(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5099498号
(P5099498)
(45)発行日　平成24年12月19日(2012.12.19)　(24)登録日　平成24年10月5日(2012.10.5)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 220Z
G06F 17/21 (2006.01)　G06F 17/21 550A

請求項の数 18　(全 39 頁)

(21)出願番号　特願2007-286269 (P2007-286269)
(22)出願日　平成19年11月2日 (2007.11.2)
(65)公開番号　特開2009-116456 (P2009-116456A)
(43)公開日　平成21年5月28日 (2009.5.28)
審査請求日　平成22年10月7日 (2010.10.7)

(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　100130111
弁理士　新保　斉
(72)発明者　村田　真樹
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　金丸　敏幸
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内

審査官　吉田　誠

最終頁に続く

(54)【発明の名称】データ処理装置及びデータ処理方法

(57)【特許請求の範囲】
【請求項１】
所定の対象データに関連する関連データについて、相前後する第１関連データ及び第２関連データの順序を検出するデータ処理装置において、
ネットワーク上又はローカルの記憶手段に格納された同一又は異なるファイルから第１関連データ及び第２関連データをそれぞれ抽出する関連データ抽出手段と、
ネットワーク上又はローカルの記憶手段から該第１関連データ及び該第２関連データが共起する関連データ共起ファイルを抽出する関連データ共起ファイル抽出手段と、
該関連データ共起ファイルから、所定の関連データ間関係規則を参照して、第１関連データ及び第２関連データ間の前後を検出する関連データ間関係検出手段と、
該検出結果を出力する出力手段と
を備えたことを特徴とするデータ処理装置。
【請求項２】
前記関連データ間関係規則が、少なくとも前記第１関連データと前記第２関連データとの間に含まれる、又は含まれない、文字列に係る情報であって、
前記関連データ間関係検出手段が、前記関連データ共起ファイル内において該第１関連データと該第２関連データとの間の文字列を抽出し、該関連データ間関係規則と照合する
請求項１に記載のデータ処理装置。
【請求項３】
前記関連データ間関係規則が、予め前後関係が分かっている２つの教師用関連データが

共起する複数の教師用ファイルを用い、該教師用ファイルにおける２つの教師用関連データの出現位置、又は同時に含まれる若しくは含まれない文字列、又は同時に含まれるタグ情報の少なくともいずれかを素性として機械学習した学習結果であって、
前記関連データ間関係検出手段が、前記関連データ共起ファイルから該素性を抽出すると共に、前記第１関連データ及び前記第２関連データを入力として、該学習結果を参照して該第１関連データ及び該第２関連データ間の前後を算出する
請求項１に記載のデータ処理装置。
【請求項４】
前記データ処理装置であって、
前記関連データ抽出手段が、ネットワーク上又はローカルの記憶手段から前記対象データと共起する第１関連データ及び第２関連データをそれぞれ抽出する
請求項１ないし３のいずれかに記載のデータ処理装置。
【請求項５】
所定の対象データに関連する関連データについて、相前後する第１関連データ及び第２関連データの順序を検出するデータ処理装置において、
ネットワーク上又はローカルの記憶手段に格納された同一又は異なるファイルから第１関連データ及び第２関連データをそれぞれ抽出する関連データ抽出手段と、
該第１関連データが含まれるファイルから該第１関連データと共起する単数又は複数の第１共起データを抽出すると共に、該第２関連データが含まれるファイルから該第２関連データと共起する単数又は複数の第２共起データを抽出する共起データ抽出手段と、
該第１共起データ及び該第２共起データ間の前後に関する所定の共起データ間関係規則を参照して、第１共起データ及び第２共起データ間の前後を検出する共起データ間関係検出手段と、
該検出結果をそれらと共起している第１関連データ及び第２関連データの前後として出力する出力手段と
を備えたことを特徴とするデータ処理装置。
【請求項６】
前記データ処理装置において、
ネットワーク上又はローカルの記憶手段から該第１共起データ及び該第２共起データが共起する共起データ共起ファイルを抽出する共起データ共起ファイル抽出手段を備えた
請求項５に記載のデータ処理装置。
【請求項７】
前記共起データ間関係規則が、少なくとも前記第１共起データと前記第２共起データとの間に含まれる、又は含まれない、文字列に係る情報であって、
前記共起データ間関係検出手段が、前記共起データ共起ファイル内において該第１共起データと該第２共起データとの間の文字列を抽出し、該共起データ間関係規則と照合する
請求項６に記載のデータ処理装置。
【請求項８】
前記共起データ間関係規則が、予め前後関係が分かっている２つの教師用関連データのそれぞれと複数の文字列とが含まれる複数の教師用ファイルを用い、該教師用ファイルにおける該教師用関連データと共に含まれる単数又は複数の文字列を素性として機械学習した学習結果であって、
前記共起データ間関係検出手段が、前記第１共起データ及び前記第２共起データを素性として入力し、該学習結果を参照して該第１共起データ及び該第２共起データ間の前後を算出する
請求項５又は６に記載のデータ処理装置。
【請求項９】
前記データ処理装置で処理する対象データが宛名であり、関連データが宛先である構成において、
前記第1関連データと前記第２関連データとの間で、その新旧関係を検出する

ことを特徴とする請求項１ないし８のいずれかにデータ処理装置。
【請求項１０】
所定の対象データに関連する関連データについて、相前後する第１関連データ及び第２関連データの順序を検出するコンピュータのデータ処理方法であって、
関連データ抽出手段が、ネットワーク上又はローカルの記憶手段に格納された同一又は異なるファイルから第１関連データ及び第２関連データをそれぞれ抽出する関連データ抽出ステップ、
関連データ共起ファイル抽出手段が、ネットワーク上又はローカルの記憶手段から該第１関連データ及び該第２関連データが共起する関連データ共起ファイルを抽出する関連データ共起ファイル抽出ステップ、
関連データ間関係検出手段が、該関連データ共起ファイルから、所定の関連データ間関係規則を参照して、第１関連データ及び第２関連データ間の前後を検出する関連データ間関係検出ステップ、
出力手段が、該検出結果を出力する出力ステップ
を有することを特徴とするデータ処理方法。
【請求項１１】
前記関連データ間関係規則が、少なくとも前記第１関連データと前記第２関連データとの間に含まれる、又は含まれない、文字列に係る情報であって、
前記関連データ間関係検出手段が、前記関連データ共起ファイル内において該第１関連データと該第２関連データとの間の文字列を抽出し、該関連データ間関係規則と照合する
請求項１０に記載のデータ処理方法。
【請求項１２】
前記関連データ間関係規則が、予め前後関係が分かっている２つの教師用関連データが共起する複数の教師用ファイルを用い、該教師用ファイルにおける２つの教師用関連データの出現位置、又は同時に含まれる若しくは含まれない文字列、又は同時に含まれるタグ情報の少なくともいずれかを素性として機械学習した学習結果であって、
前記関連データ間関係検出手段が、前記関連データ共起ファイルから該素性を抽出すると共に、前記第１関連データ及び前記第２関連データを入力として、該学習結果を参照して該第１関連データ及び該第２関連データ間の前後を算出する
請求項１０に記載のデータ処理方法。
【請求項１３】
前記データ処理方法であって、
前記関連データ抽出ステップにおいて関連データ抽出手段が、ネットワーク上又はローカルの記憶手段から前記対象データと共起する第１関連データ及び第２関連データをそれぞれ抽出する
請求項１０ないし１２のいずれかに記載のデータ処理方法。
【請求項１４】
所定の対象データに関連する関連データについて、相前後する第１関連データ及び第２関連データの順序を検出するコンピュータのデータ処理方法であって、
関連データ抽出手段が、ネットワーク上又はローカルの記憶手段に格納された同一又は異なるファイルから第１関連データ及び第２関連データをそれぞれ抽出する関連データ抽出ステップ、
共起データ抽出手段が、該第１関連データが含まれるファイルから該第１関連データと共起する単数又は複数の第１共起データを抽出すると共に、該第２関連データが含まれるファイルから該第２関連データと共起する単数又は複数の第２共起データを抽出する共起データ抽出ステップ、
共起データ間関係検出手段が、該第１共起データ及び該第２共起データ間の前後に関する所定の共起データ間関係規則を参照して、第１共起データ及び第２共起データ間の前後を検出する共起データ間関係検出ステップ、
出力手段が、該検出結果をそれらと共起している第１関連データ及び第２関連データの

前後として出力する出力ステップ
を有することを特徴とするデータ処理方法。
【請求項１５】
前記データ処理方法において、前記共起データ抽出ステップの次に、
共起データ共起ファイル抽出手段が、ネットワーク上又はローカルの記憶手段から該第１共起データ及び該第２共起データが共起する共起データ共起ファイルを抽出する共起データ共起ファイル抽出ステップを有する
請求項１４に記載のデータ処理方法。
【請求項１６】
前記共起データ間関係規則が、少なくとも前記第１共起データと前記第２共起データとの間に含まれる、又は含まれない、文字列に係る情報であって、
前記共起データ間関係検出手段が、前記共起データ共起ファイル内において該第１共起データと該第２共起データとの間の文字列を抽出し、該共起データ間関係規則と照合する
請求項１５に記載のデータ処理方法。
【請求項１７】
前記共起データ間関係規則が、予め前後関係が分かっている２つの教師用関連データのそれぞれと複数の文字列とが含まれる複数の教師用ファイルを用い、該教師用ファイルにおける該教師用関連データと共に含まれる単数又は複数の文字列を素性として機械学習した学習結果であって、
前記共起データ間関係検出手段が、前記第１共起データ及び前記第２共起データを素性として入力し、該学習結果を参照して該第１共起データ及び該第２共起データ間の前後を算出する
請求項１４又は１５に記載のデータ処理方法。
【請求項１８】
前記データ処理方法が、対象データが宛名であり、関連データが宛先である構成において、
前記第1関連データと前記第２関連データとの間で、その新旧関係を検出する
ことを特徴とする請求項１０ないし１７のいずれかにデータ処理方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、所定の対象データに関連する関連データ間の順序を検出するデータ処理装置と方法に関し、より詳しくは所定のルールや機械学習に基づいて生成される規則に従って、順序を検出する技術に関わる。
【背景技術】
【０００２】
企業や個人の連絡先を調べる際に、インターネットで検索したり、ローカルなハードディスクに蓄積されたデータベースを検索することは日常的に行われている。このようなデータは、一度蓄積されるとなかなか消去されることがなく、企業が移転をしても従前の住所が検索結果として出力されることが少なくない。
【０００３】
このような住所に関する情報の他、企業名の変更や、企業の人事情報や、個人の勤務先情報、製品の型番情報など、ある対象データに関連する関連データが更新された場合に、どちらが新しい関連データなのかを解決すべき場面は多い。
【０００４】
ところで、非特許文献１および２に示されるように、ウェブページなど文書データから企業の所在地住所を取り出す研究や、単一の文書から企業内の人事の情報を取り出す研究は従来から知られている。しかし、企業名、人名を入力として企業の住所の変化情報、人の所属の変化情報を、文書の日付を自動推定する技術や教師あり機械学習手法を含めた高度な自然言語処理技術を駆使してウェブの複数の文書を総合的に扱って取り出す先行技術

はない。
【０００５】
【非特許文献１】佐藤理史、ワールドワイドウェブを利用した住所探索、情報処理学会論文誌，Vol.42，No.1，pp.59-67，2001年
【非特許文献２】関根聡，テキストからの情報抽出－文書から特定の情報を抜き出す－，情報処理，Vol.40，No.4，pp.370-373，1999年
【発明の開示】
【発明が解決しようとする課題】
【０００６】
本発明は、上記従来技術の有する問題点に鑑みて創出されたものであり、対象データに関連する２つの関連データの前後を精度良く検出する技術を提供することを目的とする。
【課題を解決するための手段】
【０００７】
本発明は次のようなデータ処理装置を提供することもできる。
すなわち、請求項１に記載の発明は、所定の対象データに関連する関連データについて、相前後する第１関連データ及び第２関連データの順序を検出するデータ処理装置であって、ネットワーク上又はローカルの記憶手段に格納された同一又は異なるファイルから第１関連データ及び第２関連データをそれぞれ抽出する関連データ抽出手段と、ネットワーク上又はローカルの記憶手段から該第１関連データ及び該第２関連データが共起する関連データ共起ファイルを抽出する関連データ共起ファイル抽出手段と、関連データ共起ファイルから、所定の関連データ間関係規則を参照して、第１関連データ及び第２関連データ間の前後を検出する関連データ間関係検出手段と、検出結果を出力する出力手段とを備えたことを特徴とする。
【０００８】
請求項２に記載の発明によれば上記の関連データ間関係規則が、少なくとも前記第１関連データと前記第２関連データとの間に含まれる、又は含まれない、文字列に係る情報であって、関連データ間関係検出手段が、関連データ共起ファイル内において該第１関連データと該第２関連データとの間の文字列を抽出し、該関連データ間関係規則と照合することを特徴とする。
【０００９】
請求項３に記載の発明によれば、上記の関連データ間関係規則が、予め前後関係が分かっている２つの教師用関連データが共起する複数の教師用ファイルを用い、該教師用ファイルにおける２つの教師用関連データの出現位置、又は同時に含まれる若しくは含まれない文字列、又は同時に含まれるタグ情報の少なくともいずれかを素性として機械学習した学習結果であって、関連データ間関係検出手段が、前記関連データ共起ファイルから該素性を抽出すると共に、前記第１関連データ及び前記第２関連データを入力として、該学習結果を参照して該第１関連データ及び該第２関連データ間の前後を算出することを特徴とする。
【００１０】
請求項４に記載の発明によれば、上記請求項１ないし３のいずれかのデータ処理装置であって、関連データ抽出手段が、ネットワーク上又はローカルの記憶手段から前記対象データと共起する第１関連データ及び第２関連データをそれぞれ抽出することを特徴とする。
【００１１】
請求項５に記載の発明によれば、所定の対象データに関連する関連データについて、相前後する第１関連データ及び第２関連データの順序を検出するデータ処理装置において、ネットワーク上又はローカルの記憶手段に格納された同一又は異なるファイルから第１関連データ及び第２関連データをそれぞれ抽出する関連データ抽出手段と、第１関連データが含まれるファイルから該第１関連データと共起する単数又は複数の第１共起データを抽出すると共に、該第２関連データが含まれるファイルから該第２関連データと共起する単

数又は複数の第２共起データを抽出する共起データ抽出手段と、第１共起データ及び第２共起データ間の前後に関する所定の共起データ間関係規則を参照して、第１共起データ及び第２共起データ間の前後を検出する共起データ間関係検出手段と、検出結果をそれらと共起している第１関連データ及び第２関連データの前後として出力する出力手段とを備えたことを特徴とするデータ処理装置を提供する。

