

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2004-265169

(P2004-265169A)

(43) 公開日 平成16年9月24日(2004.9.24)

(51) Int. Cl. ⁷	F I	テーマコード (参考)
G06F 17/30	G06F 17/30 210A	5B075
G06F 17/28	G06F 17/30 220A	5B091
	G06F 17/28 U	

審査請求 有 請求項の数 10 O L (全 29 頁)

(21) 出願番号 特願2003-55193 (P2003-55193)	(71) 出願人 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
(22) 出願日 平成15年3月3日(2003.3.3)	(74) 代理人 100085419 弁理士 大垣 孝
特許法第30条第1項適用申請有り 2002年9月13日 発行の「2002年(平成14年)情報科学技術フォーラム一般講演論文集第2分冊」に発表	(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
	Fターム(参考) 5B075 ND03 NK32 NS01 QM08 UU06 5B091 AA15 AB06 CA02 EA24

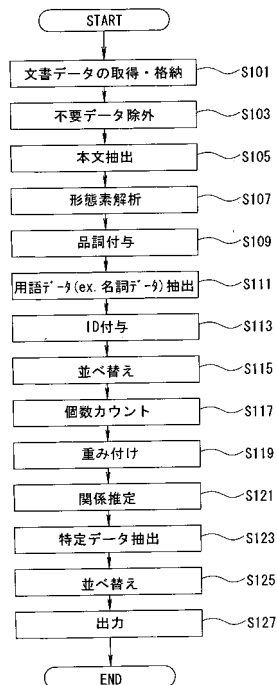
(54) 【発明の名称】 特定データの抽出方法、抽出装置、およびプログラム

(57) 【要約】 (修正有)

【課題】 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、人名と役職、場所と人物など)があった。

【解決手段】 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、複数の文書データの中から各用語データを抽出する工程S111と、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工程S119と、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、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を算出する工程S121と、前記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、前記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工程S123と、を含むことを特徴とする。

【選択図】 図4



特定データの抽出工程を示すフローチャート

【特許請求の範囲】

【請求項 1】

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、

複数の文書データの中から各用語データを抽出する工程と、

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工程と、

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、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を算出する工程と、

前記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、前記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工程と、

を含むことを特徴とする特定データの抽出方法。

【請求項 2】

前記類似度を算出する工程は、重み付き補完類似度を算出するための式に基づいて行われ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】

前記重み付き補完類似度を算出するための式は、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 とする場合における、以下の式(6)であることを特徴とする請求項 2 に記載の特定データの抽出方法。

【数 1】

$$\vec{F}_g = \{(f_{g1}), (f_{g2}), \dots, (f_{gn})\} \quad ((f_{gi}) = 0.0 \text{ through } 1.0) \quad \dots (4)$$

$$\vec{T}_g = \{(t_{g1}), (t_{g2}), \dots, (t_{gn})\} \quad ((t_{gi}) = 0.0 \text{ through } 1.0) \quad \dots (5)$$

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

ただし、

$$a_g = \sum_{i=1}^n (f_{gi}) \times (t_{gi}), \quad b_g = \sum_{i=1}^n (1 - (f_{gi})) \times (t_{gi}),$$

$$c_g = \sum_{i=1}^n (f_{gi}) \times (1 - (t_{gi})), \quad d_g = \sum_{i=1}^n (1 - (f_{gi})) \times (1 - (t_{gi})), \quad \dots (40)$$

$$F_g = \sum_{i=1}^n (f_{gi}), \quad T_g = \sum_{i=1}^n (t_{gi}), \quad (T_g)_2 = \sum_{i=1}^n (t_{gi})^2$$

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

10

20

30

40

50

度とする。

【請求項 4】

前記各用語データを抽出する工程で、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 5】

前記各用語データを抽出する工程の前工程として、

複数の文書データを取得する工程と、

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工程と、

取得し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工程と、を含む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】

前記特定データとして抽出する工程で、特定データとして抽出する数、または、前記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 7】

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、

複数の文書データの中から各用語データを抽出する工程と、

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工程と、

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、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を算出する工程と、

前記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、前記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特定データの抽出方法。

【請求項 8】

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、

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、用語データ毎に、他の用語データとの類似度を算出し、前記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、前記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。

【請求項 9】

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、

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、用語データ毎に、他の用語

10

20

30

40

50

データとの類似度を算出し、前記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、前記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ことができ、また、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ができ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。

【請求項10】

コンピュータにより、請求項8または9に記載の特定データの抽出装置を実現するためのプログラム。

10

【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】

20

【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】

30

【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】

40

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、第1の発明に係る特定データの抽出方法は、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方法において、複数の文書データの中から各用語データを抽出する工程と、各文書データ中出现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工程と、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

