

(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2009-116456
(P2009-116456A)

(43) 公開日 平成21年5月28日(2009.5.28)

(51) Int.Cl. F I テーマコード(参考)
G06F 17/30 (2006.01) G06F 17/30 380E 5B075

審査請求 未請求 請求項の数 18 O L (全 38 頁)

(21) 出願番号	特願2007-286269 (P2007-286269)	(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19年11月2日(2007.11.2)	(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
		Fターム(参考)	5B075 KK02 ND20 NK32 PQ02 PQ46 PR04 UU09

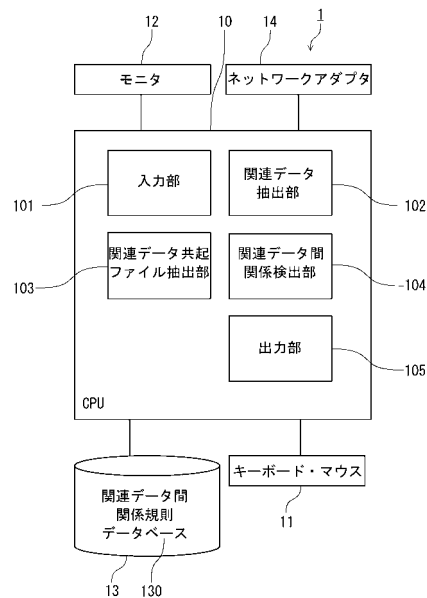
(54) 【発明の名称】 データ処理装置及びデータ処理方法

(57) 【要約】 (修正有)

【課題】 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。

【解決手段】 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1を提供する。該装置には、ネットワーク上やローカルの記憶手段に格納されたコンテンツファイルから第1関連データ・第2関連データを抽出する関連データ抽出手段102、各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103、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130を参照して、各関連データの前後を検出する関連データ間関係検出手段104、検出結果を出力する出力手段105を備える。

【選択図】 図1



【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。

10

【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。

20

【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。

30

【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。

40

【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

50

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。

10

20

30

40

50

【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に第 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 関連データと共に第 1 共起データ又は複数の第 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 関連データと共に第 2 共起データ又は複数の第 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に起している第 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】

10

20

30

40

50

【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。

10

【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。

20

【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。

30

【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。

40

【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】

50

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。

【 0 0 1 0 】

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。

10

【 0 0 1 1 】

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。

20

【 0 0 1 2 】

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。

【 0 0 1 3 】

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。

30

【 0 0 1 4 】

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。

40

【 0 0 1 5 】

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。

【 0 0 1 6 】

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

50

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。

10

【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

20

【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。

30

【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。

40

【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

50

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】

10

20

30

40

50

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、それらと共に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、各関連データと共に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。

10

【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。

20

【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。

30

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】

40

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

50

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を書く相手方の氏名、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

10

20

30

40

50

ット等のサーバ上に格納されていてもいづれでも良いが、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。

10

【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。

20

【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：xxxxxx」(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：xxxxxx」(25b)の順番に記載されている。

30

【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xxxxxx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

40

50

【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xxxxxx(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xxxxxx(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	存在文字列	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】

10

20

30

40

50

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)

$$\text{score}(D) = \sum_w \left(\text{tf}(w,D) * \log(N/\text{df}(w)) \right)$$

w Wで加算

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

10

20

30

40

50

、 で積を取る前の tf 項と idf 項の積が Okapi のウェイト法になって、この値を単語の重みに使う。

【 0 0 6 8 】

Okapi の式なら

(数 2)

$$\text{score}(D) = \left(\frac{tf(w,D)}{tf(w,D) + \text{length}/\text{delta}} * \log(N/df(w)) \right)$$

w W で加算

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。

10

【 0 0 6 9 】

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、 $tf \cdot idf$ 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。

【 0 0 7 0 】

【非特許文献 4】S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford Okapi at TREC-3, TREC-3, 1994 年

【非特許文献 5】Amit Singhal AT&T at TREC-6, TREC-6, 1997 年

20

【 0 0 7 1 】

これらの方法では、 $tf \cdot idf$ 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。

【 0 0 7 2 】

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。

【 0 0 7 3 】

【非特許文献 6】J. J. Rocchio, Relevance feedback in information retrieval, The SMART retrieval System, Edited by G. Salton, Prentice Hall, Inc., page 313-323, 1971 年

