９６巻、１５７号，１－６頁，NLC96-9
山本英子、梅村恭司，コーパス中の一対多関係を推定する問題における類似尺度，自然言語処理，２００２年　4月１０日，9巻、2号，４５－７５頁
松本兼一、梅村恭司，補間類似度を用いた固有名詞のグルーピングの試み，情報処理学会研究報告，１９９６年　7月１８日，９６巻、６５号，１－６頁，96-NL-114-1
鈴木猛雄、力宗幸男，頻出名詞を用いた文書分類・検索システム，電子情報通信学会技術研究報告，２０００年１１月１０日，１００巻、４３９号，１７－２３頁，OFS2000-48

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