

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7628266号
(P7628266)

(45)発行日 令和7年2月10日(2025.2.10)

(24)登録日 令和7年1月31日(2025.1.31)

(51)国際特許分類

F I

G 0 6 Q 10/063 (2023.01)

G 0 6 Q 10/063

請求項の数 12 (全31頁)

(21)出願番号	特願2024-117155(P2024-117155)	(73)特許権者	515102378
(22)出願日	令和6年7月22日(2024.7.22)		株式会社W A C U L
(62)分割の表示	特願2020-97382(P2020-97382)の分割		東京都千代田区神田小川町3-26-8
原出願日	令和2年6月4日(2020.6.4)	(74)代理人	- 2 F
(65)公開番号	特開2024-160705(P2024-160705 A)		100179970
(43)公開日	令和6年11月14日(2024.11.14)	(74)代理人	弁理士 桐山 大
審査請求日	令和6年7月22日(2024.7.22)		100071205
早期審査対象出願		(74)代理人	弁理士 野本 陽一
		(72)発明者	垣内 勇威
			東京都千代田区神田小川町3-26-8
		(72)発明者	- 2 F 株式会社W A C U L 内
			鎌谷 崇広
			東京都千代田区神田小川町3-26-8
			- 2 F 株式会社W A C U L 内
		審査官	塩田 徳彦

最終頁に続く

(54)【発明の名称】 データ分析システム及び分類処理方法

(57)【特許請求の範囲】

【請求項1】

ウェブサイトにおけるウェブページに紐付くページ情報を、過去における複数の前記ページ情報を用いた機械学習により生成された分類モデルの入力とし、入力とした前記ページ情報を性質の類似性に基づく複数のグループのうちの何れかに分類する分類処理手段を有する分類装置と、

前記分類装置による分類で得られたデータに分析処理を施す分析装置と、を有し、

前記分類装置は、

前記ウェブページにおける、表示画面上に配置される各オブジェクトの情報を含む表示関連データを用いて、該ウェブページを各オブジェクトの種別に対応するパーツ情報を含むワイヤフレーム画像に変換し、該ワイヤフレーム画像から、オブジェクトごとの、該オブジェクトの種別を示す種別情報と、該オブジェクトの位置及びサイズを示す配置情報との組み合わせによる識別情報を生成し、生成したオブジェクトごとの識別情報に基づいて固定長のベクトル表現であるフレーム特徴量を前記ページ情報として生成する前処理手段を有する、データ分析システム。

【請求項2】

ウェブサイトにおけるウェブページに紐付くページ情報を、過去における複数の前記ページ情報を用いた機械学習により生成された分類モデルの入力とし、入力とした前記ページ情報を性質の類似性に基づく複数のグループのうちの何れかに分類する分類処理手段を有する分類装置と、

前記分類装置による分類で得られたデータに分析処理を施す分析装置と、を有し、
前記分類装置は、

前記ウェブページにおける、表示画面上に配置される各オブジェクトの情報を含む表示
関連データを用いて、該ウェブページを各オブジェクトの種別に対応するパーツ情報を含む
ワイヤフレーム画像に変換し、該ワイヤフレーム画像から、オブジェクトごとの、該オ
ブジェクトの種別を示す種別情報と、該オブジェクトの位置及びサイズを示す配置情報と
の組み合わせによる識別情報を生成し、生成したオブジェクトごとの識別情報に基づいて
固定長のベクトル表現であるフレーム特徴量を生成する前処理手段を有し、

前記分類処理手段は、

前記ウェブページの位置を示すアドレス情報を、過去における複数の前記アドレス情報
を用いた機械学習により生成されたアドレス分類モデルの入力としてアドレス特徴量を抽
出するアドレス分類手段と、

前記フレーム特徴量を、過去における複数の前記フレーム特徴量を用いた機械学習によ
り生成されたページ分類モデルの入力としてページ特徴量を抽出するページ分類手段と、

前記ウェブページごとの前記アドレス特徴量及び前記ページ特徴量を前記ページ情報と
して前記分類モデルに入力して各ウェブページのグルーピングを行う統括分類手段と、を
有する、データ分析システム。

【請求項 3】

ウェブサイトにおけるウェブページに紐付くページ情報を、過去における複数の前記ペ
ージ情報を用いた機械学習により生成された分類モデルの入力とし、入力とした前記ペ
ージ情報を性質の類似性に基づく複数のグループのうちの何れかに分類する分類処理手段を
有する分類装置と、

前記分類装置による分類で得られたデータに分析処理を施す分析装置と、を有し、

前記分類装置は、

前記ウェブページにおける、表示画面上に配置される各オブジェクトの情報を含む表示
関連データを用いて、該ウェブページを各オブジェクトの種別に対応するパーツ情報を含む
ワイヤフレーム画像に変換し、該ワイヤフレーム画像から、オブジェクトごとの、該オ
ブジェクトの種別を示す種別情報と、該オブジェクトの位置及びサイズを示す配置情報と
の組み合わせによる識別情報を生成し、生成したオブジェクトごとの識別情報に基づいて
固定長のベクトル表現であるフレーム特徴量を生成する前処理手段を有し、

前記分類処理手段は、

前記ウェブページ内のテキストの情報であるテキストデータを、過去における複数の前
記テキストデータを用いた機械学習により生成されたテキスト分類モデルの入力としてテ
キスト特徴量を抽出するテキスト分類手段と、

前記フレーム特徴量を、過去における複数の前記フレーム特徴量を用いた機械学習によ
り生成されたページ分類モデルの入力としてページ特徴量を抽出するページ分類手段と、

前記ウェブページごとの前記テキスト特徴量及び前記ページ特徴量を前記ページ情報と
して前記分類モデルに入力して各ウェブページのグルーピングを行う統括分類手段と、を
有する、データ分析システム。

【請求項 4】

ウェブサイトにおけるウェブページに紐付くページ情報を、過去における複数の前記ペ
ージ情報を用いた機械学習により生成された分類モデルの入力とし、入力とした前記ペ
ージ情報を性質の類似性に基づく複数のグループのうちの何れかに分類する分類処理手段を
有する分類装置と、