【 0 0 7 4 】

30

この方法は、 $\log(N/df(w))$ のかわりに、

(数 3)

$$\{E(t) + k_{af} * (\text{RatioC}(t) - \text{RatioD}(t))\} * \log(N/df(w))$$

を使う。

【 0 0 7 5 】

$E(t) = 1$ (対象データ)

$= 0$ (それ以外)

$\text{RatioC}(t)$ は関連データ群 B での t の出現率

$\text{RatioD}(t)$ 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 $\text{score}(D)$ 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。

40

【 0 0 7 6 】

$\text{score}(D)$ の 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。

【 0 0 7 7 】

また、他の方法として、 $\text{score}(D)$ の 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。

【 0 0 7 8 】

ここでは rocchio 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

50

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。

10

【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)。

20

【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-siryuu-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.

30

【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。

40

【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。

50

【 0 0 8 6 】

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。

【 0 0 8 7 】

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。

【 0 0 8 8 】

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。

10

【 0 0 8 9 】

【数4】

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

【 0 0 9 0 】

【数5】

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

20

【 0 0 9 1 】

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 ($F, 1 \leq j \leq k$)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($f_j|a$)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_j を持つ確率を意味する。P($f_j|a$)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、スムージングを行う。ここでは、以下の数6を用いてスムージングを行ったものを用いる。

【 0 0 9 2 】

【数6】

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

30

【 0 0 9 3 】

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

【 0 0 9 4 】

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を与えられたときに、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。

40

【 0 0 9 5 】

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 ($F, 1 \leq j \leq k$)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。

【 0 0 9 6 】

(数7)

$$p(a|b) = p(a|f_{\max})$$

50

ただし、 f_{\max} は以下の数 8 によって与えられる。

【 0 0 9 7 】

【 数 8 】

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

【 0 0 9 8 】

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。

10

【 0 0 9 9 】

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 ($1 \leq j \leq k$) 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。

【 0 1 0 0 】

【 数 9 】

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

20

$$\text{for } \forall f_j (1 \leq j \leq k)$$

【 数 1 0 】

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

30

【 0 1 0 1 】

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。

【 0 1 0 2 】

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。

40

【 0 1 0 3 】

【 非特許文献 1 0 】 Eric Sven Ristad, Maximum Entropy Modeling for Natural Language, (ACL/EACL Tutorial Program, Madrid, 1997)

【 0 1 0 4 】

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。

50

【 0 1 0 5 】

このとき、二つの分類が正例と負例からなるものとする、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。

【 0 1 0 6 】

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。

【 0 1 0 7 】

この拡張された方法は、以下の識別関数(数 1 1)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。

【 0 1 0 8 】

【 数 1 1 】

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

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

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

【 0 1 0 9 】

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 ($i=1, \dots, l, y_j \in \{1, -1\}$)は学習データの文脈と分類先を意味し、関数 sgn は、

$$\operatorname{sgn}(x) = \begin{cases} 1 & (x \geq 0) \\ -1 & (\text{otherwise}) \end{cases}$$

であり、また、各 α_i は数 1 3 と数 1 4 の制約のもと数 1 2 を最大にする場合のものである。

【 0 1 1 0 】

【 数 1 2 】

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

【 数 1 3 】

$$0 \leq \alpha_i \leq C \quad (i=1, \dots, l)$$

【 数 1 4 】

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

【 0 1 1 1 】

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を用いる。

10

20

30

40

50

【 0 1 1 2 】

(数 1 5)

$$K(x,y)=(x \cdot y+1)^d$$

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、 $x_i > 0$ 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。

【 0 1 1 3 】

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。

【 0 1 1 4 】

【非特許文献11】Nello Cristianini and John Shawe-Taylor, An Introduction to Support Vector Machines and other kernel-based learning methods, (Cambridge University Press, 2000)

【非特許文献12】Taku Kudoh, Tinsvm: Support Vector machines, (<http://chasen.org/~taku/software/TinySVM/>, 2002年)