【００１２】

請求項６に記載の発明によれば、ネットワーク上又はローカルの記憶手段から該第１共起データ及び該第２共起データが共起する共起データ共起ファイルを抽出する共起データ共起ファイル抽出手段を備えた処理装置を提供してもよい。

【００１３】

請求項７に記載の発明によれば、上記の共起データ間関係規則が、少なくとも前記第１共起データと前記第２共起データとの間に含まれる、又は含まれない、文字列に係る情報であって、共起データ間関係検出手段が、前記共起データ共起ファイル内において該第１共起データと該第２共起データとの間の文字列を抽出し、該共起データ間関係規則と照合することを特徴とする。

【００１４】

請求項８に記載の発明によれば、上記の共起データ間関係規則が、予め前後関係が分かっている２つの教師用関連データのそれぞれと複数の文字列とが含まれる複数の教師用ファイルを用い、該教師用ファイルにおける該教師用関連データと共に含まれる単数又は複数の文字列を素性として機械学習した学習結果であって、共起データ間関係検出手段が、前記第１共起データ及び前記第２共起データを素性として入力し、該学習結果を参照して該第１共起データ及び該第２共起データ間の前後を算出することを特徴とする。

【００１５】

請求項９に記載の発明によれば、データ処理装置で処理する対象データが宛名であり、関連データが宛先である構成において、第1関連データと前記第２関連データとの間で、その新旧関係を検出することを特徴とする。

【００１６】

本発明は、次のようなデータ処理方法を提供することもできる。

請求項１０に記載の発明は、所定の対象データに関連する関連データについて、相前後する第１関連データ及び第２関連データの順序を検出するコンピュータのデータ処理方法であって、関連データ抽出手段が、ネットワーク上又はローカルの記憶手段に格納された同一又は異なるファイルから第１関連データ及び第２関連データをそれぞれ抽出する関連データ抽出ステップ、関連データ共起ファイル抽出手段が、ネットワーク上又はローカルの記憶手段から該第１関連データ及び該第２関連データが共起する関連データ共起ファイルを抽出する関連データ共起ファイル抽出ステップ、関連データ間関係検出手段が、該関連データ共起ファイルから、所定の関連データ間関係規則を参照して、第１関連データ及び第２関連データ間の前後を検出する関連データ間関係検出ステップ、出力手段が、該検出結果を出力する出力ステップを有することを特徴とする。

【００１７】

本発明が決定する順序は、関連データ間の新旧、前後、評価、重要度などいかなる順序でもよいが、数値の大小など自然法則によって一義的に定まるものは関係規則による必要はないから、本発明の対象としない。すなわち、本発明が対象とするのは、住所変更による住所の新旧、人手によって並べられたデータの前後、アンケート結果から得られた評価、作成者によってばらばらに決定された重要度など、順序が何らかの作為あるいは精神作用によって決定づけられたものである。

【００１８】

請求項１１に記載の発明は、上記の関連データ間関係規則が、少なくとも前記第１関連データと前記第２関連データとの間に含まれる、又は含まれない、文字列に係る情報であって、上記関連データ間関係検出手段が、前記関連データ共起ファイル内において該第１関連データと該第２関連データとの間の文字列を抽出し、該関連データ間関係規則と照合

することを特徴とする。

【００１９】

請求項１２に記載の発明は、上記の関連データ間関係規則が、予め前後関係が分かっている２つの教師用関連データが共起する複数の教師用ファイルを用い、該教師用ファイルにおける２つの教師用関連データの出現位置、又は同時に含まれる若しくは含まれない文字列、又は同時に含まれるタグ情報の少なくともいずれかを素性として機械学習した学習結果とする構成である。該機械学習には、サポートベクトルマシンや最大エントロピー法の教師有り機械学習処理を行う公知の機械学習モジュールを用いる。

そして、関連データ間関係検出手段が、前記関連データ共起ファイルから該素性を抽出すると共に、前記第１関連データ及び前記第２関連データを入力として、該学習結果を参照して該第１関連データ及び該第２関連データ間の前後を算出することを特徴とする。

【００２０】

請求項１３に記載の発明は、上記の関連データ抽出ステップにおいて関連データ抽出手段が、ネットワーク上又はローカルの記憶手段から対象データと共起する第１関連データ及び第２関連データをそれぞれ抽出することを特徴とする。

【００２１】

請求項１４に記載の発明は、所定の対象データに関連する関連データについて、相前後する第１関連データ及び第２関連データの順序を検出するコンピュータのデータ処理方法であって、関連データ抽出手段が、ネットワーク上又はローカルの記憶手段に格納された同一又は異なるファイルから第１関連データ及び第２関連データをそれぞれ抽出する関連データ抽出ステップ、共起データ抽出手段が、該第１関連データが含まれるファイルから該第１関連データと共起する単数又は複数の第１共起データを抽出すると共に、該第２関連データが含まれるファイルから該第２関連データと共起する単数又は複数の第２共起データを抽出する共起データ抽出ステップ、共起データ間関係検出手段が、該第１共起データ及び該第２共起データ間の前後に関する所定の共起データ間関係規則を参照して、第１共起データ及び第２共起データ間の前後を検出する共起データ間関係検出ステップ、出力手段が、該検出結果をそれらと共起している第１関連データ及び第２関連データの前後として出力する出力ステップを有することを特徴とする。

【００２２】

請求項１５に記載の発明は、上記のデータ処理方法において、前記共起データ抽出ステップの次に、共起データ共起ファイル抽出手段が、ネットワーク上又はローカルの記憶手段から該第１共起データ及び該第２共起データが共起する共起データ共起ファイルを抽出する共起データ共起ファイル抽出ステップを有することを特徴とする。

【００２３】

請求項１６に記載の発明は、上記のデータ処理方法において、共起データ間関係規則が、少なくとも前記第１共起データと前記第２共起データとの間に含まれる、又は含まれない、文字列に係る情報であって、共起データ間関係検出手段が、前記共起データ共起ファイル内において該第１共起データと該第２共起データとの間の文字列を抽出し、該共起データ間関係規則と照合することを特徴とする。

【００２４】

請求項１７に記載の発明は、上記の共起データ間関係規則が、予め前後関係が分かっている２つの教師用関連データのそれぞれと複数の文字列とが含まれる複数の教師用ファイルを用い、該教師用ファイルにおける該教師用関連データと共に含まれる単数又は複数の文字列を素性として機械学習した学習結果とする構成である。該機械学習には、サポートベクトルマシンや最大エントロピー法の教師有り機械学習処理を行う公知の機械学習モジュールを用い、関連データ間関係検出ステップの前に実行処理することができる。

そして、共起データ間関係検出手段が、第１共起データ及び第２共起データを素性として入力し、該学習結果を参照して該第１共起データ及び該第２共起データ間の前後を算出することを特徴とする。

【００２５】

請求項１８に記載の発明は、上記のデータ処理方法が、対象データが宛名であり、関連データが宛先である構成において、第1関連データと第２関連データとの間で、その新旧関係を検出することを特徴とする。

【発明の効果】

【００２６】

本発明は、上記構成を備えることにより次のような効果を奏する。

すなわち、請求項１又は１０に記載の発明によれば、対象データと関連のある２つの関連データ間の順序を関連データ間関係規則に基づいて高精度に決定することができ、従来は人手によって前後の文脈から判断していた処理を自動化することができる。

【００２７】

本発明が対象とする順序は、上記の通り人間の作為や精神作用によって決定づけられたものであるため、本来はコンピュータの処理になじみにくい。

これに対して本発明はまず対象データに関連する２つの関連データを抽出し、さらにそれらが共起する関連データ共起ファイルを抽出する。この方法によれば大量のデータを対象として順序の検出に最適な関連データを抽出し、それと関係規則から高精度に順序を検出することができる。

【００２８】

請求項２又は１１に記載の発明は、関連データ間関係規則として、２つの関連データとの間の文字列が含まれること、あるいは含まれないことを用いるので、コンピュータの文字列比較により簡便に順序を検出することができる。

【００２９】

請求項３又は１２に記載の発明によれば、機械学習を用いて教師データから関連データ間の順序を学習すると共に、その結果を関連データ間関係規則として用いるのでさらに高精度な検出に寄与する。

【００３０】

請求項４又は１３に記載の発明によれば、上記において対象データだけを抽出するのではなく、対象データと共起するデータを抽出することで、関連データと対象データとの関連性がより確実になり、また、共起するデータに限定することで処理すべき対象データ数が抑制される。これにより順序検出の高精度化、処理の高速化を図ることができる。

【００３１】

請求項５又は１４に記載の発明によれば、各関連データから直接順序を検出するのではなく、それらと共起する共起データにより順序を検出することができる。これによって関連データの性質上、順序を決定しにくい場合にも、その共起データを比較することで高精度に検出することができる。

【００３２】

請求項６又は１５に記載の発明によれば、各関連データと共起する共起データが共に出現するファイルを用いることで、共起データ間の関係を正確に把握することができる。

【００３３】

請求項７又は１６に記載の発明によれば、共起データ間関係規則として、２つの共起データとの間の文字列が含まれること、あるいは含まれないことを用いるので、コンピュータの文字列比較により簡便に順序を検出することができる。

【００３４】

請求項８又は１７に記載の発明によれば、機械学習を用いて教師データから共起データ間の順序を学習すると共に、その結果を共起データ間関係規則として用いるのでさらに高精度な検出に寄与する。

【００３５】

請求項９又は１８に記載の発明によれば、対象データとして宛名、関連データとして宛先を用い、変更されることが多く、しかも関連データを較べただけではどちらが新しいかの判定が難しい住所データに対して本発明を適用することができる。

【発明を実施するための最良の形態】

【００３６】
以下、本発明の実施形態を、図面に示す実施例を基に説明する。なお、実施形態は下記に限定されるものではない。
（実施例１）
図１は本発明に係るデータ処理装置（以下、本装置と呼ぶ）の構成図である。本発明は公知のパーソナルコンピュータにより容易に実現することが可能であり、演算処理や機械学習、テキスト処理などを司るＣＰＵ（１０）によって本発明の各ステップを実行処理する。ＣＰＵ（１０）は周知のようにメモリ（図示しない）と協働して動作し、キーボードやマウス（１１）などの入力手段の他、出力結果を表示するモニタ（１２）、ハードディスク等の外部記憶装置（１３）などを備えている。
また、テキストデータ、ファイル等の取得などのためにデータの取得入力手段としてインターネット等のネットワークと接続するネットワークアダプタ（１４）を備える。
【００３７】
そして、ＣＰＵ（１０）には入力部（１０１）、関連データ抽出部（１０２）、関連データ共起ファイル抽出部（１０３）、関連データ間関係検出部（１０４）、出力部（１０５）が設けられている。
そして、公知のプログラミング言語によって記載されたプログラムがＣＰＵ（１０）及びそれと連動するハードウェアを動作させて、以下に説述する各部（１０１）～（１０５）の機能が実現される。
【００３８】
以下、図２に示す処理フローチャートを用いて、請求項１ないし４等に係る本発明の各処理を詳細に説述する。
まず、入力部（１０１）が外部記憶装置（３）あるいはインターネット、ＬＡＮ(Local Area Network)等のサーバ上からネットワークアダプタ（４０）を介して第１コンテンツファイル（２０）及び第２コンテンツファイル（２１）を取得し、ＣＰＵ（１０）内に取り込む処理を行う。
各コンテンツファイル（２０）（２１）は同一の記憶装置やサーバ上にあってもよいし、それぞれ別に格納されているものでもよい。
【００３９】
コンテンツファイル（２０）（２１）の例としては、住所録などの複数の項目に対してそれぞれデータ（氏名・会社名・住所・電話番号）を割り当ててあるデータベースや、ＨＴＭＬ（HyperTextMarkup Language）で記載されたウェブページのソーステキスト、特許公報のウェブページのように、ウェブページであっても項目と内容が正確に対応づけられたデータを含むテキストなどを用いることができる。
分かりやすくするために、以下では対象データを「会社名」、関連データを「住所」として説明を続ける。このように本発明の請求項９等に記載の通り、対象データを宛名、関連データを宛先としたときに、その前後関係として例えば宛先の新旧関係を検出するのに用いることができる。
【００４０】
なお、宛名とは手紙や証書等に書く相手方の氏名、会社名等であり、宛先とは宛名の場所である。例えば、「△△株式会社」が宛名であり、その住所である「東京都中央区駅前１-２-３」が宛先である。
【００４１】
関連データ抽出部（１０２）では、予め定義してある対象データ（２２）と関連する関連データを異なるファイルである各コンテンツファイル（２０）（２１）から抽出する。
（関連データ抽出ステップ：Ｓ１０）
説明上第１関連データ（２４）、第２関連データ（２５）と呼ぶが、これらの順序は未知であり、本発明により対象データ（２２）に関連した２つの関連データ間の順序を検出するものである。
【００４２】

例えば、会社が移転した場合を想定して、「旧住所」（序列が前）、「新住所」（序列が後）を考える。この場合、第１コンテンツファイルは、旧住所が記載されたウェブページ、第２コンテンツファイルは、新住所が記載されたウェブページがあり、それらから旧住所と新住所が第１関連データ、第２関連データとして抽出されることになる。詳しくは後述するが、ここでは対象データ（２２）自体が各コンテンツファイル（２０）（２１）に出現していることは必要なく、そのページ自体に会社名が記載されていなくても予め人手により関連があることは選定され、その上でコンテンツファイルが入力される場合にも本発明は適用される。

【００４３】

各コンテンツファイル（２０）（２１）に含まれるデータが１つであって、予め対象データ（２２）に関連することが確実な各関連データ（２４）（２５）を抽出する構成が最もシンプルな構成であるが、通常はコンテンツファイル（２０）（２１）には複数のデータが含まれる。例えば、旧住所と共に、電話番号や担当者名、製品情報など順序の検出と関係のないデータが含まれている。

本発明では、大量の関連データを比較して順序を決定しても良いので、この段階で真に関連データとして必要であるかを選定する必要はない。

【００４４】

もっとも、処理の高速化、必要な結果のみを得るために、予め選別して抽出を行っても良い。例えば、住所だけを抽出したいのであれば、ＣＰＵ（１０）により公知のテキスト処理を行い、都道府県名や都市名に続き、数字等で終わるテキストなどを抽出すれば住所だけを簡単に抽出することもできる。