前記分類装置による分類で得られたデータに分析処理を施す分析装置と、を有し、

前記分類装置は、

前記ウェブページ内のテキストの情報であるテキストデータに解析処理を施して解析デ
ータを生成するテキスト前処理手段と、

前記ウェブページにおける、表示画面上に配置される各オブジェクトの情報を含む表示
関連データを用いて、該ウェブページを各オブジェクトの種別に対応するパーツ情報を含

10

20

30

40

50

むワイヤフレーム画像に変換し、該ワイヤフレーム画像から、オブジェクトごとの、該オブジェクトの種別を示す種別情報と、該オブジェクトの位置及びサイズを示す配置情報との組み合わせによる識別情報を生成し、生成したオブジェクトごとの識別情報に基づいて固定長のベクトル表現であるフレーム特徴量を生成するページ前処理手段と、を有し、

前記分類処理手段は、

前記解析データを、過去における複数の前記解析データを用いた機械学習により生成されたテキスト分類モデルの入力としてテキスト特徴量を抽出するテキスト分類手段と、

前記フレーム特徴量を、過去における複数の前記フレーム特徴量を用いた機械学習により生成されたページ分類モデルの入力としてページ特徴量を抽出するページ分類手段と、

前記ウェブページごとの前記テキスト特徴量及び前記ページ特徴量を前記ページ情報として前記分類モデルに入力して各ウェブページのグルーピングを行う統括分類手段と、を有する、データ分析システム。

【請求項 5】

ウェブサイトにおけるウェブページに紐付くページ情報を、過去における複数の前記ページ情報を用いた機械学習により生成された分類モデルの入力とし、入力とした前記ページ情報を性質の類似性に基づく複数のグループのうちの何れかに分類する分類処理手段を有する分類装置と、

前記分類装置による分類で得られたデータに分析処理を施す分析装置と、を有し、

前記分類装置は、

前記ウェブページにおける、表示画面上に配置される各オブジェクトの情報を含む表示関連データを用いて、該ウェブページを各オブジェクトの種別に対応するパーツ情報を含むワイヤフレーム画像に変換し、該ワイヤフレーム画像から、オブジェクトごとの、該オブジェクトの種別を示す種別情報と、該オブジェクトの位置及びサイズを示す配置情報との組み合わせによる識別情報を生成し、生成したオブジェクトごとの識別情報に基づいて固定長のベクトル表現であるフレーム特徴量を生成する前処理手段を有し、

前記分類処理手段は、

前記ウェブページの位置を示すアドレス情報を、過去における複数の前記アドレス情報を用いた機械学習により生成されたアドレス分類モデルの入力としてアドレス特徴量を抽出するアドレス分類手段と、

前記ウェブページ内のテキストの情報であるテキストデータを、過去における複数の前記テキストデータを用いた機械学習により生成されたテキスト分類モデルの入力としてテキスト特徴量を抽出するテキスト分類手段と、

前記フレーム特徴量を、過去における複数の前記フレーム特徴量を用いた機械学習により生成されたページ分類モデルの入力としてページ特徴量を抽出するページ分類手段と、

前記ウェブページごとの前記アドレス特徴量、前記テキスト特徴量、及び前記ページ特徴量を前記ページ情報として前記分類モデルに入力して各ウェブページのグルーピングを行う統括分類手段と、を有する、データ分析システム。

【請求項 6】

ウェブサイトにおけるウェブページに紐付くページ情報を、過去における複数の前記ページ情報を用いた機械学習により生成された分類モデルの入力とし、入力とした前記ページ情報を性質の類似性に基づく複数のグループのうちの何れかに分類する分類処理手段を有する分類装置と、

前記分類装置による分類で得られたデータに分析処理を施す分析装置と、を有し、

前記分類装置は、

前記ウェブページ内のテキストの情報であるテキストデータに解析処理を施して解析データを生成するテキスト前処理手段と、

前記ウェブページにおける、表示画面上に配置される各オブジェクトの情報を含む表示関連データを用いて、該ウェブページを各オブジェクトの種別に対応するパーツ情報を含むワイヤフレーム画像に変換し、該ワイヤフレーム画像から、オブジェクトごとの、該オブジェクトの種別を示す種別情報と、該オブジェクトの位置及びサイズを示す配置情報と

10

20

30

40

50

の組み合わせによる識別情報を生成し、生成したオブジェクトごとの識別情報に基づいて固定長のベクトル表現であるフレーム特徴量を生成するページ前処理手段と、を有し、

前記分類処理手段は、

前記ウェブページの位置を示すアドレス情報を、過去における複数の前記アドレス情報を用いた機械学習により生成されたアドレス分類モデルの入力としてアドレス特徴量を抽出するアドレス分類手段と、

前記解析データを、過去における複数の前記解析データを用いた機械学習により生成されたテキスト分類モデルの入力としてテキスト特徴量を抽出するテキスト分類手段と、

前記フレーム特徴量を、過去における複数の前記フレーム特徴量を用いた機械学習により生成されたページ分類モデルの入力としてページ特徴量を抽出するページ分類手段と、

前記ウェブページごとの前記アドレス特徴量、前記テキスト特徴量、及び前記ページ特徴量を前記ページ情報として前記分類モデルに入力して各ウェブページのグルーピングを行う統括分類手段と、を有する、データ分析システム。

【請求項 7】

ウェブサイトにおけるウェブページに紐付くページ情報を、過去における複数の前記ページ情報を用いた機械学習により生成された分類モデルの入力とし、入力とした前記ページ情報を性質の類似性に基づく複数のグループのうちの何れかに分類する分類処理手段を有する分類装置と、

前記分類装置による分類で得られたデータに分析処理を施す分析装置と、を有し、

前記分類装置は、

前記ウェブページにおける、表示画面上に配置される各オブジェクトの情報を含む表示関連データを用いて、該ウェブページを各オブジェクトの種別に対応するパーツ情報を含むワイヤフレーム画像に変換し、該ワイヤフレーム画像から、オブジェクトごとの、該オブジェクトの種別を示す種別情報と、該オブジェクトの位置及びサイズを示す配置情報との組み合わせによる識別情報を生成し、生成したオブジェクトごとの識別情報に基づいて固定長のベクトル表現であるフレーム特徴量を生成するページ前処理手段を有し、