【 0 1 1 5 】

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。

【 0 1 1 6 】

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。

【 0 1 1 7 】

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。

【 0 1 1 8 】

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。

【 0 1 1 9 】

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。

【 0 1 2 0 】

10

20

30

40

50

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。

10

【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。

20

【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。

30

【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。

40

【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

50

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

10

20

30

40

50

、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

10

【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。

20

【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、組み合わせで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。

30

【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)

40

【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。

50

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

10

上記において、B-???は、ハイフン以下の固有表現の種類が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が始まり以外を意味するタグであり、0 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。

20

【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。

30

【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・力行促音便 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。

40

【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のような素性が出現するかを調べ、どのような素性が出現しているときにどのような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。

50

【 0 1 5 0 】

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 1 5 参照)。

【 0 1 5 1 】

【非特許文献 1 5】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

【 0 1 5 2 】

また、例えば、以下の非特許文献 1 6 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。

【 0 1 5 3 】

【非特許文献 1 6】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

【 0 1 5 4 】

(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。

【 0 1 5 5 】

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。

【 0 1 5 6 】

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。

【 0 1 5 7 】

(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。

【 0 1 5 8 】

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。

【 0 1 5 9 】

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。

【 0 1 6 0 】

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

10

20

30

40

50

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】

10

20

30

40

50

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決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はなく、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「東京都中央区駅前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無い限り、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、「(042

10

20

30

40

50

4)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。

10

【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。

20

【0179】

【非特許文献17】Eric Brill, Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-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。

30

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。

40

【0183】

【表 3】

RES NO: 18	URL: http://www.jba.or.jp/publish/publication/pdf/004_b14bio.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/~nana/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 asystem, co, jp, 株式会社 アシシステム horifuchi-color, co, jp, 株式会社 ... キノメック 株式会社 d kids, co, jp, 株式会社 ディ・キッズ el sciences, co, jp, セレスター通信株式会社 ...
RES NO: 20	URL: http://kaznak.web.infoseek.co.jp/tasanyou/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、富士通信および セレスター通信 は、今後、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/G0386568.htm 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 ... SUMMARY: ... 上層バイオアシンメトリプロジェクト（総括責任者 上層洋文、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、フア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 工場生産能力を増強、 ...

10

20

以降省略

【 0 1 8 4 】

30

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。

【 0 1 8 5 】

【表 4】

```

/((平成|昭和|大正|明治)?([0-9]{0-9}元一二三四五六七八九〇十)年([0-9]{0-9}一三四五六七八九〇十)月)?([0-9]{0-9}元一二三四五六七八九〇十)日)?/

```

【 0 1 8 6 】

40

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。

【 0 1 8 7 】

【表 5】

10 千葉県美浜区 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：千葉県美浜区、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) 独自のシーズ展開事業 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プロジェクト(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...@

10

【 0 1 8 8 】

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みだけ、「社名」と「住所の部分表現」のAND検索用のデータを作成する。

20

【 0 1 8 9 】

【表 6】

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

【 0 1 9 0 】

入力 1、2 を検索エンジンで A N D 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。

30

【 0 1 9 1 】

【表 7】

<p>RES NO: 1</p> <p>URL: http://pr.fujitsu.com/jp/news/2003/01/24-1.html</p> <p>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 ...</p> <p>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 (以下共同) で、複数の蛋白質間で ...</p>	
<p>RES NO: 2</p> <p>URL: http://www.esci.co.jp/jp/release/2003/news03001_a.html</p> <p>TITLE: セレスター通信株式会社</p> <p>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 (以下共同) 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 (特許) のチューニングを行ない、 ...</p>	10
<p>RES NO: 3</p> <p>URL: http://www.venturewatch.jp/event/company_data/015.html</p> <p>TITLE: NEDO 技術開発機構 成果展示会 2005</p> <p>SUMMARY: 【ブース番号】 ライト 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@esci.co.jp ...</p>	
<p>RES NO: 4</p> <p>URL: http://www.business-i.jp/bio/jyushou/past.html</p> <p>TITLE: 日本バイオベンチャー大賞 > 受賞者紹介 > 過去の受賞者</p> <p>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 ジェネティクス (札幌市中央区北 10 条西 15 丁目 28) ...</p>	20
<p>RES NO: 5</p> <p>URL: http://www.daisankyo.co.jp/files/cgi-bin/cs/view_obj.php?b_newsrelease_n1/196/051205-001j-v1.pdf</p> <p>TITLE: Research Memo & Minutes</p> <p>SUMMARY: ファイルタイプ: PDF/Adobe Acrobat</p>	