さらに、「本社」に続く文字列だけを抽出することで、複数の住所が記載されているページから、内容の等価性が予想される１つの関連データを特定して抽出してもよい。

【００４５】

抽出された第１関連データ（２４）及び第２関連データ（２５）を用い、関連データ共起ファイル抽出部（１０３）において、コンテンツデータ（２３）からそれらが共起する関連データ共起ファイル（２３１）を抽出する。（関連データ共起ファイル抽出ステップ：Ｓ１１）

【００４６】

コンテンツデータ（２３）は、外部記憶装置（１３）に格納されていても、インターネット等のサーバ上に格納されていてもいずれでも良いが、データ量が多いほど共起するデータが確実に抽出できることから、後者の方が好ましい。

共起するファイルが複数ある場合には、全てを抽出して次の処理に進んでもよいし、ファイルの作成日時が最新のもの１つ、あるいは２つの関連データの占める割合が高いものとして全体のデータ容量が最も小さいもの１つを選んでもよい。

【００４７】

関連データ共起ファイル抽出部（１０３）は、予めどこを検索するかを定めておく場合に限らず、まず公知の検索エンジンのサイトに、各関連データを送信し、それらが共起するウェブページを検索した上で、そのウェブページを関連データ共起ファイル（２３１）として抽出してもよい。

【００４８】

次いで、関連データ間関係検出部（１０４）において、外部記憶装置に格納された関連データ間関係規則（１３０）を参照し、関連データ間の順序を検出する。（関連データ間関係検出ステップ：Ｓ１２）

本発明では請求項２等に記載の発明のようにルールベースによる方法と、請求項３等に記載の発明のように機械学習による方法の２つを提案する。

【００４９】

まずルールベースによる方法から説明する。

ルールベースの場合、予め人手によって規則を定めておき、それに従って判定を行うが、本発明のように自動的にコンテンツデータ（２３）を参照して関連データ間の順序を決

定することは、コンテンツデータ（２３）が膨大であると事実上不可能である。本発明はこのような場合にも高精度に順序を検出することができる。

【００５０】

本発明請求項に係る関連データ共起ファイル（２３１）の例を図３ないし図５に示す。なおこの関連データ共起ファイル（２３１）は後述の各実施例においても共通に用いることのできる例である。

図３（Ａ）は抽出されたウェブページの１例（２３１ａ）を示しており、様々なテキストの中で「△△は、下記に移転します。」（２２ａ）との表示の後に、「新住所：×××××」（２５ａ）、「現住所：●●●●●」（２４ａ）の順番に記載されている。

図３（Ｂ）は、別の表示例（２３１ｂ）であり、「△△社屋移転のお知らせ」（２２ｂ）との表示の後に、「旧住所：●●●●●」（２４ｂ）、「新住所：××××××」（２５ｂ）の順番に記載されている。

【００５１】

明らかなように、本発明の対象データは△△であり、第１関連データは旧住所の●●●●●、第２関連データは新住所の××××××である。

実際には住所変更の場合には多くの表記方法があるが、それらも含めて次のような関連データ間関係規則（１３０）を用意する。本実施例の関連データ間関係規則（１３０）は請求項２や７などのルールベースによる方法で共通に用いることの出来る関連データ間関係規則の一例である。

【００５２】

【表１】

関連データ間関係規則の例１

番号	介在文字列	前出の関連データ	後出の関連データ
1	現住所	後	前
2	旧住所	後	前
3	・・から	前	後
4	・・より	前	後
5	・・へ	後	前
6	新住所	前	後
7	移転先	前	後

【００５３】

関連データ間関係検出部（１０４）では、図３（Ａ）の場合には第２関連データ（２５ａ）が前出、第１関連データ（２４ａ）が後出であり、その間に含まれる文字列から関連データ間関係規則（１３０）に含まれる文字列「現住所」が発見できることから、番号１の規則を適用して、第２関連データ××××××（２５ａ）が後、第１関連データ●●●●●（２４ａ）が前と検出する。

【００５４】

また図３（Ｂ）の場合には第１関連データ（２４ｂ）が前出、第２関連データ（２５ｂ）が後出であり、その間に含まれる文字列から関連データ間関係規則（１３０）に含まれる文字列「新住所」が発見できることから、番号６の規則を適用して、第１関連データ●●●●●（２４ｂ）が前、第２関連データ××××××（２５ｂ）が後、と検出する。

なお、上記関連データ間関係規則（１３０）では含まれる文字列のみを定義したが、逆に含まれない文字列を定義してもよい。

【００５５】

次に図４ではウェブページにおける表を利用して、項目名として「旧住所」「新住所」が記載されて、その下欄に第１関連データ（２４ｃ）と第２関連データ（２５ｃ）が記載されている。

このような場合に、ＨＴＭＬにおけるタグを利用して関連データ間関係規則（１３０）とすることもできる。例えば、関連データ間に表の枠線のタグが介在する場合には、その

左側の関連データを前、右側の関連データを後とすることができる。そのほか、「旧住所」の文字列の下欄又は左欄にある関連データを前、「新住所」の文字列の下欄又は左欄にある関連データを後と定義してもよい。

【００５６】

その他、図５のように第１関連データ（２４ｄ）と第２関連データ（２５ｄ）の文字の大きさが異なる場合に、関連データ間関係規則（１３０）に、文字サイズが小さなものを前、大きなものを後とする規則を備えておいて、文字サイズを指定するタグから順序を検出してもよい。

【００５７】

本発明では関連データ共起ファイル（２３１）は複数抽出してもよいから、以上のような関連データ間関係規則（１３０）に複数の条件が合致する場合がある。このような場合には単純には多数決により「前」と判定された数が多い関連データが前、「後」と判定された数が多い関連データを後とすればよい。

また、関連データ間関係規則（１３０）に表２のように重みを定義しておき、例えば番号２と４と５が抽出された場合には、前出関連データが前である確度は０．４、後である確度は０．８＋０．４＝１．２（後出関連データについてはこの逆）として、確度の高い後、と判定するようにしてもよい。

【００５８】

【表２】

関連データ間関係規則の例２

番号	介在文字列	前出の関連データ	後出の関連データ	重み
1	現住所	後	前	0.5
2	旧住所	後	前	0.8
3	・・から	前	後	0.7
4	・・より	前	後	0.4
5	・・へ	後	前	0.4
6	新住所	前	後	0.8
7	移転先	前	後	0.9

【００５９】

本発明は、このように関連データ間関係規則（１３０）を使うとしても単にルールに従って判定するだけでなく、多量のデータに基づいて、どちらがより前らしいか、後らしいかを含めて検出することができる点に特徴を有する。

【００６０】

検出結果は出力部（１０５）から出力される。（出力ステップ：Ｓ１３）

本発明における出力としては、モニタ（１２）からの表示や、外部記憶装置（１３）への記録、ネットワークアダプタ（１４）を介して外部サーバに出力などいずれでもよい。本発明のデータ処理装置（１）を、データ検索装置に装備し、検索結果の表示順を本装置（１）の検出した順序に合わせて変更するように利用してもよい。

【００６１】

図２において、本実施例では異なる２つのコンテンツファイル（２０）（２１）を入力したが、同一のコンテンツファイルを関連データ抽出部（１０２）に入力して、２つの関連データを抽出してもよい。この場合、そもそもコンテンツファイルにおいて関連データが共起していることから、これも他のコンテンツデータ（２３）と共に、関連データ間関係検出部（２０４）で用いてもよい。

【００６２】

また、請求項４等に記載の発明の実施態様として、関連データ抽出部（１０２）では対象データと共起する関連データを抽出する構成でもよい。上記したとおり、第１コンテンツファイル（２０）等が予め対象データと関連があることが分かっている場合には必要ないが、コンテンツファイルをインターネット等から抽出する場合には、対象データ（２２

）が出現するファイルを抽出し、これらをコンテンツファイルとする必要がある。
【００６３】
この場合、単に対象データ（２２）が出現するコンテンツファイル内の全ての文字列を関連データ（２４）（２５）としてもよいが、より好ましくは、対象データ（２２）が出現する前後所定の文字数内の文字列を関連データ（２４）（２５）としてもよい。これにより、一般的に関連が高いと思われる近傍の文字列を関連データとすることができる。
【００６４】
本発明における関連データや共起データの抽出には次のような高度な手法を適用することもできる。
共起データを例に挙げると、関連データを構成する単語群Ａ（単語群は単数又は複数の単語を言う。）を、多く含む共起データの抽出方法を説明する。
【００６５】
(1) 基本的な方法 (TF・IDF 法) の説明
（数１）

$$score(D) = \Sigma_{w \in W \text{ で加算}} (tf(w,D) * \log(N/df(w)))$$

Wは関連データの集合、tf(w,D)はコンテンツデータ中におけるwの出現回数、df(w)は全文書でWが出現した文書の数、Nは文書の総数
数１に示す式において、score(D) が高い文書データを共起データとして出力する。このようにすることで、関連データとして一般的な語句を多数抽出してしまった場合、意味のない共起データが多数抽出されることを防ぐことができる。
同様に関連データを抽出する際にも有意な関連データを抽出するのに寄与させることができる。
【００６６】
(2)Robertson らの Okapi weightingの説明
本方法は、非特許文献３に記載されている。
【００６７】
【非特許文献３】村田真樹,馬青,内元清貴,小作浩美,内山将夫,井佐原均“位置情報と分野情報を用いた情報検索”自然言語処理(言語処理学会誌) 2000年 4月,7 巻,2 号, p.141 ～ p.160　該非特許文献１３における数２が性能がよいことが知られている。　そして、Σで積を取る前の tf 項とidf 項の積が Okapiのウェイティング法になって、この値を単語の重みに使う。
【００６８】
Okapi の式なら
（数２）

$$score(D) = \Sigma_{w \in W \text{ で加算}} (tf(w,D)/(tf(w,D) + length/delta) * \log(N/df(w)))$$

lengthはデータＤの長さ、delta はデータの長さの平均、データの長さは、データのバイト数、また、データに含まれる単語数などを使う。
【００６９】
さらに、以下の情報検索を行うこともできる。
(Okapi の参考文献)
非特許文献４, ５に開示されるようなOkapiの式、SMARTの式を用いることもできる。より高度な情報検索の方法として、tf・idf を使うだけの式でなく、これらの OkapiのSMARTの式を用いてもよい。
【００７０】
【非特許文献４】S. E. Robertson, S. Walker, S. Jones, M. M.Hancock-Beaulieu, and M. GatfordOkapi at TREC-3, TREC-3, 1994年
【非特許文献５】Amit Singhal AT&T at TREC-6, TREC-6,1997 年

【００７１】
これらの方法では、tf・idf だけでなく、コンテンツデータの長さなども利用して、より高精度な情報検索を行うことができる。
【００７２】
今回の、単語群Ａをより多く含む共起データの抽出方法では、さらに、Rocchio's formula （非特許文献６）を使うことができる。
【００７３】
【非特許文献６】J. J. Rocchio,Relevance feedback ininformation retrieval,The SMART retrieval System, Edited by G. Salton,PrenticeHall, Inc.,page 313-323, 1971年
【００７４】
この方法は、log(N/df(w))のかわりに、
（数３）
{E(t)+ k_af * (RatioC(t) - RatioD(t))} *log(N/df(w))
を使う。
【００７５】
E(t) = 1 (対象データ)
= 0 (それ以外)
RatioC(t) は関連データ群Ｂでのt の出現率
RatioD(t) はコンテンツデータ群Cでのt の出現率
log(N/df(w))を上式でおきかえた式でscore(D)を求めて、その値が大きいものほど単語群Aをより多く含む共起データとして取り出すものである。
【００７６】
score(D)のΣの加算の際に足す単語wの集合Wは、元の対象データと、単語群Aの両方とする。ただし、元の対象データと、単語群Aは重ならないようにする。
【００７７】
また、他の方法として、score(D)のΣの加算の際に足す。単語wの集合Wは、単語群Ａのみとする。ただし、元の対象データと、単語群Ａは重ならないようにする。
【００７８】
ここでは roccio の式で複雑な方法をとったが、単純に、単語群Aの単語の出現回数の和が大きいものほど、単語群Aをより多く含む共起データとして取り出すようにしてもよいし、また、単語群Aの出現の異なりの大きいものほど、単語群Aをより多く含む共起データとして取り出すようにしてもよい。
以上の方法により、単語群Aを含む共起データを取り出すことができる。
【００７９】
（実施例２）
ルールベースを用いた実施例１に続いて、機械学習を用いた手法を実施例２として説明する。以下の実施例は本発明の請求項３等の技術に関する。
本実施例における関連データ間関係検出部（１０４）のさらに詳細な構成を図６に示す。ここでは教師データ入力部（1041）、解-素性対抽出部（1042）、機械学習処理部（1043）、関係判定部（1044）がそれぞれ設けられる。
【００８０】
機械学習の手法は公知の機械学習モジュールにおける学習過程と、それを用いた解の推定過程とが一体的に成り立つものである。本発明の実施において、学習過程は必ずしも必須ではなく外部記憶装置（１３）には機械学習の結果形成された関連データ間関係規則（１３０）を備えておくだけでもよい。その場合には、関連データ間関係検出部（１０４）に必要なのは各機械学習手法に従って順序を判定する関係判定部（1044）だけである。
【００８１】
機械学習の手法は、様々なものが公知であるが、ここでは各手法を簡単に説明する。問題-解の組のセットを多く用意し、それで学習を行ない、どういう問題のときにどういう解になるかを学習し、その学習結果を利用して、新しい問題のときも解を推測できるよう

にする方法である(例えば、下記の非特許文献７～非特許文献９参照)。

【００８２】

【非特許文献７】村田真樹,機械学習に基づく言語処理,龍谷大学理工学部.招待講 演.2004.http://www2.nict.go.jp/x/x161/member/murata/ps/kougi-ml-siryou-new2.pdf

【非特許文献８】サポートベクトルマシンを用いたテンス・アスペクト・モダリティの日英翻訳,村田真樹,馬青,内元清貴,井佐原均,電子情報通信学会言語理解とコミュニケーション研究会 NLC2000-78 ,2001年.

【非特許文献９】SENSEVAL2J辞書タスクでのCRLの取り組み,村田真樹,内山将夫,内元清貴,馬青,井佐原均,電子情報通信学会言語理解とコミュニケーション研究会 NLC2001-40 ,2001年.