50

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を算出する工程と、前記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、前記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工程と、を含むことを特徴とする。

【0008】

第1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、高い類似度を算出できるので、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、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抽出することができる。

10

【0009】

また、第2の発明に係る特定データの抽出方法は、第1の発明に加え、前記類似度を算出する工程が、重み付き補完類似度を算出するための式に基づいて行われることを特徴とする。

【0010】

また、第3の発明に係る特定データの抽出方法は、第2の発明に加え、前記重み付き補完類似度を算出するための式が、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 とする場合における、以下の式(6)であることを特徴とする。

20

【0011】

【数2】

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

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

30

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

ただし、

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

40

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

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

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

50

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。

10

【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。

20

【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を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、第 4 の発明に係る特定データの抽出方法は、第 1 の発明に加え、前記各用語データを抽出する工程で、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。

30

【0018】

第 4 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、類似度を算出する工程において、用語データ毎の、他の用語データとの類似度の算出を容易に行うことができる。

【0019】

また、第 5 の発明に係る特定データの抽出方法は、第 1 の発明に加え、前記各用語データを抽出する工程の前工程として、複数の文書データを取得する工程と、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工程と、取得し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工程と、を含むことを特徴とする。

40

【0020】

第 5 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、第 6 の発明に係る特定データの抽出方法は、第 1 の発明に加え、前記特定データとして抽出する工程で、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

50

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】

第6の発明は、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、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、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。

10

【0023】

また、第7の発明に係る特定データの抽出方法は、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方法において、複数の文書データの中から各用語データを抽出する工程と、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工程と、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、用語データ毎に、他の用語データとの類似度を算出する工程と、前記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、前記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。

20

【0024】

第7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、第8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、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、用語データ毎に、他の用語データとの類似度を算出し、前記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、前記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。第8の発明は、第1の発明を実施するための装置の構成を明示している。

30

【0026】

また、第9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、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、用語データ毎に、他の用語データとの類似度を算出し、前記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、前記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ことができ、また、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ができる特定データ抽出手段と、を有することを特徴

40

50

とする。第9の発明は、第7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、第10の発明に係るプログラムは、コンピュータにより、第8の発明または第9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。第10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、第8の発明または第9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。

10

【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。

20

【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。

30

40

【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。

50

【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。

10

【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。

20

【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、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。

30

【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。

40

【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 $k_1 \sim k_3$ 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 $w_{eigh}t(t_f)$ 」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

50

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）が一对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对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、一对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。一对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

10

20

30

40

50

説明する。

【 0 0 5 7 】

【 数 3 】

$$\vec{F} = \{ f_1, f_2, \dots, f_n \} \quad (f_j=0 \text{ or } 1) \quad \dots (1)$$

$$\vec{T} = \{ t_1, t_2, \dots, t_n \} \quad (t_j=0 \text{ or } 1) \quad \dots (2)$$

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

ただし、

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

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

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

である。

【 0 0 5 8 】

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。

【 0 0 5 9 】

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つパターンの中で、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。

【 0 0 6 0 】

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。

【 0 0 6 1 】

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

10

20

30

40

50

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。

10

【0063】

【数4】

$$\vec{F}_g = \{(f_{g1}), (f_{g2}), \dots, (f_{gn})\} \quad ((f_{gi}) = 0.0 \text{ through } 1.0) \quad \dots (4)$$

$$\vec{T}_g = \{(t_{g1}), (t_{g2}), \dots, (t_{gn})\} \quad ((t_{gi}) = 0.0 \text{ through } 1.0) \quad \dots (5)$$

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

20

ただし、

$$a_g = \sum_{i=1}^n (f_{gi}) \times (t_{gi}), \quad b_g = \sum_{i=1}^n (1 - (f_{gi})) \times (t_{gi}),$$

$$c_g = \sum_{i=1}^n (f_{gi}) \times (1 - (t_{gi})), \quad d_g = \sum_{i=1}^n (1 - (f_{gi})) \times (1 - (t_{gi})),$$

30

$$F_g = \sum_{i=1}^n (f_{gi}), \quad T_g = \sum_{i=1}^n (t_{gi}), \quad (T_g)_2 = \sum_{i=1}^n (t_{gi})^2$$

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。

40

【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、用語データ間の一对多の

50

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。

10

【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。

20

【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。

30

【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】

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

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

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

ただし、

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

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

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

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、一对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