以下省略

【 0 1 9 2 】

【表 8】

<p>RES_NO: 1</p> <p>URL: http://www.jst.go.jp/pr/info/info365/index.html</p> <p>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 ...)</p> <p>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 ...</p>	
<p>RES_NO: 2</p> <p>URL: http://www-06.ibm.com/jp/government/hpc/casestudies/</p> <p>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</p> <p>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月に設立された創業 ...</p>	10
<p>RES_NO: 3</p> <p>URL: http://www.ncdo.go.jp/kengyou/gyoumuka/tenjikai/h17/tokyo2005/pdf/pamphlet_pdf/jitsuyou/life/lj_05.pdf</p> <p>TITLE: タンパク解析用超電導 NMR プロブの開発 個の医療と創薬のための生命 ...</p> <p>SUMMARY: ファイルタイプ: PDF/Adobe Acrobat - HTMLバージョン</p>	
<p>RES_NO: 4</p> <p>URL: http://phonebook.yahoo.co.jp/a108/g116/g20093/g32163000/?b=3&h=s</p> <p>TITLE: Yahoo!電話帳 - 茨城県 - 情報、通信 - 研究機関 - 自然科学研究所</p> <p>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 ...</p>	20
<p>RES_NO: 5</p> <p>URL: http://local.yahoo.co.jp/a108/g107/g20781/?k1_14</p> <p>TITLE: Yahoo!地域情報 - 茨城県 - 企業 - 研究機関</p> <p>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) ...</p>	

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】

<pre> /((平成 昭和 大正 明治)?(0 90 - 9元 + 三四五六七八九〇十)年(0 90 - 9 + 三四五六七八九〇十日)?(10 90 - 9元 + 三四五六七八九〇十日)?)/ </pre>

【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】

10

20

30

40

【表 1 0】

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 1-1-2) . 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】 7A-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 電子メール 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 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 電子メール info@cl-sciences.co.jp ホームページ : ... @
80 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... 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... @

10

【 0 1 9 7】

【表 1 1】

入力 2 の場合

36 茨城県つくば市東光台 5 丁目 9-9... 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). 029-817-1781. 茨城県つくば市東光台 5 丁 目 9-9... @
48 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... http://www.jst.go.jp/pr/info/info365/index.html@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 ... @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... @

30

【 0 1 9 8】

表 5 の出力における住所を表 1 0、1 1 の表現で補完してさらに、日付の新しい順に出
力する。

【 0 1 9 9】

【表 1 2】

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

40

【 0 2 0 0】

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。

【 0 2 0 1】

【表 1 3】

<p>RES NO: 1</p> <p>URL: http://www.traders.co.jp/stocks_data/data/name_change/name_change.asp</p> <p>TITLE: 社名変更 株式情報満載のサイト トレーダーズ・ウェブ</p> <p>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</p> <p>社名、旧社名、10/01、9861、東1、吉野家ホールディングス、吉野家ディールアンドシー、10/01、</p> <p>5726、東1、大阪チタニウムテクノロジー、住友チタニウム ...</p>	
<p>RES_NO: 2</p> <p>URL: http://www.aisatsujo.com/houjin/syamei.html</p> <p>TITLE: 社名変更 の捺捺状印刷/捺捺状ドットコム</p> <p>SUMMARY: 社名変更 捺捺状の印刷なら捺捺状ドットコム。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)</p>	10
<p>RES_NO: 3</p> <p>URL: http://www.softbanktelecom.co.jp/release/2006/aug/0828/index.html</p> <p>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</p> <p>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。 ...</p>	
<p>RES NO: 4</p> <p>URL: http://d.hatena.ne.jp/keyword/社名変更</p> <p>TITLE: 社名変更 とは はてなダイアリー</p> <p>SUMMARY: ビジネスに役立つ無料サイト：定型文書はコピーで効率上...</p> <p>internet.watch.impress.co.jp 31 users &mdot; アップル、「アップルジャパン」へ 社名変更 (MYCOM ジャーナル) journal.my</p> <p>com.co.jp 18 users &mdot; カネボウが「クラシエ」に 社名変更 社内などから募り ...</p>	20
<p>RES_NO: 5</p> <p>URL: http://www.itmedia.co.jp/news/0211/18/njbt_11.html</p> <p>TITLE: News: アスキーが 社名変更</p> <p>SUMMARY: アスキーは11月18日、社名を「メディアリヴス」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。 ...</p>	