【００８３】

どういう問題のときに、という、問題の状況を機械に伝える際に、素性(解析に用いる情報で問題を構成する各要素)が必要になる。問題を素性によって表現するのである。例えば、日本語文末表現の時制の推定の問題において、

問題:「彼が話す。」---解「現在」

が与えられた場合に、素性の一例は、「彼が話す。」「が話す。」「話す。」「す。」「。」となる。

【００８４】

すなわち、機械学習の手法は、素性の集合-解の組のセットを多く用意し、それで学習を行ない、どういう素性の集合のときにどういう解になるかを学習し、その学習結果を利用して、新しい問題のときもその問題から素性の集合を取り出し、その素性の場合の解を推測する方法である。

【００８５】

図６に示すようにＣＰＵ（１０）において、関係判定部（1044）で処理する前段として、解-素性対抽出部（1042）と、機械学習処理部（1043）を備える。ここで機械学習処理は、図７のように分散したテキストデータをどのように分類するのか、その分類結果（解）を得る。

機械学習処理部（1043）における機械学習の手法として、例えば、ｋ近傍法、シンプルベイズ法、決定リスト法、最大エントロピー法、サポートベクトルマシン法などの手法を用いる。

【００８６】

ｋ近傍法は、最も類似する一つの事例のかわりに、最も類似するｋ個の事例を用いて、このｋ個の事例での多数決によって分類先(解)を求める手法である。ｋは、あらかじめ定める整数の数字であって、一般的に、１から９の間の奇数を用いる。

【００８７】

シンプルベイズ法は、ベイズの定理にもとづいて各分類になる確率を推定し、その確率値が最も大きい分類を求める分類先とする方法である。

【００８８】

シンプルベイズ法において、文脈bで分類aを出力する確率は、以下の数４で与えられる。

【００８９】

【数４】

$$p(a|b)=\frac{p(a)}{p(b)}\,p(b|a)$$

【００９０】

【数５】

$$\cong \frac{\tilde{p}(a)}{p(b)} \prod_i \tilde{p}(f_i|a)$$

【００９１】

ただし、ここで文脈bは、あらかじめ設定しておいた素性f_j (∈F,1≦j≦k)の集合である。p(b)は、文脈bの出現確率である。ここで、分類aに非依存であって定数のために計算しない。P(a)(ここでPはpの上部にチルダ)とP(f_i|a)は、それぞれ教師データから推定された確率であって、分類aの出現確率、分類aのときに素性f_iを持つ確率を意味する。P(f_i|a)として最尤推定を行って求めた値を用いると、しばしば値がゼロとなり、数５の値がゼロで分類先を決定することが困難な場合が生じる。そのため、スームージングを行う。ここでは、以下の数６を用いてスームージングを行ったものを用いる。

【００９２】

【数６】

$$p(f_i|a) = \frac{freq(f_i,a) + 0.01 * freq(a)}{freq(a) + 0.01 * freq(a)}$$

【００９３】

ただし、freq(f_i,a)は、素性f_iを持ちかつ分類がaである事例の個数、freq(a)は、分類がaである事例の個数を意味する。

【００９４】

決定リスト法は、素性と分類先の組とを規則とし、それらをあらかじめ定めた優先順序でリストに蓄えおき、検出する対象となる入力が与えられたときに、リストで優先順位の高いところから入力のデータと規則の素性とを比較し、素性が一致した規則の分類先をその入力の分類先とする方法である。

【００９５】

決定リスト方法では、あらかじめ設定しておいた素性f_j(∈F,1≦j≦k)のうち、いずれか一つの素性のみを文脈として各分類の確率値を求める。ある文脈bで分類aを出力する確率は以下の数７によって与えられる。

【００９６】

(数７)

p(a|b)=p(a|fmax)

ただし、fmax は以下の数８によって与えられる。

【００９７】

【数８】

$$f_{\max} = \arg\max_{f_j \in F} \max_{a_i \in A} \tilde{p}(a_i|f_j)$$

【００９８】

また、P(a_i|f_j)(ここでPはpの上部にチルダ)は、素性f_jを文脈に持つ場合の分類a_iの出現の割合である。

【００９９】

最大エントロピー法は、あらかじめ設定しておいた素性f_j (1≦j≦k)の集合をFとするとき、以下所定の条件式（数９)を満足しながらエントロピーを意味する数１０を最大にするときの確率分布p(a,b)を求め、その確率分布にしたがって求まる各分類の確率のうち、最も大きい確率値を持つ分類を求める分類先とする方法である。

【０１００】

【数９】

$$\sum_{a\in A,b\in B} p(a,b)g_j(a,b)=\sum_{a\in A,b\in B} \tilde{p}(a,b)g_j(a,b)$$

$$for\ \forall f_j(1\le j\le k)$$

【数１０】

$$H(p) = -\sum_{a\in A,b\in B} p(a,b)\log(p(a,b))$$

【０１０１】

ただし、A、Bは分類と文脈の集合を意味し、$g_j(a,b)$は文脈bに素性f_jがあって、なおかつ分類がaの場合1となり、それ以外で0となる関数を意味する。また、$P(a_i|f_j)$(ここでPはpの上部にチルダ)は、既知データでの(a,b)の出現の割合を意味する。

【０１０２】

数９は、確率pと出力と素性の組の出現を意味する関数gをかけることで出力と素性の組の頻度の期待値を求めることになっており、右辺の既知データにおける期待値と、左辺の求める確率分布に基づいて計算される期待値が等しいことを制約として、エントロピー最大化(確率分布の平滑化) を行なって、出力と文脈の確率分布を求めるものとなっている。最大エントロピー法の詳細については、以下の非特許文献１０に記載されている。

【０１０３】

【非特許文献１０】Eric Sven Ristad, Maximum Entropy Modelingfor Natural Language ,(ACL/EACL Tutorial Program, Madrid, 1997

【０１０４】

サポートベクトルマシン法は、空間を超平面で分割することにより、二つの分類からなるデータを分類する手法である。図８にサポートベクトルマシン法のマージン最大化の概念を示す。図８において、白丸は正例、黒丸は負例を意味し、実線は空間を分割する超平面を意味し、破線はマージン領域の境界を表す面を意味する。図８（Ａ）は、正例と負例の間隔が狭い場合(スモールマージン)の概念図、図８（Ｂ）は、正例と負例の間隔が広い場合(ラージマージン)の概念図である。

【０１０５】

このとき、二つの分類が正例と負例からなるものとすると、学習データにおける正例と負例の間隔(マージン) が大きいものほどオープンデータで誤った分類をする可能性が低いと考えられ、図８（Ｂ）に示すように、このマージンを最大にする超平面を求めそれを用いて分類を行なう。

【０１０６】

基本的には上記のとおりであるが、通常、学習データにおいてマージンの内部領域に少数の事例が含まれてもよいとする手法の拡張や、超平面の線形の部分を非線型にする拡張(カーネル関数の導入) がなされたものが用いられる。

【０１０７】

この拡張された方法は、以下の識別関数（数１１)を用いて分類することと等価であり、その識別関数の出力値が正か負かによって二つの分類を判別することができる。

【０１０８】

【数１１】

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l}\alpha_i y_i K(x_i, x)+b\right)$$

$$b = -\frac{\max_{i,y_i=-1} b_i + \min_{i,y_i=-1} b_i}{2}$$

$$b_i = \sum_{j=1}^{l}\alpha_j y_j K(x_j, x_i)$$

【０１０９】

ただし、xは識別したい事例の文脈(素性の集合）を、x_iとy_j（i=1,...,l,y_j∈{1,-1}）は学習データの文脈と分類先を意味し、関数sgnは、

sgn(x)=1(x≧0)

-1(otherwise)

であり、また、各α_iは数１３と数１４の制約のもと数１２を最大にする場合のものである。

【０１１０】

【数１２】

$$L(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i, x_j)$$

【数１３】

$$0 \le \alpha_i \le C \quad (i = 1,...,l)$$

【数１４】

$$\sum_{i=1}^{l}\alpha_i y_i = 0$$

【０１１１】

また、関数Kはカーネル関数と呼ばれ、様々なものが用いられるが、本形態では以下の多項式のものを用いる。

【０１１２】

（数１５）

K(x,y)=(x・y+1)d

C、dは実験的に設定される定数である。例えば、Cはすべての処理を通して1に固定した。また、dは、1と2の二種類を試している。ここで、α_i>0となるx_iは、サポートベクトルと呼ばれ、通常、数８の和をとっている部分は、この事例のみを用いて計算される。つまり、実際の解析には学習データのうちサポートベクトルと呼ばれる事例のみしか用いられない。

【０１１３】

なお、拡張されたサポートベクトルマシン法の詳細については、以下の非特許文献１１および非特許文献１２に記載されている。

【０１１４】

【非特許文献１１】Nello Cristianini and John Shawe-Taylor, AnIntroduction to Support Vector Machines and other kernel-based learningmethods,(Cambridge University Press,2000)
【非特許文献１２】Taku Kudoh, Tinysvm:Support Vectormachines,(http://chasen.org/~taku/software/TinySVM/,2002年)
【０１１５】
サポートベクトルマシン法は、分類の数が２個のデータを扱うものである。したがって、分類の数が3個以上の事例を扱う場合には、通常、これにペアワイズ法またはワンVSレスト法などの手法を組み合わせて用いることになる。
【０１１６】
ペアワイズ法は、n個の分類を持つデータの場合に、異なる二つの分類先のあらゆるペア(n(n-1)/2個)を生成し、各ペアごとにどちらがよいかを二値分類器、すなわちサポートベクトルマシン法処理モジュールで求めて、最終的に、n(n-1)/2個の二値分類による分類先の多数決によって、分類先を求める方法である。
【０１１７】
ワンVSレスト法は、例えば、a、b、cという三つの分類先があるときは、分類先aとその他、分類先bとその他、分類先cとその他、という三つの組を生成し、それぞれの組についてサポートベクトルマシン法で学習処理する。そして、学習結果による推定処理において、その三つの組のサポートベクトルマシンの学習結果を利用する。推定するべき候補が、その三つのサポートベクトルマシンではどのように推定されるかを見て、その三つのサポートベクトルマシンのうち、その他でないほうの分類先であって、かつサポートベクトルマシンの分離平面から最も離れた場合のものの分類先を求める解とする方法である。例えば、ある候補が、「分類先aとその他」の組の学習処理で作成したサポートベクトルマシンにおいて分離平面から最も離れた場合には、その候補の分類先は、aと推定する。
【０１１８】
以上のように機械学習の手法は様々であるが、本発明はそのいずれも関連データ間関係検出部（１０４）に利用することができる。すなわち、関連データ間の関係、例えば２つの関連データを連続して入力した時に、その順序が正しければ１、正しくなければ０という解、さらにその確からしさを解として求めることができる。
【０１１９】
学習の際には教師データ入力部（1041）が予め用意してある教師データを入力する。教師データは、外部記憶装置等に格納しておけばよい。教師データには、予め前後の分かっている２つの関連データが含まれており、解としては例えば含まれている順序が正しければ１、含まれている順序と正解が逆のときには０と考えればよい。この解の定め方は機械学習方法や必要となる結果に応じて適宜設計することができる。
その上で、上記した機械学習方法のいずれかによって解の求め方は次のように異なる。
【０１２０】
例えば、本発明の実施の形態において、機械学習処理部（1043）が、機械学習の手法としてk近傍法を用いる場合、機械学習処理部（1043）は、教師データ入力部（1041）で入力した教師データから抽出された素性の集合のうち重複する素性の割合(同じ素性をいくつ持っているかの割合)にもとづく事例同士の類似度を定義して、前記定義した類似度と事例とを学習結果情報として関連データ間関係規則（１３０）に記憶しておく。
【０１２１】
そして、関係判定部（1044）は、関連データ共起ファイル（２３１）から解-素性対抽出部（1042）が抽出したデータについて、関連データ間関係規則（１３０）において定義された前後関係の正誤の確率と、素性とを参照して、そのデータが正解である可能性が高い順にk個の素性を関連データ間関係規則（１３０）の事例から選択し、選択したk個の素性での多数決によって正しいか否かという分類先を、解として推定する。
【０１２２】
すなわち、関係判定部（1044）では、抽出された各データに対して、どのような解(分

類先)になりやすいかの度合いを、選択したk個の素性での多数決の票数、ここでは「正しい（関連データ共起ファイル内の関連データの序列が正しい順序である）」という分類が獲得した票数とする。この票数が過半数以下であれば、逆に出現順と逆が正しいことになる。

【０１２３】

また、機械学習手法として、シンプルベイズ法を用いる場合には、機械学習処理部（1043）は、教師データの事例について、前記事例の解と素性の集合との組を学習結果情報として関連データ間関係規則（１３０）に記憶する。

【０１２４】

そして、関係判定部（1044）は、関連データ共起ファイル抽出部（１０３）が関連データ共起ファイル（２３１）を抽出したときに、関連データ間関係規則（１３０）の学習結果情報の解と素性の集合との組をもとに、ベイズの定理にもとづいて解-素性対抽出部（1042）で取得した素性の集合について、出現順が正解か否かに係わる各分類になる確率を算出して、その確率の値が最も大きい分類を、そのデータについての素性の分類(解)と推定する。

【０１２５】

すなわち、関係判定部（1044）では、抽出されたデータについての素性の集合の場合にある解となりやすさの度合いを、各分類になる確率、ここでは「出現順が正しい」という分類になる確率とする。

【０１２６】

機械学習手法として決定リスト法を用いる場合には、機械学習処理部（1043）は、教師データの事例について、素性と分類先との規則を所定の優先順序で並べたリストを関連データ間関係規則（１３０）に記憶する。そして、関連データ共起ファイル抽出部（１０３）が関連データ共起ファイル（２３１）を抽出したときに、関係判定部（1044）は、関連データ間関係規則（１３０）のリストの優先順位の高い順に、抽出された表現対の候補の素性と規則の素性とを比較し、素性が一致した規則の分類先をその候補の分類先(解)として推定する。

【０１２７】

すなわち、関係判定部（1044）では、抽出されたデータについてその素性の集合の場合にある解となりやすさの度合いを、所定の優先順位またはそれに相当する数値、尺度、ここでは「出現順が正しい」という分類になる確率のリストにおける優先順位とする。

【０１２８】

また、機械学習手法として最大エントロピー法を使用する場合には、機械学習処理部（1043）は、教師データの事例から解となりうる分類を特定し、所定の条件式を満足しかつエントロピーを示す式を最大にするときの素性の集合と解となりうる分類の二項からなる確率分布を求めて関連データ間関係規則（１３０）に記憶する。そして、関連データ共起ファイル抽出部（１０３）が関連データ共起ファイル（２３１）を抽出したときに、関係判定部（1044）は、関連データ間関係規則（１３０）の確率分布を利用して、抽出されたファイルについてその素性の集合についてその解となりうる分類の確率を求めて、最も大きい確率値を持つ解となりうる分類を特定し、その特定した分類をその候補の解と推定する。すなわち、関係判定部（1044）では、抽出されたデータについてその素性の集合の場合にある解となりやすさの度合いを、各分類になる確率、ここでは「出現順が正しい」という分類になる確率とする。