前記分類処理手段は、

前記ウェブページの位置を示すアドレス情報、及び該ウェブページに対応する前記フレーム特徴量を、前記ページ情報として前記分類モデルの入力とし、入力とした前記ページ情報を複数の前記グループのうちの何れかに分類するものである、データ分析システム。

【請求項 8】

ウェブサイトにおけるウェブページに紐付くページ情報を、過去における複数の前記ページ情報を用いた機械学習により生成された分類モデルの入力とし、入力とした前記ページ情報を性質の類似性に基づく複数のグループのうちの何れかに分類する分類処理手段を有する分類装置と、

前記分類装置による分類で得られたデータに分析処理を施す分析装置と、を有し、

前記分類装置は、

前記ウェブページにおける、表示画面上に配置される各オブジェクトの情報を含む表示関連データを用いて、該ウェブページを各オブジェクトの種別に対応するパーツ情報を含むワイヤフレーム画像に変換し、該ワイヤフレーム画像から、オブジェクトごとの、該オブジェクトの種別を示す種別情報と、該オブジェクトの位置及びサイズを示す配置情報との組み合わせによる識別情報を生成し、生成したオブジェクトごとの識別情報に基づいて固定長のベクトル表現であるフレーム特徴量を生成するページ前処理手段を有し、

前記分類処理手段は、

前記ウェブページ内のテキストの情報であるテキストデータ、及び該ウェブページに対応する前記フレーム特徴量を、前記ページ情報として前記分類モデルの入力とし、入力とした前記ページ情報を複数の前記グループのうちの何れかに分類するものである、データ分析システム。

【請求項 9】

ウェブサイトにおけるウェブページに紐付くページ情報を、過去における複数の前記ペ

10

20

30

40

50

ージ情報を用いた機械学習により生成された分類モデルの入力とし、入力とした前記ページ情報を性質の類似性に基づく複数のグループのうちの何れかに分類する分類処理手段を有する分類装置と、

前記分類装置による分類で得られたデータに分析処理を施す分析装置と、を有し、

前記分類装置は、

前記ウェブページ内のテキストの情報であるテキストデータに解析処理を施して解析データを生成するテキスト前処理手段と、

前記ウェブページにおける、表示画面上に配置される各オブジェクトの情報を含む表示関連データを用いて、該ウェブページを各オブジェクトの種別に対応するパーツ情報を含むワイヤフレーム画像に変換し、該ワイヤフレーム画像から、オブジェクトごとの、該オブジェクトの種別を示す種別情報と、該オブジェクトの位置及びサイズを示す配置情報との組み合わせによる識別情報を生成し、生成したオブジェクトごとの識別情報に基づいて固定長のベクトル表現であるフレーム特徴量を生成するページ前処理手段と、を有し、

前記分類処理手段は、

前記ウェブページに対応する前記解析データ、及び該ウェブページに対応する前記フレーム特徴量を、前記ページ情報として前記分類モデルの入力とし、入力とした前記ページ情報を複数の前記グループのうちの何れかに分類するものである、データ分析システム。

【請求項 10】

ウェブサイトにおけるウェブページに紐付くページ情報を、過去における複数の前記ページ情報を用いた機械学習により生成された分類モデルの入力とし、入力とした前記ページ情報を性質の類似性に基づく複数のグループのうちの何れかに分類する分類処理手段を有する分類装置と、

前記分類装置による分類で得られたデータに分析処理を施す分析装置と、を有し、

前記分類装置は、

前記ウェブページにおける、表示画面上に配置される各オブジェクトの情報を含む表示関連データを用いて、該ウェブページを各オブジェクトの種別に対応するパーツ情報を含むワイヤフレーム画像に変換し、該ワイヤフレーム画像から、オブジェクトごとの、該オブジェクトの種別を示す種別情報と、該オブジェクトの位置及びサイズを示す配置情報との組み合わせによる識別情報を生成し、生成したオブジェクトごとの識別情報に基づいて固定長のベクトル表現であるフレーム特徴量を生成する前処理手段を有し、

前記分類処理手段は、

前記ウェブページの位置を示すアドレス情報、該ウェブページ内のテキストの情報であるテキストデータ、及び該ウェブページに対応する前記フレーム特徴量を、前記ページ情報として前記分類モデルの入力とし、入力とした前記ページ情報を複数の前記グループのうちの何れかに分類するものである、データ分析システム。

【請求項 11】

ウェブサイトにおけるウェブページに紐付くページ情報を、過去における複数の前記ページ情報を用いた機械学習により生成された分類モデルの入力とし、入力とした前記ページ情報を性質の類似性に基づく複数のグループのうちの何れかに分類する分類処理手段を有する分類装置と、

前記分類装置による分類で得られたデータに分析処理を施す分析装置と、を有し、

前記分類装置は、

前記ウェブページ内のテキストの情報であるテキストデータに解析処理を施して解析データを生成するテキスト前処理手段と、

前記ウェブページにおける、表示画面上に配置される各オブジェクトの情報を含む表示関連データを用いて、該ウェブページを各オブジェクトの種別に対応するパーツ情報を含むワイヤフレーム画像に変換し、該ワイヤフレーム画像から、オブジェクトごとの、該オブジェクトの種別を示す種別情報と、該オブジェクトの位置及びサイズを示す配置情報との組み合わせによる識別情報を生成し、生成したオブジェクトごとの識別情報に基づいて固定長のベクトル表現であるフレーム特徴量を生成するページ前処理手段と、を有し、

10

20

30

40

50

前記分類処理手段は、

前記ウェブページの位置を示すアドレス情報、該ウェブページに対応する前記解析データ、及び該ウェブページに対応する前記フレーム特徴量を、前記ページ情報として前記分類モデルの入力とし、入力とした前記ページ情報を複数の前記グループのうちの何れかに分類するものである、データ分析システム。