30

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 1 4】

<p>アスキー・メディアリーヴス 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。</p>
<p>DDI ホケット WILLCOM 2005年2月 DDI ホケットは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。</p>
<p>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</p>
<p>ポータフォン ソフトバンクモバイル ポータフォンは、10月1日から社名を「ソフトバンクモバイル」に変更する</p>
<p>等原健治氏 1999年の創業より、株式会社イー・マーカ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等原健治氏は、1999年の創業より、株式会社イー・マーカ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、...</p>

10

【 0 2 0 4】

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】

20

【 0 2 0 5】

- 【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 (S V M) 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 1 0】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 1 1】関連データ共起ファイルの例
- 【図 1 2】共起データ共起ファイルの例
- 【図 1 3】関連データ共起ファイルの例
- 【図 1 4】共起データ共起ファイルの例
- 【図 1 5】共起データ共起ファイルの例

30

【符号の説明】

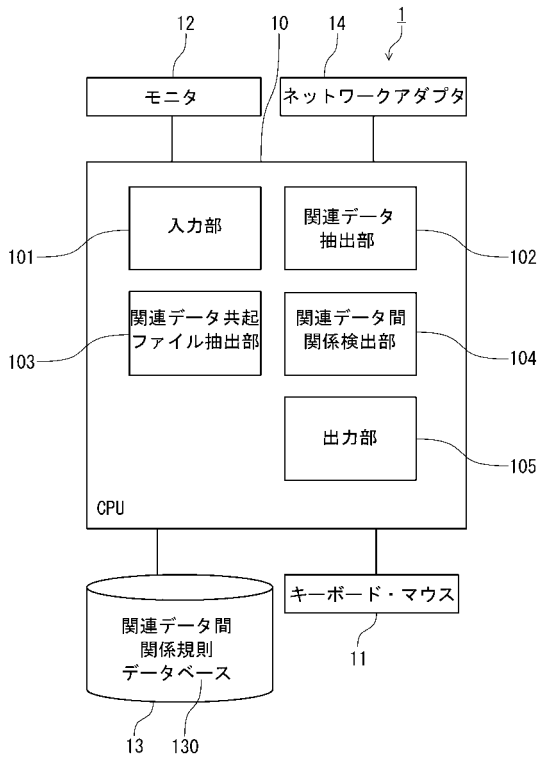
【 0 2 0 6】

- 1 データ処理装置
- 1 0 C P U
- 1 1 キーボード・マウス
- 1 2 モニタ
- 1 3 ハードディスク
- 1 4 ネットワークアダプタ
- 1 0 1 入力部
- 1 0 2 関連データ抽出部
- 1 0 3 関連データ共起ファイル抽出部
- 1 0 4 関連データ間関係検出部
- 1 0 5 出力部
- 1 3 0 関連データ間関係規則データベース