【０１２９】

機械学習手法としてサポートベクトルマシン法を使用する場合には、機械学習処理部（1043）は、教師データの事例から解となりうる分類を特定し、分類を正例と負例に分割して、カーネル関数を用いた所定の実行関数にしたがって事例の素性の集合を次元とする空間上で、その事例の正例と負例の間隔を最大にし、かつ正例と負例を超平面で分割する超平面を求めて関連データ間関係規則（１３０）に記憶する。

本実施例の関連データ間関係規則（１３０）は請求項３等の機械学習を用いた方法で共

通に用いることの出来る関連データ間関係規則の一例である。

【０１３０】

そして関連データ共起ファイル抽出部（１０３）が関連データ共起ファイル（２３１）を抽出したときに、関係判定部（1044）は、関連データ間関係規則（１３０）の超平面を利用して、抽出されたデータについての素性の集合が超平面で分割された空間において正例側か負例側のどちらにあるかを特定し、その特定された結果にもとづいて定まる分類を、その候補の解と推定する。

【０１３１】

すなわち、関係判定部（1044）では、抽出されたデータについてその素性の集合の場合にある解となりやすさの度合いを、分離平面からの正例(出現順が正しいデータ)の空間への距離の大きさとする。より詳しくは、出現順が正しいデータを正例、風評情報ではないデータを負例とする場合に、分離平面に対して正例側の空間に位置するデータが「出現順が正しいデータ」と判断され、その事例の分離平面からの距離をそのデータの出現順が正しい度合いとする。

【０１３２】

さらに、本発明では機械学習の手法として、公知のニューラルネットワークによる方法、重回帰分析による方法を用いることもできる。

例えば、求める分類が２種類であれば重回帰分析を利用することができる。重回帰分析をコンピュータ上で実行する方法については、非特許文献１３に詳しい。

【０１３３】

【非特許文献１３】「Excelで学ぶ時系列分析と予測」３章,オーム社

【０１３４】

重回帰分析の場合は、素性の数だけ説明変数xを用意し、素性のありなしを、その説明変数xの値を1,0で表現する。目的変数（被説明変数）は、ある分類の場合を値１、他の分類の場合を値0として求めればよい。

【０１３５】

以上に説述した通り、本発明は公知の任意の機械学習手法を備えた機械学習モジュールを用いることで、関連データ間関係規則（１３０）を生成した上で、関係判定部（1044）が、出現順が正解か否かを的確に判定する。

出現順が正解か否かは、上述したように機械学習手法によって「出現順が正しい」「出現順と正しい順序は逆である」のいずれかで出力される場合もあるし、「出現順が正しい確率」が出力される場合もある。「出現順が正しい確率」が大きな順にその確率と共に出力されてもよい。また、確率を示すための書式、例えば、文字色や文字サイズ、あるいは確率を示すマークなどと共に出力されてもよい。

【０１３６】

本実施例において、ルールベースに基づく方法や機械学習を用いる方法のいずれにおいても、確率が最も高いものや、高い方から順に所定の個数を取り出すことができる。また、ある閾値を設定して、その閾値以上のものを抽出することもできる。所定の閾値以上のもので、かつ確率が高い方から所定の個数だけを抽出してもよい。このようにデータをどのような基準で抽出するかは本発明において任意である。

閾値や所定の個数は予め本装置に備えて固定してもよいし、ユーザが変更できるようにしてもよい。

【０１３７】

本発明では、素性として２つの関連データの出現位置を用いることができる。ここでいう出現位置とは絶対的な位置の他、２つの関連データの相対的な位置も含まれる。絶対的な出現位置とは、例えばファイル内の関連データが始まる文字数、行、列などである。相対的な位置とは、どちらが前後にあるかの他、何文字前（後）にあるかを示す文字数などである。

【０１３８】

素性としては、関連データと同時に含まれる文字列を用いることもできる。例えば前述

のルールベースで示したような「新住所」「旧住所」などの文字列が含まれているか、あるいは何が含まれているかを素性とすることができる。
また、逆に含まれていない、ということも素性にすることができる。すなわち文字列「変更」や「移転」が含まれていないことは、その関連データ共起ファイルにおいてそもそも関連データ間の前後関係を示していない可能性を示唆するものであり、それによって確率を算出する素性として用いることができる。
【０１３９】
同時に含まれる文字列は、関連データの一部の文字列であってもよい。例えば、教師用関連データの一部に、古いビル名と新しいビル名が含まれているような場合、それらの文字列を素性としておくことで、新しいビル名が後のデータであることの検出に寄与する。市町村合併などによる住居表示変更の場合にも同様に検出することができる。
【０１４０】
タグ情報を素性としてもよい。上述した構成と同様に、関連データの表の枠線のタグや、フォントを設定するタグなどを素性とすることができる。
これらは単独で用いるだけでなく、組み合わせて素性とすることができる。例えば、図４のような表において、「旧住所」「新住所」という同時に含まれる文字列と、それぞれの直下に各関連データが配置されているというタグ情報、さらに第１関連データ（２４ｃ）が前で、第２関連データ（２５ｃ）がその直後という相対的位置関係をすべて素性とすることができる。
【０１４１】
本発明の関連データ抽出部（１０２）において特徴的な関連データを効率よく抽出するために、固有表現の抽出技術を用いてもよい。すなわち、本発明のＣＰＵ（１０）に図示しない判定対象名詞抽出部を備えて、第１コンテンツファイル（２０）、第２コンテンツファイル（２１）からそれぞれ固有表現を抽出する。以下簡単に説明する。
【０１４２】
(１) 固有表現抽出のために機械学習を用いる手法
機械学習を用いて固有表現を抽出する手法がある(例えば、以下の非特許文献１４参照)。
【０１４３】
【非特許文献１４】浅原正幸,松本裕治,日本語固有表現抽出における冗長的な形態素解析の利用情報処理学会自然言語処理研究会 NL153-7 2002年
【０１４４】
まず、例えば、「日本の首相は小泉さんです。」という文を、各文字に分割し、分割した文字について、以下のように、 B-LOCATION、 I-LOCATION等の正解タグを付与することによって、正解を設定する。以下の一列目は、分割された各文字であり、各文字の正解タグは二列目である。

日 B-LOCATION
本 I-LOCATION
の 0
首 0
相 0
は 0
小 B-PERSON
泉 I-PERSON
さ 0
ん 0
で 0
す 0
。 0

上記において、B-???は、ハイフン以下の固有表現の種類の始まりを意味するタグである。例えば、B-LOCATIONは、地名という固有表現の始まりを意味しており、B-PERSONは、人名という固有表現の始まりを意味している。また、I-???は、ハイフン以下の固有表現の種類の始まり以外を意味するタグであり、O はこれら以外である。従って、例えば、文字「日」は、地名という固有表現の始まりに該当する文字であり、文字「本」までが地名という固有表現である。

【０１４５】

このように、各文字の正解を設定しておき、このようなデータから学習し、新しいデータでこの正解を推定し、この正解のタグから、各固有表現の始まりと、どこまでがその固有表現かを認識して、固有表現を推定する。

【０１４６】

この各文字に設定された正解のデータから学習するときには、システムによってさまざまな情報を素性という形で利用する。例えば、

日 B-LOCATION

の部分は、

日本-B 名詞-B

などの情報を用いる。日本-B は、日本という単語の先頭を意味し、名詞-Bは、名詞の先頭を意味する。単語や品詞の認定には、例えば前述したChasenによる形態素解析を用いる。上述したChasenは各単語の品詞も推定することができるので、「学校へ行く」を入力すると以下の結果を得る。

【０１４７】

学校　ガッコウ　学校　名詞-一般
へ　　へ　　へ　助詞-格助詞-一般
行く　イク　　行く　動詞-自立 五段・カ行促音便 基本形
EOS

このように各行に一個の単語が入るように分割され、各単語に読みや品詞の情報が付与される。

【０１４８】

なお、例えば、上記の非特許文献１４では、素性として、入力文を構成する文字の、文字自体(例えば、「小」という文字)、字種(例えば、ひらがなやカタカナ等)、品詞情報、タグ情報(例えば、「B-PERSON」等)を利用している。

【０１４９】

これら素性を利用して学習する。タグを推定する文字やその周辺の文字にどういう素性が出現するかを調べ、どういう素性が出現しているときにどういうタグになりやすいかを学習し、その学習結果を利用して新しいデータでのタグの推定を行なう。機械学習には、例えばサポートベクトルマシンを用いる。

【０１５０】

固有表現抽出には、上記の手法の他にも種々の手法がある。例えば、最大エントロピーモデルと書き換え規則を用いて固有表現を抽出する手法がある(非特許文献１５参照)。

【０１５１】

【非特許文献１５】内元清貴,馬青,村田真樹,小作浩美,内山将夫,井佐原均,最大エ ントロピーモデルと書き換え規則に基づく固有表現抽出,言語処理学会誌, Vol.7, No.2, 2000年

【０１５２】

また、例えば、以下の非特許文献１６に、サポートベクトルマシンを用いて日本語固有表現抽出を行う手法について記載されている。

【０１５３】

【非特許文献１６】山田寛康,工藤拓,松本裕治,SupportVector Machineを用いた日本語固

有表現抽出,情報処理学会論文誌, Vol.43, No.1", 2002年
【０１５４】
(2)作成したルールを用いる手法
人手でルールを作って固有表現を取り出すという方法もある。
例えば、
名詞+「さん」だと人名とする
名詞+「首相」だと人名とする
名詞+「株式会社」だと企業名とする
名詞+「町」だと地名とする
名詞+「市」だと地名とする
などである。
【０１５５】
以上の方法によって固有表現を抽出し、抽出された表現のうち、例えば人名や企業名などを解-素性対抽出部（1042）において抽出することができる。
【０１５６】
このように固有表現だけを関連データとして抽出することで、前後関係を検出する必要のない関連データを抽出することを防止でき、特に対象データにとって重要な関連データについて本発明の順序の検出を行うことができる。
【０１５７】
（実施例３）
本発明は、関連データからその前後を検出する上記の方法に限らず、関連データと共起する共起データから前後を検出する方法を提供することもできる。以下、請求項５ないし８等に記載の本発明の実施例について説述する。
図９は本実施例に係るデータ処理装置（１’）の構成図である。上記実施例１と同一の構成部については同一符号を付し、説明を省略する。
【０１５８】
ＣＰＵ（１０）には入力部（１０６）、関連データ抽出部（１０７）、共起データ抽出部（１０８）、共起データ間関係検出部（１０９）、出力部（１０５）が設けられている。本構成により図１０に示す処理を実行する。
【０１５９】
まず、入力部（１０６）が外部記憶装置（３）あるいはインターネット、ＬＡＮ(Local Area Network)等のサーバ上からネットワークアダプタ（４０）を介して第１コンテンツファイル（２０）及び第２コンテンツファイル（２１）を取得し、ＣＰＵ（１０）内に取り込む処理を行う。
【０１６０】
請求項５に記載の関連データ抽出手段である関連データ抽出部（１０７）では、予め定義してある対象データ（２２）と関連する関連データを異なるファイルである各コンテンツファイル（２０）（２１）から抽出する。（関連データ抽出ステップ：Ｓ１０）　本処理は実施例１と同様である。コンテンツファイルは同一のファイルでもよい。
【０１６１】
そして共起データ抽出手段である共起データ抽出部（１０８）において、第１コンテンツファイル（２０）やコンテンツデータ（２３）から、第１関連データ（２４）と共起している単数又は複数の文字列である第１共起データ（３０）、第２コンテンツファイル（２１）やコンテンツデータ（２３）から、第２関連データ（２５）と共起している単数又は複数の文字列である第２共起データ（３１）を抽出する。（共起データ抽出ステップ：Ｓ２０）
【０１６２】
なお、別実施例として請求項６等に記載のように、図示しない共起データ共起ファイル抽出ステップを、上記共起データ抽出ステップ（Ｓ２０）の直後に設けて、第１共起データと第２共起データとが共起する共起データ共起ファイルを抽出してもよい。

抽出された共起データや共起データ共起ファイルから共起データ間関係検出手段である共起データ間関係検出部（１０９）が共起データ間関係規則（１３１）を参照して、第１共起データ（３０）と第２共起データ（３１）の順序を検出する。（共起データ間関係検出ステップ：Ｓ２１）

【０１６３】

本実施例では、関連データ間関係検出ステップ（Ｓ１２）に代わって共起データ間関係検出を行っているが、関連データについて行う場合と全く同様に共起データについて処理すればよい。共起データ間関係規則（１３１）についても関連データ間関係規則（１３０）と異なるところはない。

【０１６４】

共起データ間関係検出ステップ（Ｓ２１）においても、請求項７等に記載のようにルールベースで作成された共起データ間関係規則（１３１）を用いてもよいし、請求項８等に記載のように機械学習により作成された共起データ間関係規則（１３１）を用いてもよい。

ルールベースの作成方法、機械学習方法についても上記実施例と同様である。

【０１６５】

さらに、出力部（１１０）では、前後の決定された共起データに合わせて、関連データの前後を出力する。（出力ステップ：Ｓ２２）

すなわち、第１共起データが後、第２共起データが前と検出された場合には、第１関連データを後、第２関連データを前として出力する。

【０１６６】

上記の処理について具体例を用いて説明すると、図１１（Ａ）に示すような第１コンテンツファイル（２０ｅ）には対象データ（２２ｅ）と第１関連データ（２４ｅ）が含まれる。また第２コンテンツファイル（２１ｅ）には同じ対象データ（２２ｅ）と第２関連データ（２５ｅ）が含まれる。これらを関連データ抽出ステップ（Ｓ１０）において抽出する。

【０１６７】

次にコンテンツデータ（２３）中の２つのファイル（２３１ｅ）（２３１ｆ）から、それぞれ第１関連データ（２４ｅ）と共起する第１共起データ（３０ｅ）、第２関連データ（２５ｅ）と共起する第２共起データ（３１ｅ）を共起データ抽出ステップ（Ｓ２０）で抽出する。

【０１６８】

図示するように、２つの関連データにはそれぞれ異なるビル名、●●●●●ビルと××××××ビルが記載されているだけであり、ルールベースによる実施例１や機械学習を用いた実施例２でも両者の前後関係が判定できないことがある。そのとき、本発明による共起データを用いる方法を適用する。