【請求項 12】

ウェブサイトにおけるウェブページに紐付くページ情報を分類する分類装置が、

前記ウェブページにおける、表示画面上に配置される各オブジェクトの情報を含む表示関連データを用いて、該ウェブページを各オブジェクトの種別に対応するパーツ情報を含むワイヤフレーム画像に変換し、該ワイヤフレーム画像から、オブジェクトごとの、該オブジェクトの種別を示す種別情報と、該オブジェクトの位置及びサイズを示す配置情報との組み合わせによる識別情報を生成し、

生成したオブジェクトごとの識別情報に基づいて固定長のベクトル表現であるフレーム特徴量を前記ページ情報として生成し、

生成した前記フレーム特徴量を、過去における複数の前記フレーム特徴量を用いた機械学習により生成された分類モデルの入力とし、入力とした前記フレーム特徴量を性質の類似性に基づく複数のグループのうちの何れかに分類し、

前記分類装置による分類で得られたデータに分析装置が分析処理を施す、分類処理方法。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、ウェブサイトにおけるウェブページに紐付くページ情報を分類するデータ分析システム及び分類処理方法に関する。

【背景技術】

【0002】

ウェブサイトは、企業の取り扱う商品やサービスの宣伝の他、それらの取引にも利用されている。そのため、ウェブサイトの運営者等は、ウェブサイトの各ウェブページに対するユーザの閲覧行動を、各ウェブページに紐付く情報から解析し、その解析結果から課題を見出して種々の改善を行っている。ただし、こうした情報は大量に発生するため、解析精度及び解析速度を上げるためには、解析用の情報のグルーピング（分類処理）が必要となる。例えば、特許文献1の分析システムは、類似の性質を示すページや類似の検索キーワードをグルーピングした後で、所定の分析処理を行うようになっている。

【先行技術文献】

【特許文献】

【0003】

【文献】特開2016-201080号公報

【発明の概要】

【発明が解決しようとする課題】

【0004】

しかしながら、特許文献1のような従来の構成では、ページやキーワードの類似性に関する設定を手作業で行う必要があるため、大量に発生するページ情報のグルーピングには適さない、という実情がある。

【0005】

本発明は、上述のような課題を解決するためになされたものであり、ページ情報のグルーピングを自動的に効率よく行うデータ分析システム及び分類処理方法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】

【0006】

本発明の一態様に係るデータ分析システムは、ウェブサイトにおけるウェブページに紐付くページ情報を、過去における複数のページ情報を用いた機械学習により生成された分

10

20

30

40

50

類モデルの入力とし、入力としたページ情報を性質の類似性に基づく複数のグループのうちの何れかに分類する分類処理手段を有する分類装置と、分類装置による分類で得られたデータに分析処理を施す分析装置と、を有し、分類装置は、ウェブページの各オブジェクトの情報を含む表示関連データから、オブジェクトごとに、該オブジェクトの種別を示す種別情報と、該オブジェクトの位置及びサイズを示す配置情報との組み合わせによる識別情報を生成し、生成したオブジェクトごとの識別情報に基づいてページ情報としてのフレーム特徴量を生成する前処理手段を有するものである。

【0007】

本発明の一態様に係るデータ分析システムは、ウェブサイトにおけるウェブページに紐付くページ情報を、過去における複数のページ情報を用いた機械学習により生成された分類モデルの入力とし、入力としたページ情報を性質の類似性に基づく複数のグループのうちの何れかに分類する分類処理手段を有する分類装置と、分類装置による分類で得られたデータに分析処理を施す分析装置と、を有し、分類処理手段は、ウェブページの位置を示すアドレス情報を、過去における複数のアドレス情報を用いた機械学習により生成されたアドレス分類モデルの入力としてアドレス特徴量を抽出するアドレス分類手段と、ウェブページ内のテキストの情報であるテキストデータを、過去における複数のテキストデータを用いた機械学習により生成されたテキスト分類モデルの入力としてテキスト特徴量を抽出するテキスト分類手段と、ウェブページごとのアドレス特徴量及びテキスト特徴量をページ情報として分類モデルに入力して各ウェブページのグルーピングを行う統括分類手段と、を有するものである。

【0008】

本発明の一態様に係るデータ分析システムは、ウェブサイトにおけるウェブページに紐付くページ情報を、過去における複数のページ情報を用いた機械学習により生成された分類モデルの入力とし、入力としたページ情報を性質の類似性に基づく複数のグループのうちの何れかに分類する分類処理手段を有する分類装置と、分類装置による分類で得られたデータに分析処理を施す分析装置と、を有し、分類処理手段は、ウェブページの位置を示すアドレス情報、及び該ウェブページ内のテキストの情報であるテキストデータを、ページ情報として分類モデルの入力とし、入力としたページ情報を複数のグループのうちの何れかに分類するものである。

【0009】

本発明の一態様に係る分類処理方法は、ウェブサイトにおけるウェブページに紐付くページ情報を分類する分類装置が、ウェブページの各オブジェクトの情報を含む表示関連データから、オブジェクトごとに、該オブジェクトの種別を示す種別情報と、該オブジェクトの位置及びサイズを示す配置情報との組み合わせによる識別情報を生成し、生成したオブジェクトごとの識別情報に基づいてページ情報としてのフレーム特徴量を生成し、生成したフレーム特徴量を、過去における複数のフレーム特徴量を用いた機械学習により生成された分類モデルの入力とし、入力としたフレーム特徴量を性質の類似性に基づく複数のグループのうちの何れかに分類し、分類装置による分類で得られたデータに分析装置が分析処理を施すようになっている。

【発明の効果】

【0010】

本発明は、過去における複数のページ情報に基づく分類モデルの入力としたページ情報を、性質の類似性に基づく複数のグループのうちの何れかに分類するようになっている。よって、ウェブページに紐付くページ情報のグルーピングを手作業によらず自動的に行うことができるため、処理の迅速化及び効率化を図ることができる。