10

20

30

40

50

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。

10

【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。

20

【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。

30

【0084】

重み付きの値 $w e i g h t (t f)$ は、関係を推定する対象となる用語データの文書内頻度 $T F$ 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 $T h g j$ 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 $T h g j$ が出現する全ての文書数を $d f (T h g j)$ とし、各用語データ $T h g j$ が1回だけ出現する文書数を $d f 1 (T h g j)$ とし、各用語データ $T h g j$ が2回だけ出現する文書数を $d f 2 (T h g j)$ とする。

40

【0086】

このような定義において、全ての用語データ $w j$ (ただし、 $1 \leq j \leq m$) を対象とする重み付きの値 $w e i g h t (t f)$ は、 $t f = 0$ の場合(すなわち、各用語データ $T h g j$ 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、 $t f = 1$ の場合(すなわち、各用語データ $T h g j$ 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、 $t f = 2$ の場合(すなわち、各用語データ $T h g j$ 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、 $t f = 3$ の場合(すなわち、各用語データ $T h g j$ が1回でも出現する場合)に以下の式(10)によって

50

算出された値（すなわち 1）となる。

【 0 0 8 7 】

【 数 6 】

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

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

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

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

10

【 0 0 8 8 】

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。

20

【 0 0 8 9 】

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。

【 0 0 9 0 】

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。

30

【 0 0 9 1 】

【 表 1 】

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

40

【 0 0 9 2 】

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

50

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 $S_g(F_g, T_g)$ 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】

$$\vec{F}_g = \{(f_{g1}), (f_{g2}), \dots, (f_{gn})\} \quad ((f_{gi}) = 0.0 \text{ through } 1.0) \quad \dots (4)$$

$$\vec{T}_g = \{(t_{g1}), (t_{g2}), \dots, (t_{gn})\} \quad ((t_{gi}) = 0.0 \text{ through } 1.0) \quad \dots (5)$$

10

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

ただし、

$$a_g = \sum_{i=1}^n (f_{gi}) \times (t_{gi}), \quad b_g = \sum_{i=1}^n (1 - (f_{gi})) \times (t_{gi}),$$

20

$$c_g = \sum_{i=1}^n (f_{gi}) \times (1 - (t_{gi})), \quad d_g = \sum_{i=1}^n (1 - (f_{gi})) \times (1 - (t_{gi})),$$

$$F_g = \sum_{i=1}^n (f_{gi}), \quad T_g = \sum_{i=1}^n (t_{gi}), \quad (T_g)_2 = \sum_{i=1}^n (t_{gi})^2$$

である。

30

【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 a_g , b_g , c_g , d_g 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_g , b_g , c_g , d_g 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。

40

【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 $S_g(F_g, T_g)$ 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 $S_g(F_g, T_g)$ は、重み付きのない補完類似度 $S_c(F, T)$ よりも、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 $S_c(F, T)$ と重み付き補完類似度 $S_g(F_g, T_g)$ の適合率は、以下の表2に示す通りとなる。

【0098】

【表2】

50

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

10

20

30

40

50

【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 $w e i g h t (t f)$ は、例えば、 $t f = 0$ の場合に0となり、 $t f = 1$ の場合に0.84となり、 $t f = 2$ の場合に0.95となり、 $t f = 3$ 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、例えば学术论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 $IDF \cdot TF$ を用いて算出される。

10

【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、一对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。

20

【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	

10

20

30

【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。

40

【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。

50

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】

10

20

30

40

50

次に、S 6 0 7 , S 6 0 9 において、抽出装置 1 0 の前処理部 1 0 3 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。

【 0 1 2 2 】

この後、S 6 1 3 において、抽出装置 1 0 の本処理部 1 0 5 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 I D を付与する。この I D 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。

【 0 1 2 3 】

次に、S 6 1 5 において、抽出装置 1 0 の本処理部 1 0 5 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。 10

【 0 1 2 4 】

次に、S 6 1 7 において、抽出装置 1 0 の本処理部 1 0 5 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）。

【 0 1 2 5 】

次に、S 6 1 9 において、抽出装置 1 0 の本処理部 1 0 5 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 I D 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。 20

【 0 1 2 6 】

次に、S 6 2 1 , S 6 2 3 において、抽出装置 1 0 の本処理部 1 0 5 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 1 0 の本処理部 1 0 5 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 1 0 の本処理部 1 0 5 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。

【 0 1 2 7 】

次に、S 6 2 5 において、抽出装置 1 0 の本処理部 1 0 5 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 1 0 7 に記憶する。ここで、本処理部 1 0 5 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 1 0 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 1 0 5 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 1 0 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。 30