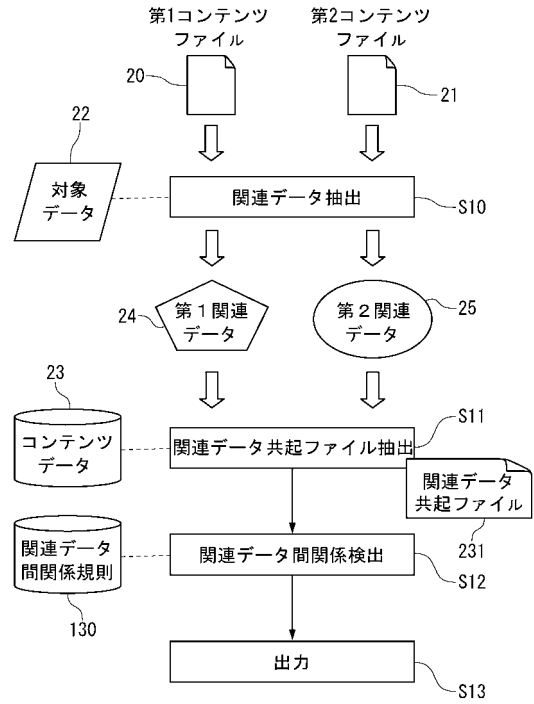
40

50

【 図 1 】

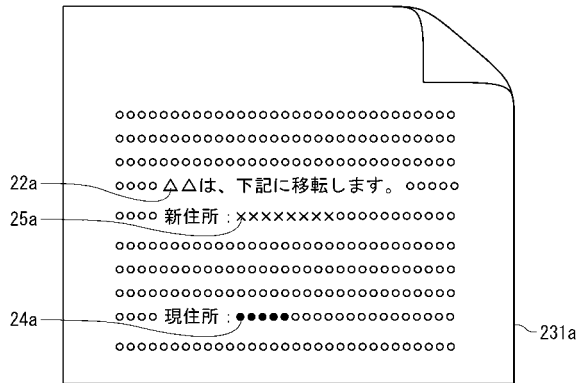


【 図 2 】

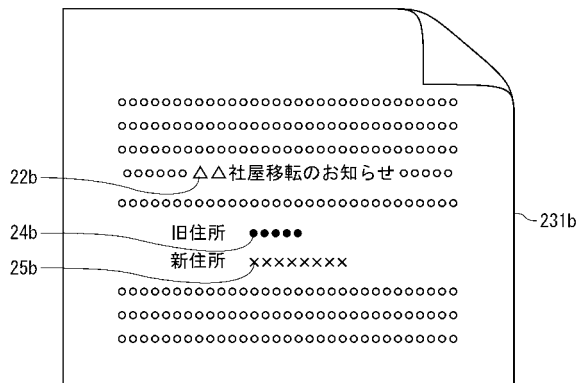


【 図 3 】

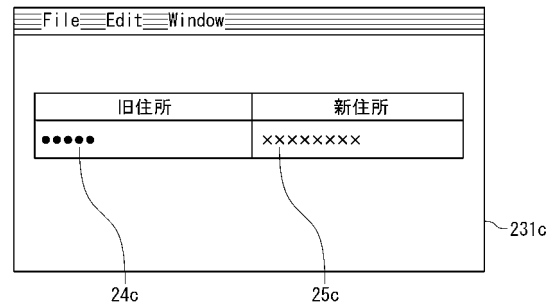
(A)



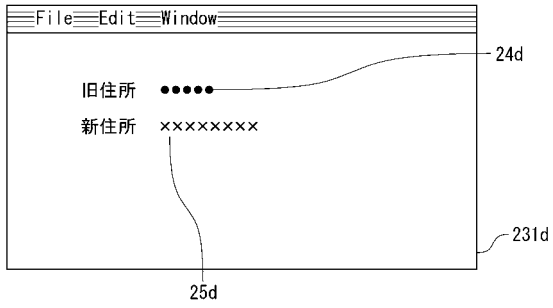
(B)