【図面の簡単な説明】

【0011】

【図1】本発明の実施の形態1に係るデータ分析システム及びその周辺機器等の構成例を示すブロック図である。

【図2】図1の管理装置の機能的な構成を例示したブロック図である。

【図 3】図 1 の端末装置の機能的な構成を例示したブロック図である。

【図 4】図 1 のデータ分析システムの機能的な構成を例示したブロック図である。

【図 5】図 1 のデータ分析システムの動作を例示したフローチャートである。

【図 6】本発明の実施の形態 2（3～5）に係るデータ分析システム及びその周辺機器等の構成例を示すブロック図である。

【図 7】図 6 のデータ分析システムの機能的な構成を例示したブロック図である。

【図 8】本発明の実施の形態 3 に係るデータ分析システムの機能的な構成を例示したブロック図である。

【図 9】図 6 の端末装置の表示部などに表示されるウェブページを例示した説明図である。

【図 10】図 9 のウェブページ中の各オブジェクトに対応するパーツ情報のみを画像として抽出したワイヤフレーム画像を例示した説明図である。

【図 11】本発明の実施の形態 4 に係るデータ分析システムの機能的な構成を例示したブロック図である。

【図 12】図 11 の分類装置におけるデータの流れを分類処理方法の各工程に対応づけて例示した説明図である。

【図 13】本発明の実施の形態 4 の変形例 4 a に係るデータ分析システムの機能的な構成を例示したブロック図である。

【図 14】本発明の実施の形態 4 の変形例 4 b に係るデータ分析システムの機能的な構成を例示したブロック図である。

【図 15】本発明の実施の形態 4 の変形例 4 c に係るデータ分析システムの機能的な構成を例示したブロック図である。

【図 16】本発明の実施の形態 5 に係るデータ分析システムの機能的な構成を例示したブロック図である。

【図 17】本発明の実施の形態 5 の変形例 5 a に係るデータ分析システムの機能的な構成を例示したブロック図である。

【図 18】本発明の実施の形態 5 の変形例 5 b に係るデータ分析システムの機能的な構成を例示したブロック図である。

【発明を実施するための形態】

【0012】

実施の形態 1.

図 1 を参照し、本実施の形態 1 におけるデータ分析システム及びその周辺機器等の構成例について説明する。図 1 に示すように、データ分析システム 100 は、分類装置 10 と、分析装置 20 と、を有している。データ分析システム 100 と、管理装置 30 と、サービス提供システム 40 と、解析システム 50 と、端末装置 80 とは、インターネットなどのネットワーク N を介して通信可能に接続されている。すなわち、データ分析システム 100 は、管理装置 30、サービス提供システム 40、解析システム 50、及び端末装置 80 との間で有線又は無線による通信を行うことができる。

【0013】

管理装置 30 は、ウェブサイトの運営者等、つまりサービス提供システム 40 を用いて商品やサービスの販売を行う運営者等が管理する PC（Personal Computer）である。PC には、タブレット PC、ノート PC、デスクトップ型 PC などが含まれる。もっとも、管理装置 30 は、スマートフォン又はタブレット端末などのモバイル端末であってもよい。

【0014】

サービス提供システム 40 は、EC サイト（Electronic Commerce site）などのウェブサイトを介して商品やサービスの情報を端末装置 80 等に提供する Web サーバである。サービス提供システム 40 が提供するウェブサイトは、ネットワーク N を介して端末装置 80 等から閲覧することができる。サービス提供システム 40 は、クラウドコンピューティングに基づくクラウドサーバ又は物理サーバなどにより構成される。

【0015】

解析システム 50 は、ウェブサイトのアクセス解析を行う機能を備えたものであり、例えば、Google（登録商標）提供のアクセス解析ツールである Google アナリティクス（G A）のサーバがこれに相当する。アクセス解析とは、ウェブサイトにおけるコンバージョン（C V：Conversion）の数を増やすために、ウェブサイトの訪問者であるユーザの特性や行動を分析することである。コンバージョンは、ウェブサイトの目的となるゴール、つまりウェブサイトでユーザに達成してもらいたい行動のことであり、登録、商品やサービスの購入、申し込み、資料請求、予約の実施などを指す。

【0016】

解析システム 50 は、ネットワークを介してのユーザの行動を示す行動データを、サービス提供システム 40 等から逐次又は定期的に収集する。そして、本実施の形態 1 の解析システム 50 は、収集した行動データにアクセス解析を施すことにより、ネットワーク N 上に存在する情報資源の位置を示すアドレス情報を生成して蓄積する。解析システム 50 は、データ分析システム 100 からの要求に応じてアドレス情報を返送する。解析システム 50 は、クラウドコンピューティングに基づくクラウドサーバ又は物理サーバなどにより構成される。

【0017】

端末装置 80 は、ユーザにより使用されるモバイル端末又は P C である。端末装置 80 は、ユーザの操作に応じて、ウェブサイトへのアクセスやログインなどを行うと共に、ウェブサイト内の種々の情報を表示し、売買取引の中継などを行う。ユーザは、1 又は複数の端末装置 80 を使用し、ネットワーク N を介してサービス提供システム 40 にアクセスする。

【0018】

データ分析システム 100 は、ウェブサイトにおけるウェブページに紐付くページ情報として、ウェブページの位置を示すアドレス情報を取得し、取得したアドレス情報に分類処理を施した上で、所定の分析処理を行うものである。データ分析システム 100 は、クラウドコンピューティングに基づくクラウドサーバ、もしくは物理サーバ、又はこれらを組み合わせたシステムなどにより構成される。なお、データ分析システム 100 は、P C により構成してもよい。以降では、ウェブサイトのことを「サイト」ともいい、ウェブページのことを「ページ」ともいう。

【0019】

次に、図 2 を参照して、管理装置 30 の機能的な構成について説明する。管理装置 30 は、管理通信部 31 と、管理制御部 32 と、管理記憶部 33、入力部 34 と、表示部 35 と、を有している。管理記憶部 33 には、管理制御部 32 の動作プログラムの他、種々の情報が記憶される。管理記憶部 33 は、R A M（Random Access Memory）及び R O M（Read Only Memory）、フラッシュメモリ等の P R O M（Programmable R O M）、又は H D D（Hard Disk Drive）などにより構成される。