【 0 1 2 8 】

次に、S 6 2 7 において、抽出装置 1 0 の本処理部 1 0 5 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 1 0 1 に一時格納する。 40

【 0 1 2 9 】

次に、抽出装置 1 0 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 S 7 0 1 ~ S 7 3 3 において、S 7 0 1 ~ S 7 2 7 は、第 2 の実施の形態における S 6 0 1 ~ S 6 2 7 と同様である。

【 0 1 3 0 】

すなわち、図 8 に示されるように、S 7 0 1 において、抽出装置 1 0 の前処理部 1 0 3 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 4 0 0 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 1 0 1 に一時格納する。 50

【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。 10

【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。 20

【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。 30

【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。 40

【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せられた用語データと、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せら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

10

20

30

40

50

(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】

第1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、高い類似度を算出できるので、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、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。

10

【0152】

また、第2の発明と第3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。

20

【0153】

また、第4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、類似度を算出する工程において、用語データ毎の、他の用語データとの類似度の算出を容易に行うことができる。

【0154】

また、第5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。

30

【0155】

また、第6の発明によれば、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、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、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、第7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、第1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。

40

【図面の簡単な説明】

50

- 【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。

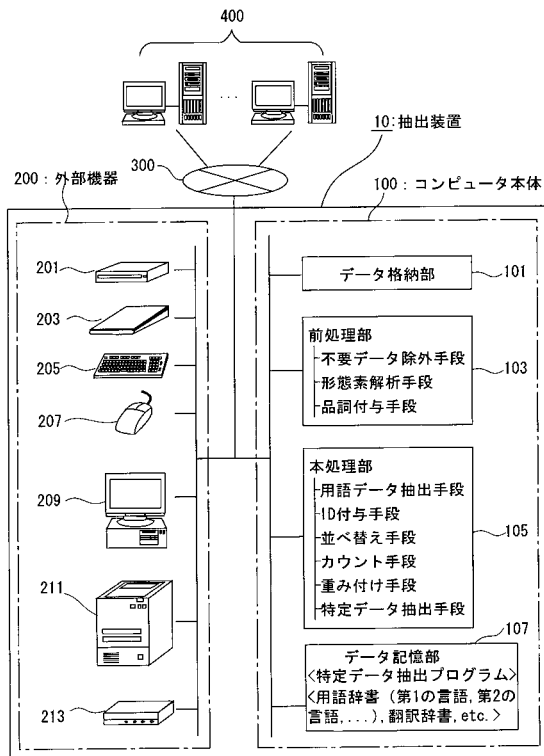
【符号の説明】

- 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

10

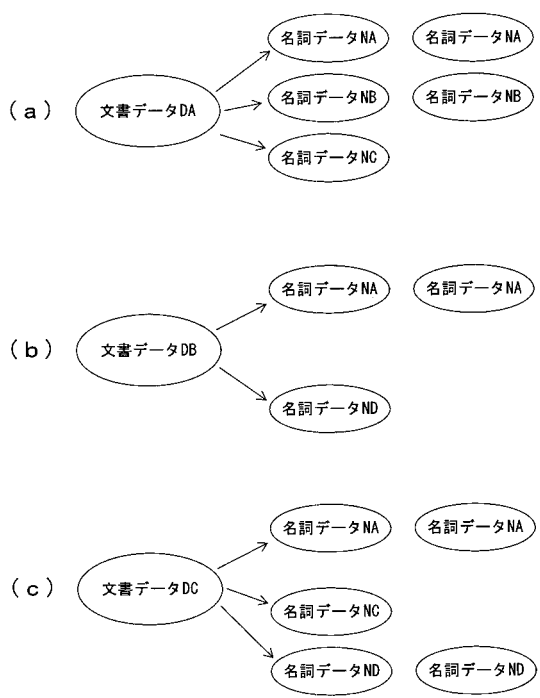
20

【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 】

	名詞データMA	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

(a)

	名詞データMA	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

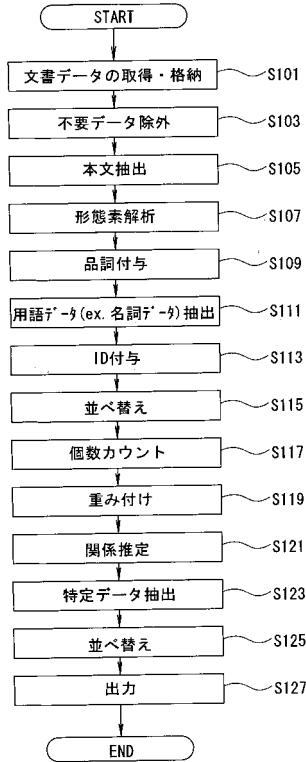
(b)