【０１６９】

すなわち、●●●●●ビルとは２００５年１０月１日という日付が、××××××ビルとは２００８年４月１日という日付がそれぞれ共起しており、それらの共起データ（３０ｅ）（３１ｅ）を比較することで、●●●●●ビルと××××××ビルとの前後を判定しようとするものである。

ここで挙げた例は単純な例であり、共起データ間関係規則に日付があったときにはその前後で共起データ間の関係を決定すると定めておけば共起データ間関係検出ステップ（Ｓ２１）において、２００８年４月１日である第２共起データ（３１ｅ）が後と検出される。

【０１７０】

その結果、出力ステップ（Ｓ２２）では、共起データが後と判定された第２関連データ（２５ｅ）が後、第１関連データ（２４ｅ）が前と出力される。

なお、共起データを抽出するコンテンツデータ（２３）は、例示した「ビル完成情報」のように定型的にビルの完成した情報が記載されたデータを用いれば、極めて高い精度で

共起データから関連データの前後を検出することができる。しかし、本発明は多数のコンテンツデータ（２３）から多数のルールベースで、あるいは機械学習により検出することができるので、これほど定型的なものでなく、ただ共起する日付が古いものが多い、新しいものが多い、というようにあいまいな複数のデータからでも検出することができる。

【０１７１】

また、図１３ないし図１５には別の実施例を挙げる。まず、図１３に示すように１つのコンテンツファイル（２０ｇ）から対象データ△△（２２ｇ）に関連する第１関連データ（２４ｇ）と第２関連データ（２５ｇ）を抽出する。（Ｓ１０）　このように関連データは同一のコンテンツファイル（２０ｇ）から抽出してもよい。

【０１７２】

図３（Ａ）で示した例と異なり、新住所と旧住所が共起していても、「新」「旧」を表すために文字でなくグラフィックを用いているような場合、前述した方法で両者の前後関係（新旧関係）を検出することはできない。

【０１７３】

そこで第１関連データ（２４ｇ）、第２関連データ（２５ｇ）と共起する第１共起データ（３０ｇ）、第２共起データ（３１ｇ）を抽出する。図１４に示すように、それぞれを１つのコンテンツファイル（２０）の一部（２０ｈ）（２０ｉ）において、各関連データから所定の文字数内（例えば前後５０文字以内）に出現する文字列を共起データとすることができる。本実施例では、旧住所「東京都中央区駅前１-２-３」（２４ｇ）と共起する「ＸＹＺビル」（３０ｇ）が第１共起データであり、新住所「東京都中央区駅間９-８-７」（２５ｇ）と共起する「ＡＢＣビル」（３１ｇ）が第２共起データである。

【０１７４】

さらに、これらの２つの共起データ（３０ｇ）（３１ｇ）が共起する共起データ共起ファイル（２３１ｇ）を抽出する。該ファイル（２３１ｇ）において、文字列「ＸＹＺビル」と文字列「ＡＢＣビル」の間には「から」が含まれており、上述したようにルールベースを用いても「ＡＢＣビル」が新しい住所と判定できる。

【０１７５】

このように共起データを用いるのは、関連データが必ずしも順序を検出するのに最適でない場合があるからである。例えば、前後を検出したい関連データが住所である場合に、住所は住居表示の変更などがない限り、住所だけを見て前後を検出することは難しい。関連データ間関係規則（１３０）を機械学習結果とする場合でも、位置などから学習して精度良く検出できる場合はあるが、位置も出現位置の前後程度しか特色がなく、あとは住所データのみが関連データの場合に、十分な確度で選択できないことがある。

【０１７６】

上記で示した例の他にも、電話が住所と共起していて、それらを共起データ（３０）（３１）とした場合に、例えば電話の市外局番の表記方法が変更になった事実から関連データの前後を検出できることが考えられる。すなわち、第１共起データ（３０）が、「(0424)12-1234」という電話番号で、第２共起データ（３１）が「「(042)321-1234」という電話番号であったとき、共起データ間関係規則（１３１）には、市外局番が３桁化された方が新しい（後）という学習結果が格納されていれば、第１関連データ（２４）と第２関連データ（２５）を較べても前後が明らかでない場合にもいずれが前後か検出することができる。

【０１７７】

さらに、各関連データと共起するデータにさらに共起するデータを用いてもよい。すなわち、コンテンツファイルが会社名・郵便番号・住所であり、対象データ（会社名）の関連データ（住所）に対応する１次の共起データが郵便番号であったとして、さらにその郵便番号と別のコンテンツファイルで共起する２次の共起データが郵便局名であるときに、その郵便局名であれば前後が容易に検出できることがあり得る。そのような場合に、２次、３次の共起データを用いることもできる。

【０１７８】

（他言語への適用）
本発明は、日本語以外の言語であっても対象とることができる。例えば、コンテンツファイルやコンテンツデータが英語のテキストファイルであって、対象データ・関連データ・共起データ等がすべて英語の単語又は節（２以上の単語からなる集合）でもよい。
英語のように分かち書きをする言語では単語の抽出は簡単であるが、形態素解析を行って品詞情報を得ることで固有名詞などを的確に抽出することができる。英語の形態素解析を行う手法として、非特許文献１７に開示される手法がある。
【０１７９】
【非特許文献１７】Eric Brill, Transformation-BasedError-Driven Learning and Natural Language Processing: A Case Study inPart-of-Speech Tagging, Computational Linguistics, Vol. 21, No. 4, p.543-565,1995.
【０１８０】
また、関連データ抽出ステップ（Ｓ１０）、関連データ共起ファイル抽出ステップ（Ｓ１１）、関連データ間関係検出ステップ（Ｓ１２）、共起データ（Ｓ２０）、共起データ間関係検出ステップ（Ｓ２１）、共起データ共起ファイル抽出ステップなど、本発明の各処理において、日本語と英語を翻訳してから実行処理することができる。
【０１８１】
例えば、対象データ"Triangle Corporation"に対して関連データ"1-2-3 Ekimae, Chuo-ku,Tokyo"が含まれるとき、それぞれを翻訳して「△△株式会社」「東京都中央区駅前１-２-３」としてから用いることができる。
このような翻訳には訳語辞書、すなわちcar-車のように訳語が対になって表記される辞書を用いて単に置き換えることもできるし、公知の翻訳ソフトウェアなどによって単語・テキスト翻訳してもよい。
【０１８２】
最後に、住所変更情報を取得するアルゴリズムの一例と、その実験例を示す。
まず、所定の検索エンジンを用いて、会社名「セレスター通信株式会社」と「本社」をＡＮＤ検索する。その検索結果の一部を次に示す。
【０１８３】

【表３】

RES NO: 18 URL:http://www.jba.or.jp/publish/publication/pdf/004_h14bio.pdf TITLE: 平成14年度 バイオ産業基盤形成事業 報告書 SUMMARY: ファイルタイプ: PDF/Adobe Acrobat
RES_NO: 19 URL: http://www2d.biglobe.ne.jp/~gama/cgi/list.cgi?ap/co.csv TITLE: List of ap/co.csv SUMMARY: shionogi.co.jp,塩野義製薬 株式会社 igs.co.jp, 株式会社 アイジーエス usystem.co.jp, 株式 会社 アシステム horiuchi-color.co.jp, 株式会社 ... キソメック 株式会社 d-kids.co.jp, 株式会社 ディ・キッズ cl-sciences.co.jp, セレスター通信株式 ...
RES NO: 20 URL: http://kaznak.web.infoseek.co.jp/tasangyou/bio/mitubishi.htm TITLE: mitubishi SUMMARY: 三菱化学とその100%出資のゲノム創薬資源の開発・ライセンス事業会社であるゾイジーン、富士通およびセレスター通信は、今般、バイオ分野における総合的な協業関係を構築することといたしました、 ...
RES NO: 21 URL: http://search.atengineer.com/kj/g061211/60386568.htm TITLE: 長いDNA鎖を正確かつ効率よく複製する技術の開発に成功 ... SUMMARY: ... 上居バイオアシンメトリプロジェクト（総括責任者 上居洋文、研究期間 平成7年～12年）の研究成果を基に、平成16年1月から平成18年3月にかけて セレスター通信株式会社 （代表取締役 松岡雅裕、 本社 茨城県つくば市… ...
RES_NO: 22 URL: http://www.melma.com/backnumber_21482_508569/ TITLE: 株式情報	melma! SUMMARY: バイオ関連のベンチャー企業、 セレスター通信 と組み、2年間に40億円の研究費をかけた。 ... 先行き不透明感が強まるなか、ファナックは強気の生産計画を打ち出した。山梨県忍野村の 本社 工場で生産能力を増強。 ...

以降省略

【０１８４】

これらの検索結果から、社名と「本社」という単語の間に、会社という単語がない場合に、「本社」以降の表現を住所の部分表現として抽出（ただし記号などは除く）する。また、日付表現を、次の正規表現で抽出する。

【０１８５】

【表４】

/((平成|昭和|大正|明治)?[0-9０-９元一二三四五六七八九〇十]+年([0-9０-９元一二三四五六七八九〇十]+月)?([0-9０-９元一二三四五六七八九〇十]+日)?)/

【０１８６】

そして、社名と「本社」という単語の間のバイト数を計算する。バイト数、住所の部分表現、日付表現、元の検索エンジンの出力データをスペースで区切って出力すると次のようになる。

【０１８７】

【表５】

10 千葉市美浜区 http://pr.fujitsu.com/jp/news/2003/01/21-1.html@ セレスター通信 と富士通信が蛋白質間の共通アミノ酸 ... @セレスター通信と富士通信が蛋白質間の共通アミノ酸配列を実用時間で検索する技術を開発.セレスター通信株式会社(以下CLS、本社:千葉市美浜区、社長:土居洋文)と富士通信株式会社は共同で、複数の蛋白質間で...@
22 茨城県つくば市 平成15年12月|平成18年12月|平成18年3月 http://www.jst.go.jp/itaku/result/success.html@成功課題(委託開発の実績) 独創的シーズ展開事業 委託開発@本開発課題は、早稲田大学名誉教授一ノ瀬昇らの研究成果を基に、平成15年12月から平成18年12月にかけて、株式会社...年1月から平成18年3月にかけてセレスター通信株式会社(代表取締役松岡雅裕、本社茨城県つくば市、資本金...@
22 茨城県つくば市 平成7年|12年|平成16年1月|平成18年3月 http://search.atengineer.com/kj/g061211/G0386568.htm@長いDNA鎖も正確かつ効率よく複製する技術の開発に成功 ... @...土居バイオアシンメトリフロジェクト(総括責任者土居洋文、研究期間平成7年～12年)の研究成果を基に、平成16年1月から平成18年3月にかけてセレスター通信株式会社(代表取締役松岡雅裕、本社茨城県つくば市…...@

【０１８８】

表５の出力から下記の入力１、入力２のデータを作成する。すなわち、住所の部分表現の種類の分だけ、「社名」と「住所の部分表現」のAND検索用のデータを作成する。

【０１８９】

【表６】

入力１	セレスター通信株式会社 千葉市美浜区
入力２	セレスター通信株式会社 茨城県つくば市

【０１９０】

入力１、２を検索エンジンでＡＮＤ検索する。このときの入力１、入力２の検索エンジンの出力はそれぞれ次の通りであった。

【０１９１】

【表７】

RES_NO: 1 URL: http://pr.fujitsushin.com/jp/news/2003/01/24-1.html TITLE: セレスター通信 と富士通信が蛋白質間の共通アミノ酸 ... SUMMARY: セレスター通信 と富士通信が蛋白質間の共通アミノ酸配列を実用時間で検索する技術を開発. セレスター通信株式会社 (以下CLS、本社： 千葉市美浜区 、社長：山田太郎)と富士通信 株式会社 は共同で、複数の蛋白質間で ...
RES NO: 2 URL: http://www.csci.co.jp/jp/release/2003/news03001_a.html TITLE: セレスター通信株式会社 SUMMARY: セレスター通信株式会社 (以下CLS、本社： 千葉市美浜区 、社長：山田太郎) と富士通信株式会社 は共同で、複数の蛋白質間で見出される、共通のアミノ酸パターンを検索するためのモチーフ検索ツールHMMER(*1) のチューニングを行ない、 ...
RES NO: 3 URL: http://www.venturewatch.jp/event/company_data/015.html TITLE: NEDO技術開発機構 成果展示会2005 SUMMARY: 【ブース番号】 ラ実 05, 【問合せ先】 セレスター通信株式会社 業務推進室取締役坂本 光宏, 〒261 8501 千葉市美浜区 中瀬1 3 幕張テクノガーデンD棟17階 TEL：043 274 1234 FAX：043 274 1234 e mail:sakamoto@clci.co.jp ...
RES NO: 4 URL: http://www.business i.jp/bio/jyushou/past.html TITLE: 日本バイオベンチャー大賞 > 受賞者紹介> 過去の受賞者 SUMMARY: 株式会社 総合医科学研究所(大阪市豊中市新千里東町1 4 2), 経済産業大臣賞, セレスター ・通信株式会社 (千葉市美浜区 中瀬1-3 幕張テクノガーデン), 文部科学大臣賞, 株式会社 ジェネディックス(札幌市中央区北10条西15丁目28) ...
RES NO: 5 URL: http://www.daiichisankyo.co.jp/11ess/cgi-bin/csvlview_obj.php?b_newsrelease_n1/196/051205-001j-v1.pdf TITLE: Research Memo & Minutes SUMMARY: ファイルタイプ: PDF/Adobe Acrobat

以下省略

【０１９２】

【表８】

RES NO: 1 URL: http://www.jst.go.jp/pr/info/info365/index.html TITLE: 長いDNA鎖も正確かつ効率よく複製する技術の開発に成功（遺伝子診断 ... SUMMARY: セレスター通信株式会社 代表取締役 松岡雅裕（マツオカ マサヒロ） 〒300 2635 茨城県 つくば市東光台 5-9-9 筑波研究コンソーシアム TEL：0298-47-1234 FAX：0298-47-1234. 独立行政法人 科学技術振興機構 産業連携事業本部 開発部 ...
RES_NO: 2 URL: http://www-06.ibm.com/jp/government/hpc/casestudies/ TITLE: IBM ハイパフォーマンス・コンピューティング	導入事例 Japan SUMMARY: 独立行政法人 産業技術総合研究所独立行政法人 産業技術総合研究所（産総研：茨城県 つくば市）のグリッド研究センターと計算科学 ... セレスター通信 株式会社セレスター通信株式会社 は2000年8月に設立された創薬 ...
RES_NO: 3 URL: http://www.nedo.go.jp/kengyou/gyoumuka/tenjikai/h17/tokyo2005/pdf/pamphlet_pdf/jitsuyou/life/lj_05.pdf TITLE: タンパク解析用超電導NMRプローブの開発 個の医療と創薬のための生命 ... SUMMARY: ファイルタイプ： PDF/Adobe Acrobat - HTML バージョン
RES NO: 4 URL: http://phonebook.yahoo.co.jp/a108/g116/g20093/g32163000/?b=3&h=s TITLE: Yahoo!電話帳 - 茨城県 - 情報、通信 - 研究機関 - 自然科学研究所 SUMMARY: 株式会社 生体分子計測研究所 (FAX), 029 839 4612, 茨城県つくば市 榎戸 807-133, 周辺地図を表示 パソコンに送信 ケータイに送信, セレスター通信 株式会社 つくばR&Dセンター (代), 029 847 1781, 茨城県 つくば市東光台５丁目９９ ...
RES NO: 5 URL: http://local.yahoo.co.jp/a108/g107/g20781/?k1 14 TITLE: Yahoo!地域情報 - 茨城県 - 企業 - 研究機関 SUMMARY: 株式会社 生体分子計測研究所, 茨城県つくば市 榎戸, 地図を表示する, 自然科学研究所 株式会社 生体分子計測研究所 (FAX), 茨城県つくば市 榎戸, 地図を表示する, 自然科学研究所, セレスター通信株式会社 つくばR&Dセンター (代) ...