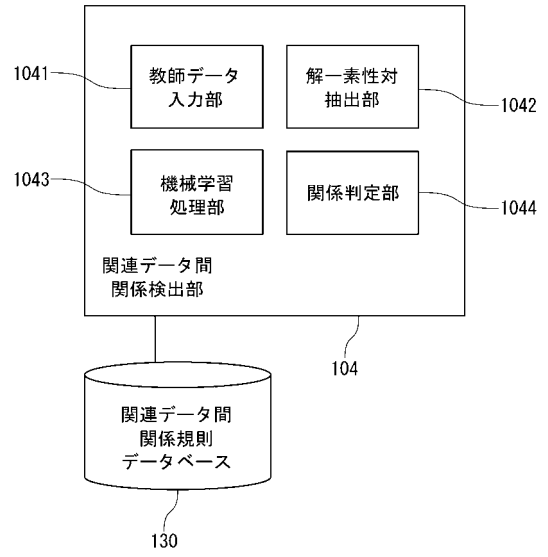
【 図 4 】



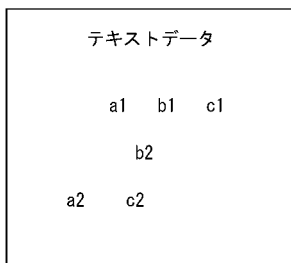
【 図 5 】



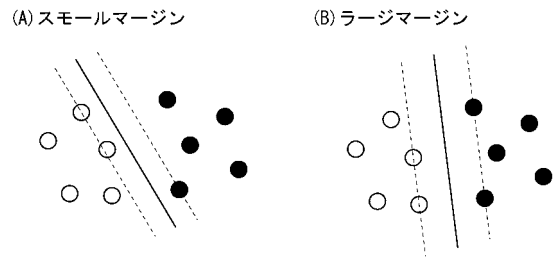
【 図 6 】



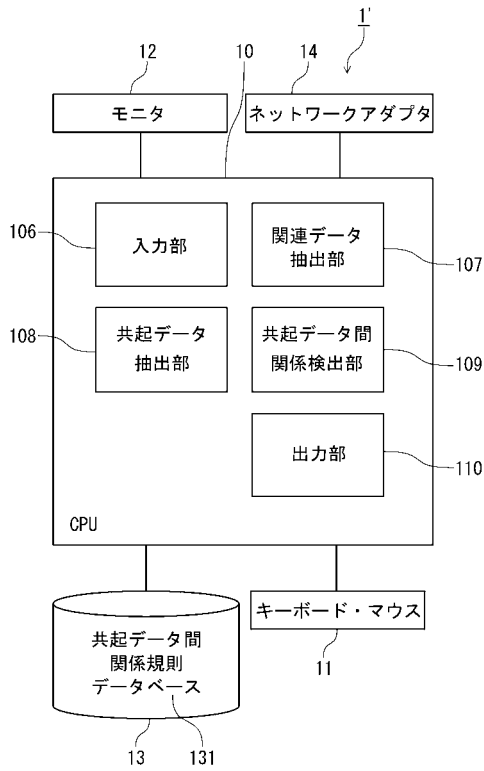
【 図 7 】



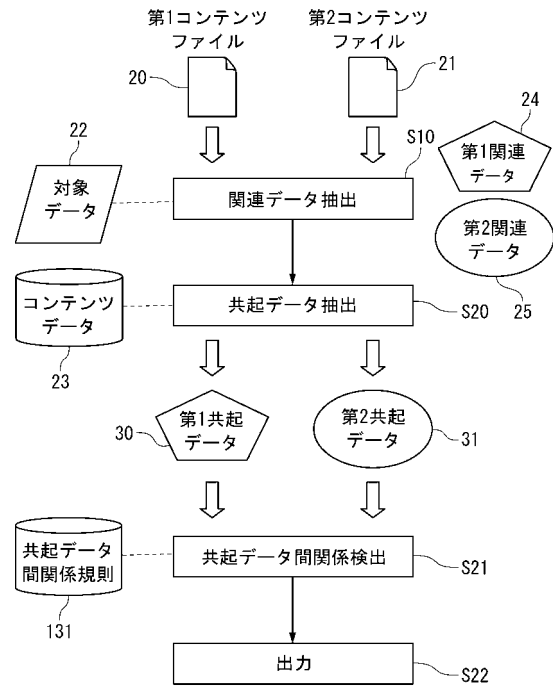
【 図 8 】



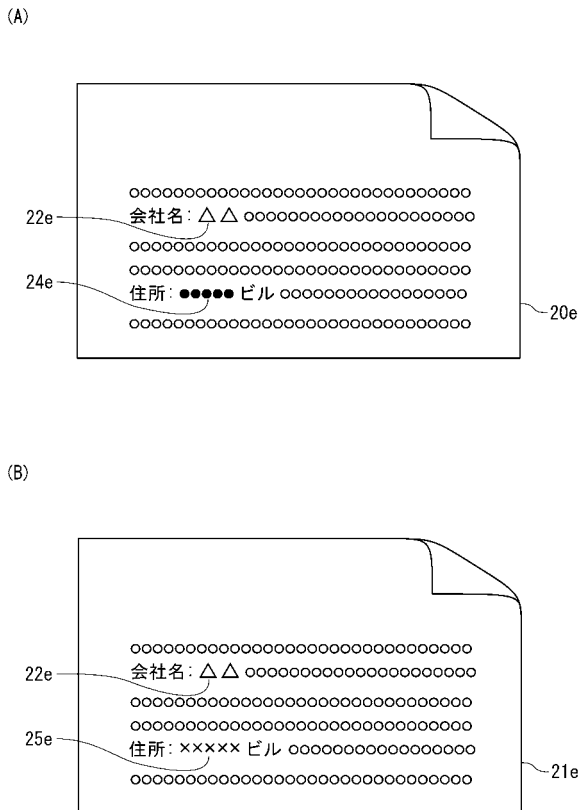