	名詞データMA	名詞データNB	名詞データNC	名詞データND
名詞データNA	—	SMAAB (FNA, INB)	SMAAC (FNA, INC)	SMAAD (FNA, IND)
名詞データNB	SMBNA (FNB, INA)	—	SMBNC (FNB, INC)	SMBND (FNB, IND)
名詞データNC	SNCNA (FNC, INA)	SNCNB (FNC, INB)	—	SNCND (FNC, IND)
名詞データND	SNDNA (FND, INA)	SNDNB (FND, INB)	SNDNC (FND, INC)	—

(c)

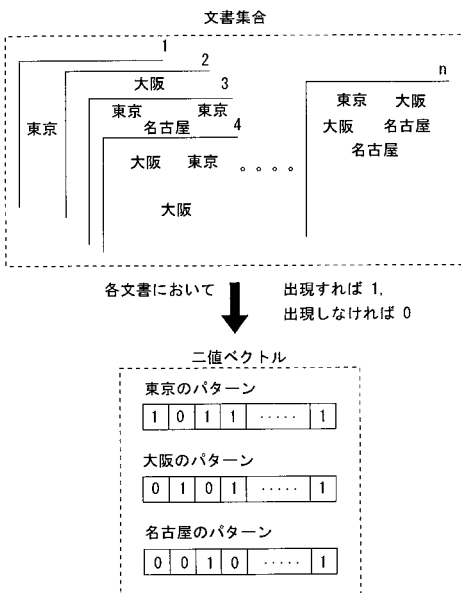
各データの関係を示す図

【 図 4 】



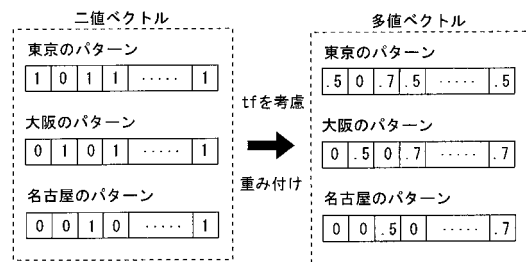
特定データの抽出工程を示すフローチャート

【 図 5 】



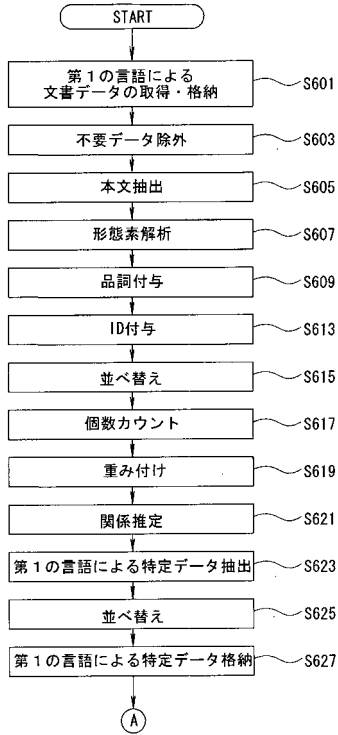
文書データ中における用語データの出現パターンを示す図

【 図 6 】



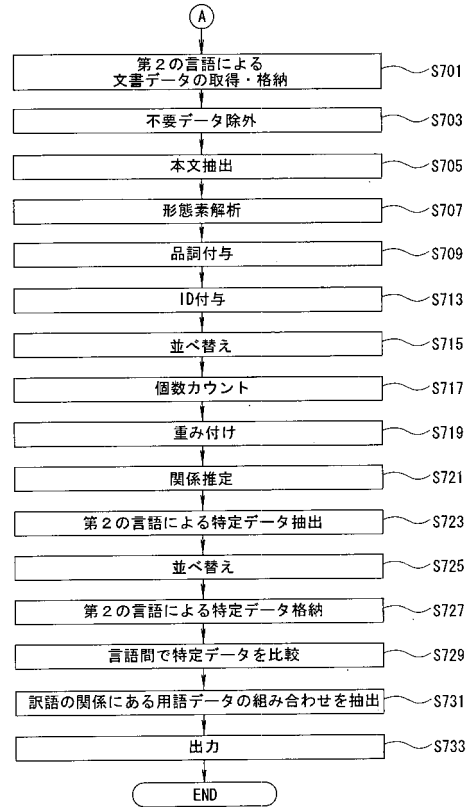
文書データ中における用語データの出現パターンを示す図

【 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