以下省略

【０１９３】

検索エンジンの結果から、住所の完全情報を取得する。そのために、社名と、住所の部分表現の間に、会社や研究所という単語がない場合に、住所の部分表現以降の表現を、句点読点、省略表現、括弧表現、空白表現を含まないまでのものを、住所の部分表現も含めて、住所表現として抽出する。また、日付表現を、下記の正規表現で抽出する。

【０１９４】

【表９】

/((平成|昭和|大正|明治)?[0-9０-９元・]+年([0-9０-９元・]+月)?([0-9０-９元・]+日)?/

【０１９５】

社名と住所の部分表現の間のバイト数を計算し、バイト数、住所表現、日付表現、元の検索エンジンの出力データをスペースで区切って出力する。

【０１９６】

【表１０】

入力１の場合

2 千葉市美浜区中瀬 1-3 幕張テクノガーデン http://www.business-i.jp/bio/jyushou/past.html@日本バイオベンチャー大賞 > 受賞者紹介 > 過去の受賞者@株式会社総合医科学研究所（大阪市豊中市新千里東町 1-4-2）,経済産業大臣賞,セレスター通信株式会社（千葉市美浜区中瀬 1-3 幕張テクノガーデン）,文部科学大臣賞,株式会社ジェネティックラボ（札幌市中央区北９条西 15 丁目 28 196）...@
35 千葉市美浜区中瀬 1-3 幕張テクノガーデン D 棟 17 階 TEL：043-274-5801FAX：043 274 5817e mail:sakamoto@cl sciences.co.jp... http://www.venturewatch.jp/event/company data/015.html@NEDO技術開発機構 - 成果展示会 2005@【ブース番号】ラ実-05,【問合せ先】セレスター通信株式会社業務推進室取締役坂本光宏,〒261-8501 千葉市美浜区中瀬 1-3 幕張テクノガーデン D 棟 17 階 TEL：043-274-5801FAX：043-274-5817e-mail:sakamoto@cl-sciences.co.jp...@
44 千葉市美浜区中瀬 1-3 幕張テクノガーデン D 棟 17 階電話番号 043-274-5801FAX 番号 043-274-5817 電子メイル info@cl-sciences.co.jp ホームページ：... http://bioencounter.bioquarry.com/modules/tinyd0/index.php?id=121@ セレスター通信株式会社 - BioEncounter@セレスター通信株式会社 CelestarLexico-Sciences,Inc.〒261-8501 千葉県千葉市美浜区中瀬１３幕張テクノガーデン D 棟 17 階電話番号 043 274 5801FAX 番号 043-274-5817 電子メイル info@cl-sciences.co.jp ホームページ：...@
80 千葉市美浜区中瀬１３幕張テクノガーデン D17,電話,043 274 5801,FAX,043 274 5817... http://www.kazusabio.net/data/cl-sciences.html@ セレスター通信株式会社 @企業名・団体名,セレスター通信株式会社,ふりがな,セレスター通信,代表者役職,代表取締役社長...所在地,千葉市美浜区中瀬 1-3 幕張テクノガーデン D17,電話,043-274-5801,FAX,043-274-5817...@

【０１９７】

【表１１】

入力２の場合

36 茨城県つくば市東光台５丁目９９... http://phonebook.yahoo.co.jp/a108/g116/g20093/g32163000/?b=3&h=s@Yahoo!電話帳　茨城県　情報、通信　研究機関　自然科学研究所@株式会社生体分子計測研究所(FAX),029 839 4612,茨城県つくば市榎戸 807 133,周辺地図を表示パソコンに送信ケータイに送信,セレスター通信株式会社つくばR&Dセンター(代),029-847-1781,茨城県つくば市東光台５丁目９９...@
48 茨城県つくば市東光台５９９筑波研究コンソーシアム TEL：0298 47 1781FAX：0298 47 1782,独立行政法人科学技術振興機構産業連携事業本部開発部... http://www.jst.go.jp/pr/info/info365/index.html@長いDNA鎖を正確かつ効率よく複製する技術の開発に成功（遺伝子診断 ... @セレスター通信株式会社代表取締役松岡雅裕（マツオカマサヒロ）〒300-2635 茨城県つくば市東光台 5-9-9 筑波研究コンソーシアム TEL：0298 47 1781FAX：0298 47 1782,独立行政法人科学技術振興機構産業連携事業本部開発部...@

【０１９８】

表５の出力における住所を表１０、１１の表現で補完してさらに、日付の新しい順に出力する。

【０１９９】

【表１２】

千葉市美浜区中瀬 1-3 幕張テクノガーデン　平成 18 年 3 月
茨城県つくば市東光台５丁目 9-9　2003

【０２００】

次に、社名変更情報の取得する方法についても実験を行った。

1.「社名」、「変更」という単語で検索エンジンでAND検索する。その結果が次の通りであった。

【０２０１】

【表１３】

RES NO: 1 URL: http://www.traders.co.jp/stocks_data/data/name_change/name_change.asp TITLE: 社名変更　株式情報満載のサイト　トレーダーズ・ウェブ SUMMARY: 株式移転に伴う 変更 は「株式移転」のページへ. 2007年. 変更 日, コード, 市場, 新 社名 , 旧 社名 . 10/01, 9861, 東1, 吉野家ホールディングス, 吉野家ディー・アンド・シー. 10/01, 5726, 東1, 大阪チタニウムテクノロジー, 住友チタニウム ...
RES_NO: 2 URL: http://www.aisatsujo.com/houjin/syamei.html TITLE: 社名変更 の挨拶状印刷／挨拶状ドットコム SUMMARY: 社名変更 挨拶状の印刷なら挨拶状ドットコム。文例豊富で手軽に簡単に注文できるあいさつ状印刷専門サイトです。(社名変更 ・会社設立・事務所移転・社長交代・転勤・転職・退職など)
RES_NO: 3 URL:http://www.softbanktelecom.co.jp/release/2006/aug/0828/index.html TITLE: 社名変更 のお知らせ ニュースリリース：ソフトバンクテレコム SUMMARY: この 社名変更 を機に、これまで追求してきたお客様との密接な関係、信頼性および安全性に加え、ソフトバンクグループ各企業との連携を深めることで、当社の強みを更に強化していきたいと考えております。 ...
RES NO: 4 URL: http://d.hatena.ne.jp/keyword/社名変更 TITLE: 社名変更 とは はてなダイアリー SUMMARY: ビジネスに役立つ無料サイト： 定型文書はコピペで効率よ... internet.watch.impress.co.jp 31 users · アップル、「アップルジャパン」へ 社名変更 (MYCOM ジャーナル) journal.mycom.co.jp 18 users · カネボウが「クラシエ」に 社名変更 社内などから募り ...
RES_NO: 5 URL: http://www.itmedia.co.jp/news/0211/18/njbi_11.html TITLE: News：アスキーが 社名変更 SUMMARY: アスキーは11月18日、 社名 を「メディアリーヴス」に 変更 した。また同日、インターネット総合研究所子会社のアイ・アール・アイコマースアンドテクノロジー (IRI C&T) は自動車ニュースサイト「オートアスキー」の営業権を取得したと発表した。 ...

以降省略

【０２０２】

2.社名と変更を含む文から

[A]は.....「[B]」..... のパターンに適合する

[A], [B]を取り出す.

[A] は元の社名で, [B] が新しい社名とする.

以下、[A]、 [B]、取り出した元の文をスペースで区切って出力する。

【０２０３】

【表１４】

アスキー・メディアワークス　アスキーは11月18日、社名を「メディアワークス」に変更した
DDIポケット WILLCOM 2005年2月 DDIポケットは10月14日、2005年2月から社名を「WILLCOM　（ウィルコム）に変更すると発表した
新社名 株式会社ライブドアホールディングス 2007年4月2日 新社名は「株式会社ライブドアホールディングス」で、変更予定日は2007年4月2日
ボーダフォン ソフトバンクモバイル　ボーダフォンは、10月1日から社名を「ソフトバンクモバイル」に変更する
笠原健治氏 1999年の創業より、株式会社イー・マーキュリーとしてインターネットを利用した新しい価値創造に尽力してまいりましたが、2004年2月に開始したSNS、『mixi』の登録者数が250万人以上となり、... 1999年|2004年2月 今回の社名変更について、同社の代表取締役の笠原健治氏は、 1999年の創業より、株式会社イー・マーキュリーとしてインターネットを利用した新しい価値創造に尽力してまいりましたが、2004年2月に開始したSNS、『mixi』の登録者数が250万人以上となり、...

【０２０４】
上記のように、この方法でも多くの社名変更情報が抽出できることがわかった。さらに既存の社名の辞書を用意しておき，[A]が既存の社名辞書にあるものだけを抽出することでさらに性能高く社名の変更情報を取得できると考えられる。

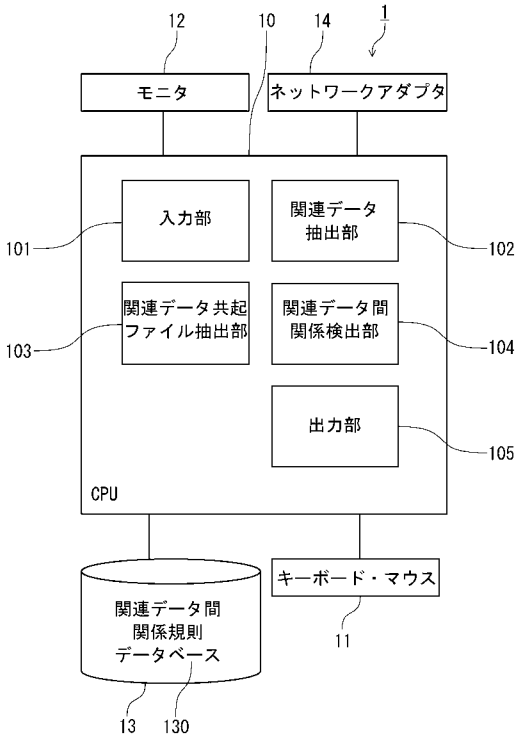
【図面の簡単な説明】

【０２０５】
【図１】本発明のデータ処理装置の全体構成図である。
【図２】本発明のデータ処理方法の処理フローチャートである。
【図３】関連データ共起ファイルの例
【図４】関連データ共起ファイルの例
【図５】関連データ共起ファイルの例
【図６】本発明の第３の実施例における関連データ間関係検出部の構成図である。
【図７】機械学習の処理を説明する説明図である。
【図８】機械学習（ＳＶＭ）の処理を説明する説明図である。
【図９】本発明の第３の実施例におけるデータ処理装置の全体構成図である。
【図１０】本発明の第３の実施例におけるデータ処理方法の処理フローチャートである。
【図１１】関連データ共起ファイルの例
【図１２】共起データ共起ファイルの例
【図１３】関連データ共起ファイルの例
【図１４】共起データ共起ファイルの例
【図１５】共起データ共起ファイルの例

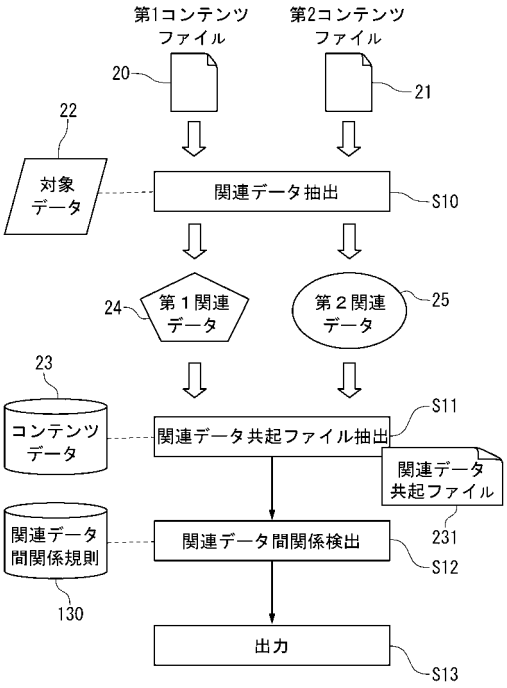
【符号の説明】

【０２０６】
１　データ処理装置
１０　ＣＰＵ
１１　キーボード・マウス
１２　モニタ
１３　ハードディスク
１４　ネットワークアダプタ
１０１　入力部
１０２　関連データ抽出部
１０３　関連データ共起ファイル抽出部
１０４　関連データ間関係検出部
１０５　出力部
１３０　関連データ間関係規則データベース