【 図 9 】



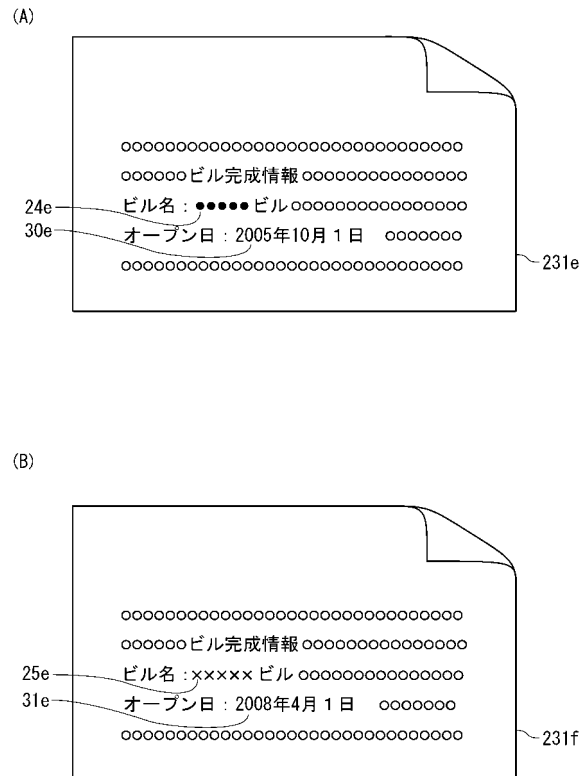
【 図 1 0 】



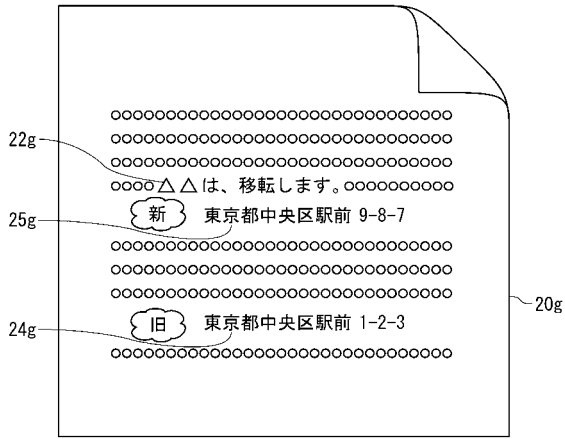
【 図 1 1 】



【 図 1 2 】

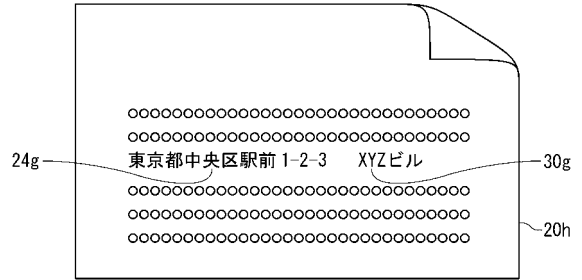


【 図 1 3 】

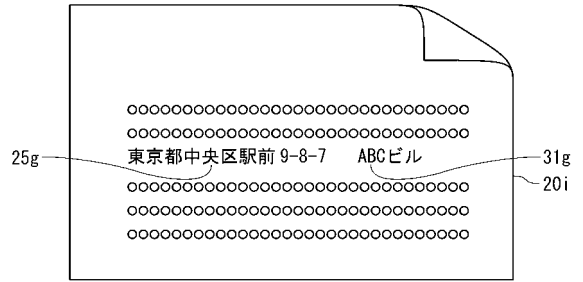


【 図 1 4 】

(A)



(B)



【 図 1 5 】