【0020】

入力部 34 は、例えば、キーボードと、マウス又はトラックボールなどのポインティングデバイスと、を含んで構成される。入力部 34 は、ユーザによる入力操作を受け付け、入力操作の内容に応じた操作信号を管理制御部 32 へ送信する。表示部 35 は、例えば液晶ディスプレイ（L C D：Liquid Crystal Display）からなり、管理制御部 32 からの指示により種々の情報を表示する。

【0021】

管理制御部 32 は、情報処理手段 32 a と、表示処理手段 32 b と、を有している。情報処理手段 32 a は、表示部 35 への表示に関する操作を受け付けると、該操作に応じた制御信号を表示処理手段 32 b へ出力する。表示処理手段 32 b は、情報処理手段 32 a からの制御信号に応じて、表示部 35 に文字や画像などを表示させる。

【0022】

管理制御部 32 は、C P U（Central Processing Unit）又は G P U（Graphics Processing Unit）などの演算装置と、こうした演算装置と協働して上記の各種機能を実現さ

10

20

30

40

50

せる動作プログラムとにより構成することができる。なお、管理装置 30 は、入力部 34 及び表示部 35 の代わりに、文字又は画像等を表示する表示パネルと、該表示パネルに積層されてタッチ操作を検出する検出手段と、を含むタッチパネルを有していてもよい。管理装置 30 は、タッチパネルと、マウス又はキーボード等とを併せ持つものであってもよい。

【0023】

次いで、図 3 を参照して、端末装置 80 の機能的な構成について説明する。端末装置 80 は、端末通信部 81 と、端末制御部 82 と、端末記憶部 83、入力部 84 と、表示部 85 と、を有している。端末記憶部 83 には、端末制御部 82 の動作プログラムの他、種々の情報が記憶される。端末記憶部 83 は、RAM 及び ROM、フラッシュメモリ等の PROM、又は HDD などにより構成される。

10

【0024】

入力部 84 は、例えば、キーボードと、マウス又はトラックボールなどのポインティングデバイスと、を含んで構成される。入力部 84 は、ユーザによる入力操作を受け付け、入力操作の内容に応じた操作信号を端末制御部 82 へ送信する。表示部 85 は、例えば液晶ディスプレイからなり、端末制御部 82 からの指示により種々の情報を表示する。

【0025】

端末制御部 82 は、情報処理手段 82a と、表示処理手段 82b と、を有している。情報処理手段 82a は、表示部 85 への表示に関する操作を受け付けると、該操作に応じた制御信号を表示処理手段 82b へ出力する。表示処理手段 82b は、情報処理手段 82a からの制御信号に応じて、表示部 85 に文字や画像などを表示させる。

20

【0026】

端末制御部 82 は、CPU 又は GPU などの演算装置と、こうした演算装置と協働して上記の各種機能を実現させる動作プログラムとにより構成することができる。なお、端末装置 80 は、入力部 84 及び表示部 85 の代わりに、文字又は画像等を表示する表示パネルと、該表示パネルに積層されてタッチ操作を検出する検出手段と、を含むタッチパネルを有していてもよい。端末装置 80 は、タッチパネルと、マウス又はキーボード等とを併せ持つものであってもよい。

【0027】

次に、図 4 を参照し、データ分析システム 100 の機能的な構成について説明する。図 4 に示すように、分類装置 10 は、サイトにおけるページに紐付くページ情報を分類するものである。分類装置 10 は、通信部 11 と、制御部 12 と、分類用記憶部 13 と、を有している。通信部 11 は、制御部 12 及び分析処理部 22 が、ネットワーク N に接続された機器等との間で有線又は無線による通信を行うためのインタフェースである。

30

【0028】

分類用記憶部 13 は、分類処理プログラム P1 を含む制御部 12 の動作プログラムの他、ページ情報の分類処理に関連する種々のデータを記憶する。例えば、分類用記憶部 13 には、ユーザの閲覧行動が反映されたページ情報が記憶される。分類用記憶部 13 には、ページ情報を分類（グルーピング）するためのアドレス分類モデル M1 が格納される。

【0029】

制御部 12 は、収集処理手段 12a と、学習処理手段 12c と、分類処理手段 12d と、を有している。収集処理手段 12a は、ページ情報の送信を要求する要求信号を定期的に解析システム 50 へ送信し、解析システム 50 から返送される複数のページ情報を分類用記憶部 13 に記憶させる。本実施の形態 1 において、収集処理手段 12a は、サイトにおけるページの位置を示すアドレス情報を、ページ情報として逐次又は定期的に収集して分類用記憶部 13 に記憶させる。ここで、情報資源には、サイト内のページが含まれ、アドレス情報は、URL (Uniform Resource Locator) に相当する。

40

【0030】

学習処理手段 12c は、分類用記憶部 13 に記憶された複数のページ情報を用いての機械学習により、ページ情報への分類設定を行うアドレス分類モデル M1 を生成する。つま

50

り、学習処理手段１２ｃは、過去における複数のページ情報を入力とする機械学習により、ページ情報の性質の類似性に基づくグルーピングを行うアドレス分類モデルＭ１を構築する。アドレス分類モデルＭ１は、入力されたページ情報を複数のグループのうちの何れかに分類し、分類先のグループに固有の識別タグである分類タグを付す学習済みモデルである。ページ情報のグループとは、類似の性質を示すページ情報が属するグループのことである。ページ情報の性質の類似性は、ページの性質の類似性に対応している。すなわち、ページ情報のグループとしては、商品詳細のページのグループ、会社説明のページのグループ、資料請求のページのグループ、商品詳細のページのグループなどがある。

【００３１】

ここで、収集処理手段１２ａは、複数のページ情報を経時的に収集して分類用記憶部１３に蓄積させるようになっている。そして、学習処理手段１２ｃは、設定されたタイミングで、分類用記憶部１３に新たに蓄積された複数のページ情報を取得すると共に、取得した複数のページ情報を入力とし、アドレス分類モデルＭ１の更新処理を実行する。