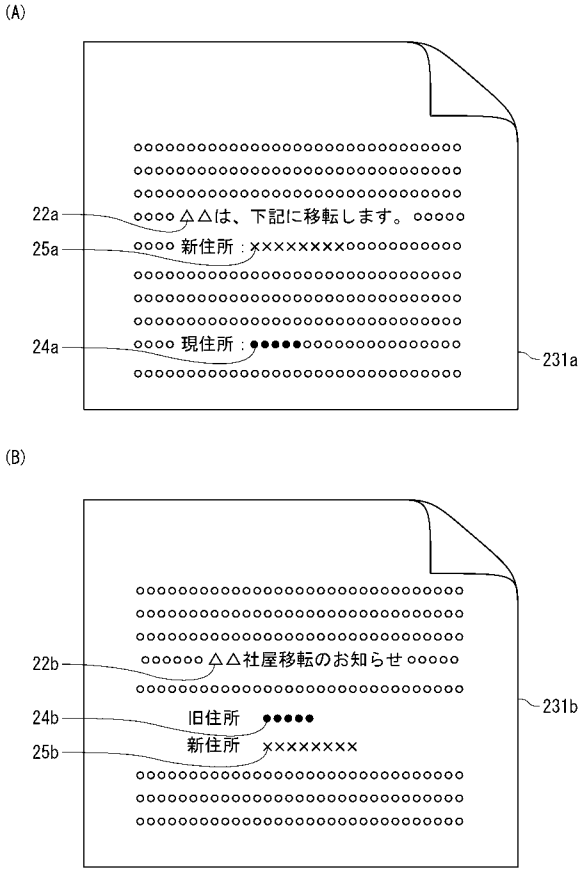
【図１】


12
10
14
1
モニタ
ネットワークアダプタ
入力部
関連データ抽出部
101
102
関連データ共起ファイル抽出部
関連データ間関係検出部
103
104
出力部
105
CPU
関連データ間関係規則データベース
キーボード・マウス
11
13
130

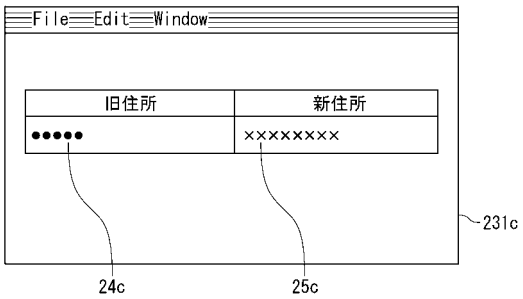
【図２】


第1コンテンツファイル
第2コンテンツファイル
20
21
22
対象データ
関連データ抽出
S10
24
第１関連データ
第２関連データ
25
23
コンテンツデータ
関連データ共起ファイル抽出
S11
関連データ共起ファイル
231
関連データ間関係規則
関連データ間関係検出
S12
130
出力
S13

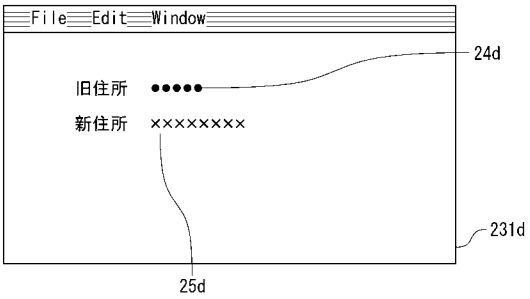
【図３】


(A)
22a
25a
24a
231a
(B)
22b
24b
25b
231b
旧住所 ●●●●●
新住所 ××××××××

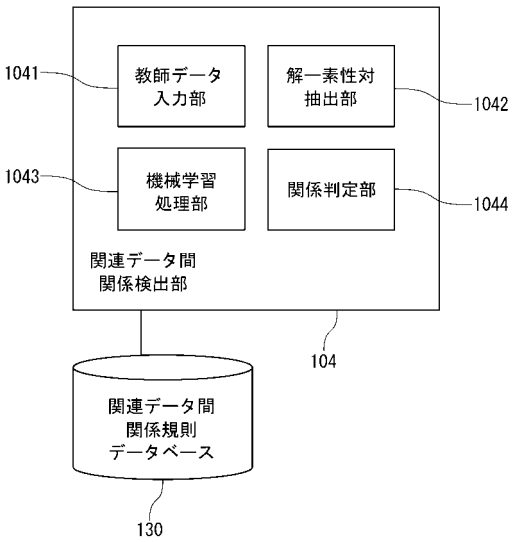
【図４】


File Edit Window
旧住所
新住所
●●●●●
××××××××
231c
24c
25c

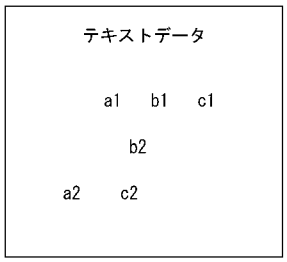
【図５】


File Edit Window
旧住所 ●●●●●
新住所 ××××××××
24d
231d
25d

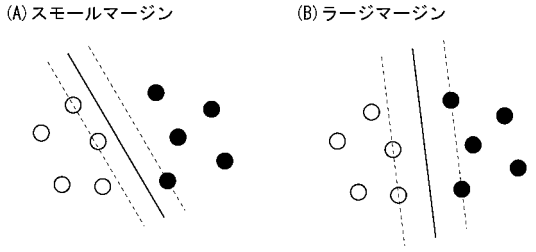
【図６】


1041
教師データ
入力部
解一素性対
抽出部
1042
1043
機械学習
処理部
関係判定部
1044
関連データ間
関係検出部
104
関連データ間
関係規則
データベース
130

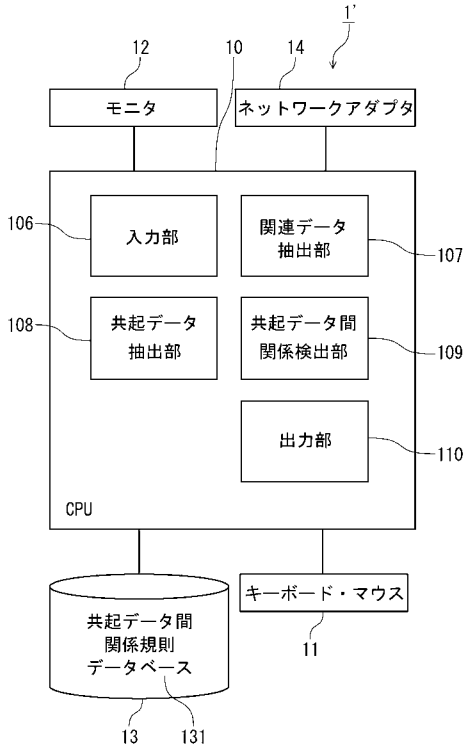
【図７】


テキストデータ
a1 b1 c1
b2
a2 c2

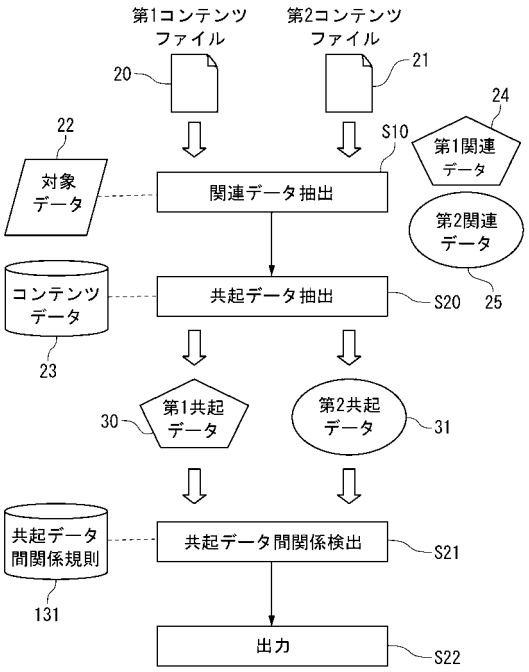
【図８】


(A)スモールマージン
(B)ラージマージン

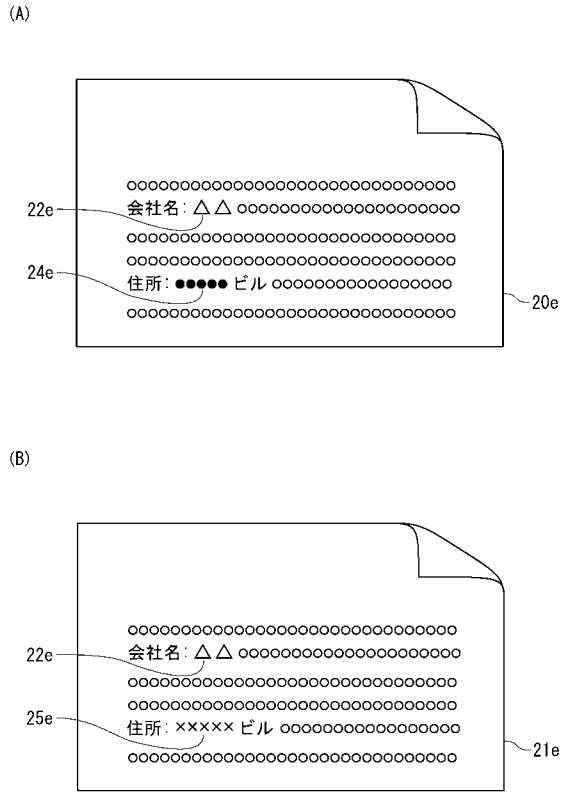
【図９】



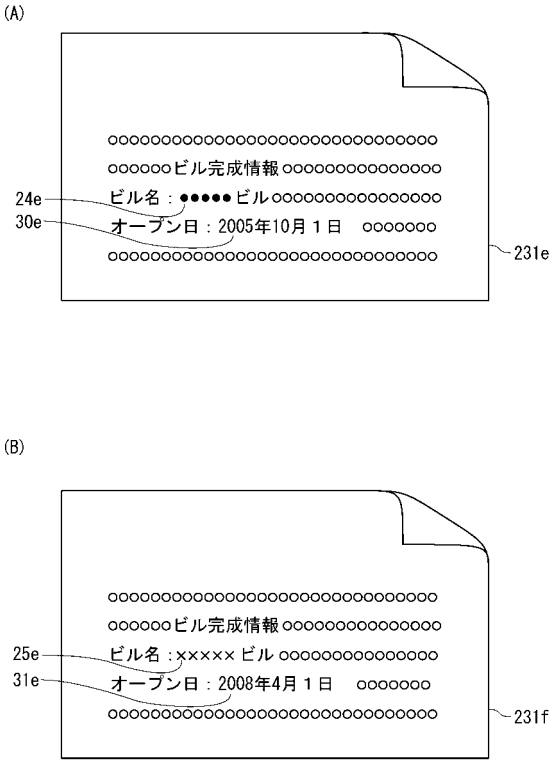
【図１０】



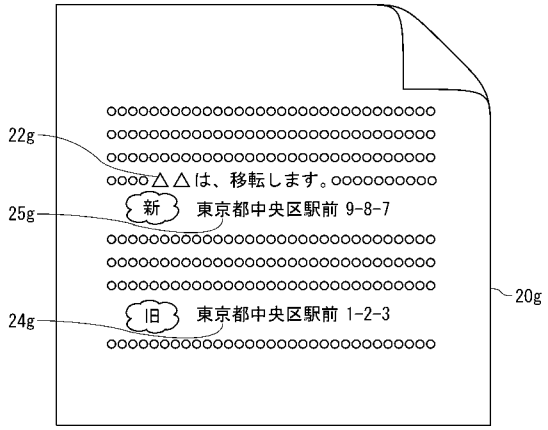
【図１１】



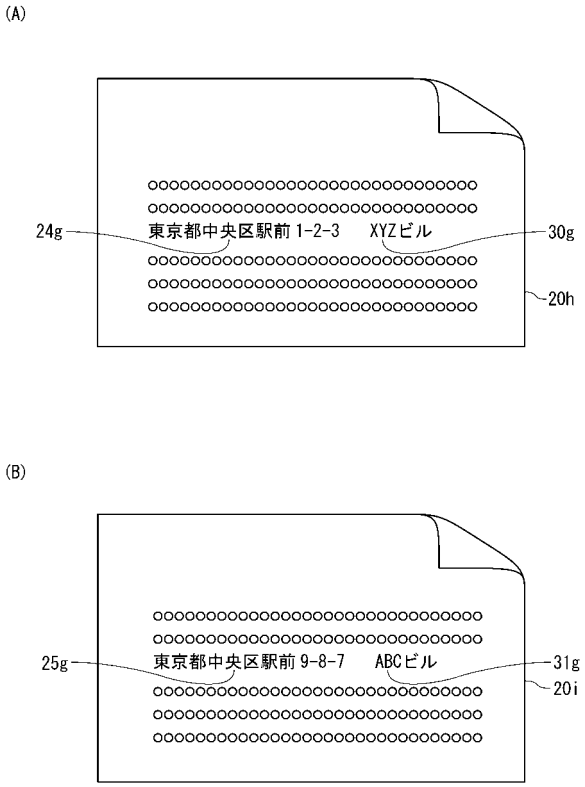
【図１２】



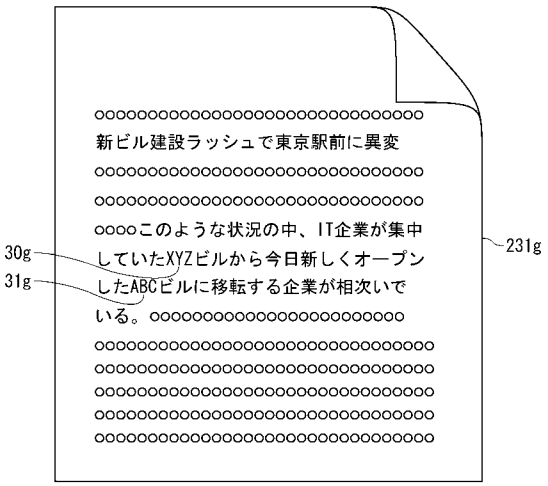
【図１３】


22g
○○○○△△は、移転します。○○○○○○○○○○
25g
新
東京都中央区駅前 9-8-7
24g
旧
東京都中央区駅前 1-2-3
20g

【図１４】


(A)
24g
東京都中央区駅前 1-2-3　XYZビル
30g
20h
(B)
25g
東京都中央区駅前 9-8-7　ABCビル
31g
20i

【図１５】


新ビル建設ラッシュで東京駅前に異変
○○○○このような状況の中、IT企業が集中
30g
していたXYZビルから今日新しくオープン
31g
したABCビルに移転する企業が相次いで
いる。○○○○○○○○○○○○○○○○○○○○
231g

フロントページの続き

(56)参考文献 特開平８－２４１３２８（ＪＰ，Ａ）
特開２００４－１０２６２８（ＪＰ，Ａ）
特開２００６－２３９６８（ＪＰ，Ａ）
賀家 智代，質問キーワードの順序依存性に基づくＷｅｂアーカイブ検索方式，日本データベース学会Ｌｅｔｔｅｒｓ，日本，日本データベース学会，２００６年 6月２２日，Ｖｏｌ．５ Ｎｏ．１，１２９－１３２ページ
小野田 透，Ｗｅｂアーカイブを用いた時系列パターンに基づく検索支援方式，電子情報通信学会 第１８回データ工学ワークショップ論文，日本，電子情報通信学会データ工学研究専門委員会，２００７年 6月 1日，１－９ページ

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