【００３２】

本実施の形態１において、学習処理手段１２ｃは、ＤＮＮ（Deep Neural Network）を用いた教師あり学習によりアドレス分類モデルＭ１を生成するようになっている。学習処理手段１２ｃは、教師なし学習又は半教師あり学習によりアドレス分類モデルＭ１を生成するものであってもよい。学習処理手段１２ｃは、ＧＢＤＴ（Gradient Boosting Decision Tree）を用いた機械学習によりアドレス分類モデルＭ１を生成してもよい。学習処理手段１２ｃは、キャラクタレベルＣＮＮ（Character-level Convolutional Neural Network）を用いた機械学習によりアドレス分類モデルＭ１を生成してもよい。加えて、学習処理手段１２ｃは、線形回帰、ロジスティック回帰、又は決定木などの回帰の手法によりアドレス分類モデルＭ１を生成するものであってよい。さらに、学習処理手段１２ｃは、ランダムフォレスト又はサポートベクターマシンなどの分類の手法によりアドレス分類モデルＭ１を生成するものであってよい。もっとも、学習処理手段１２ｃは、上述した複数の機械学習を組み合わせた手法によりアドレス分類モデルＭ１を生成してもよい。

【００３３】

分類処理手段１２ｄは、サイトにおけるページに紐づくページ情報を、過去における複数のページ情報を用いた機械学習により生成されたアドレス分類モデルＭ１の入力とするものである。そして、分類処理手段１２ｄは、アドレス分類モデルＭ１の入力としたページ情報を、性質の類似性に基づく複数のグループのうちの何れかに分類するものである。より具体的に、分類処理手段１２ｄは、分類対象である複数のアドレス情報を、順次アドレス分類モデルＭ１の入力とする。そして、分類処理手段１２ｄは、入力としたアドレス情報を複数のグループのうちの何れかに分類し、分類先のグループの分類タグを該アドレス情報に付して分類データを生成する。つまり、分類処理手段１２ｄは、類似の性質を示すページ情報には共通の分類タグを付し、異なる性質を示すページ情報には、相互に異なる分類タグを付すことになる。分類処理手段１２ｄは、アドレス情報に分類タグを付して生成した分類データを、分析用記憶部２３に記憶させてもよく、分析処理部２２へ出力してもよい。本実施の形態１の分類データは「アドレス分類データ」とも称する。

【００３４】

（アドレス情報の取得方法）

ここで、ページ情報としてのアドレス情報の、分類装置１０による取得パターンについて説明する。分類装置１０は、サービス提供システム４０が提供するサイトとの間で、解析システム５０を介しての連携（例えばサイトとのＧＡ連携：以下システム連携という。）がとれている場合、解析システム５０のＡＰＩ（Application Programming Interface）を利用し、サイト内に含まれるページのＵＲＬを列挙させて取得することができる。その際、分類装置１０は、各ページのＰＶ数（ページビュー数）などの情報も取得することができる。こうした情報をグルーピングに利用することができる。

【００３５】

一方で、システム連携がとれていない場合でも、分類装置１０は、例えば下記の２つの

手法により、ページのURLを直接収集することができる。

(1) サイト内のページを再帰的にクロールする手法：

例えば、サイトのトップページを最初にクロールし、そこに含まれるサイト内のリンクを収集し、それらもクロールする、というように、再帰的にクロールすることによりアドレス情報を収集することができる。かかる手法を採る場合、サイト内のページを再帰的にクロールするクローリング機能を収集処理手段12aにもたせてもよく、クローリング機能をもつ外部サーバなどからクローリング後のアドレス情報を収集処理手段12aが収集してもよい。

(2) サイトマップを活用する手法：

幾つかのサイトは、XMLサイトマップ(sitemap.xml)にサイト内のページのURLが列挙されている場合がある。すなわち、このような場合、収集処理手段12aは、XMLサイトマップからアドレス情報を収集することができる。

【0036】

このように、分類装置10は、システム連携がとれていない場合でも、インターネットに公開されているデータであれば、取得することができる。ただし、この場合は、各ページのPV数などの情報については取得することができない。つまり、分類装置10は、学習処理手段12cによる学習処理、及び分類処理手段12dによる分類処理に、各ページのPV数などの情報を利用することができない。そのため、グルーピングの精度の観点からは、システム連携がとれている方が好ましい。

【0037】

また、図4に示すように、分析装置20は、分類装置10による分類で得られた分類データに分析処理を施すものである。分析装置20は、分析処理部22と、分析用記憶部23と、を有している。分析用記憶部23は、分析処理プログラム23pを含む分析処理部22の動作プログラムの他、分類データの分析処理に関連する種々のデータを記憶する。分析処理部22は、分類処理手段12dにより生成された分類データに対し、入口分析処理などの種々の分析処理を実行する。

【0038】

入口分析処理とは、同一の流入元での入口ページ間のCVR(Conversion Rate)を比較し、入口ページの重要度を評価する処理である。入口ページとは、分析対象となるサイトを訪問したユーザが最初に閲覧したページのことである。流入元は、分析対象となるサイトの入口ページを訪問する際の経路元のことである。流入元には、ユーザのおおよその特性が表れる。流入元には、E-mail経由、Facebook(登録商標)経由、検索エンジン経由、他のサイトに貼られたリンク経由、ブラウザのブックマーク経由、及びURLの直打ち、などがある。CVRは、サイトの訪問者が該サイトの目標達成となる行動をどの程度とったかを表す指標であり、コンバージョン率ともいう。

【0039】

分析処理部22は、入口ページのグループ単位で入口分析処理を行うこともできる。例えば、分析処理部22は、同じ流入元のグループ内で、相対的にCVRが高い入口ページと、相対的にCVRが低い入口ページとを判別する。そして、入口ページへのリンクをユーザが設定可能な場合、分析処理部22は、よりCVRの高い入口ページへの誘導を推奨する推奨情報を管理装置30へ送信する。分析処理部22は、推奨情報に応じて設定変更した場合のCVRの伸びしろ等を示す改善情報を、推奨情報と共に管理装置30へ送信してもよい。

【0040】

一方、入口ページへのリンクをユーザが設定できない場合(入口ページが外部サイトなどによって決定されている場合など)、分析処理部22は、相対的にCVRが低い入口ページを、相対的にCVRが高い入口ページの態様に寄せる提案に係る提案情報を管理装置30へ送信する。分析処理部22は、提案情報に応じて設定変更した場合のCVRの伸びしろ等を示す改善情報を、提案情報と共に管理装置30へ送信してもよい。

【0041】

管理装置 30 の表示処理手段 32 b は、推奨情報又は提案情報に基づく情報を表示部 35 に表示させる。これにより、運営者等は、CVR等を高めるための施策を採ることができる。表示処理手段 32 b は、分析処理部 22 から改善情報が送信された場合、改善情報に基づく情報を表示部 35 に表示させてもよい。

【0042】

分類用記憶部 13 及び分析用記憶部 23 は、RAM 及び ROM、フラッシュメモリ等の PROM、又は HDD などにより構成することができる。制御部 12 は、CPU 又は GPU などの演算装置と、こうした演算装置と協働して上記の各種機能を実現させる分類処理プログラム P1 とにより構成することができる。すなわち、分類処理プログラム P1 は、コンピュータとしての制御部 12 及び分類用記憶部 13 を、収集処理手段 12 a、学習処理手段 12 c、及び分類処理手段 12 d として機能させるためのプログラムである。同様に、分析処理部 22 は、CPU 又は GPU などの演算装置と、こうした演算装置と協働して上記の各種機能を実現させる分析処理プログラム 23 p とにより構成することができる。

【0043】

図 2 では、各機能を明確に分けて説明するために、分類用記憶部 13 と分析用記憶部 23 とを別構成とし、制御部 12 と分析処理部 22 とを別構成として示したが、これに限定されない。例えば、制御部 12 と分析処理部 22 とは、1つの演算装置で構成してもよく、SoC (System-on-a-Chip) として 1つのチップで形成してもよい。同様に、分類用記憶部 13 と分析用記憶部 23 とは、一体的な構成であってもよい。以降の各図及びこれらに対応する説明についても同様である。

【0044】

次に、図 5 を参照して、分類装置 10 による分類処理方法について説明する。ここでは、分類処理手段 12 d が行う分類処理工程の一例について説明する。つまり、分類処理方法のうち、学習処理手段 12 c が行う学習処理工程は、図 5 の処理の前提となっている。

【0045】

まず、分類処理手段 12 d は、分類対象として分類用記憶部 13 に記憶されているページ情報を取得する (ステップ S101)。次いで、分類処理手段 12 d は、取得したページ情報をアドレス分類モデル M1 の入力とする (ステップ S102)。そして、分類処理手段 12 d は、アドレス分類モデル M1 により、入力としたページ情報の属するグループの識別タグを該ページ情報に付して分類データを生成し、生成した分類データを分類用記憶部 13 又は分析用記憶部 23 に記憶させる (ステップ S103)。

【0046】

分類処理手段 12 d は、分類対象のページ情報が分類用記憶部 13 に残存しているか否かを判定する (ステップ S104)。分類処理手段 12 d は、分類対象のページ情報が存在すれば (ステップ S104 / Yes)、ステップ S101 の処理へ移行する。分類処理手段 12 d は、分類対象のページ情報が存在しなければ (ステップ S104 / No)、分類処理を終了する。

【0047】

ここで、本実施の形態 1 における分類処理方法の概要について補足する。分類処理方法は、過去における複数のページ情報を用いた機械学習によりアドレス分類モデル M1 を生成する学習処理工程と、分類対象であるページ情報をアドレス分類モデル M1 の入力とし、入力としたページ情報を、性質の類似性に基づく複数のグループのうちの何れかに分類する分類処理工程と、を有している。

【0048】

以上のように、本実施の形態 1 の分類装置 10 は、サイトにおけるページに紐付くページ情報をアドレス分類モデル M1 の入力とすることで、入力としたページ情報を、性質の類似性に基づく複数のグループのうちの何れかに分類する分類処理手段 12 d を有している。分類処理手段 12 d は、アドレス分類モデル M1 の入力としたページ情報に、分類先のグループの識別タグを付すようになっている。よって、分析対象のページに紐付くページ情報のグルーピングを自動的に効率よく行うことができるため、処理の迅速化及び効率

化を図ることができる。

【0049】

また、本実施の形態1のページ情報は、ネットワークN上に存在する情報資源の位置を示す情報であり、URLに相当するアドレス情報である。ここで、アドレス情報は、特に、システム連携がとれている場合、又はXMLサイトマップ(sitemap.xml)にサイト内のページのURLが列挙されている場合等において、容易かつ迅速に取得することができる。そして、データの構成上、アドレス情報の分類処理は、比較的短時間で行うことができる。なお、制御部12は、学習処理手段12cを有さなくてもよい。この場合、分類装置10は、外部で構築されたアドレス分類モデルM1を分類用記憶部13に格納し、これを分類処理手段12dに利用させるとよい。

10

【0050】

実施の形態2.

図6及び図7を参照して、本実施の形態2におけるデータ分析システム200及びその周辺機器等の構成例について説明する。前述した実施の形態1と同等の構成及び構成部材については同一の符号を付して説明は省略する。

【0051】

図6に示すように、本実施の形態2のデータ分析システム200は、ネットワークNを介して、クロールシステム60との通信を行うようになっている。本実施の形態2のクロールシステム60は、サイト内のページをクロールすることにより、該ページのテキストデータを取得する機能を有している。クロールシステム60は、サイトが複数のページで構成されている場合、ページごとのテキストデータを取得する。

20

【0052】

図7に示すように、本実施の形態2における分類装置110の制御部112は、収集処理手段112aと、前処理手段112bと、学習処理手段112cと、分類処理手段112dと、を有している。収集処理手段112aは、サイトにおけるページ内のテキストの情報であるテキストデータを、逐次又は定期的にクロールシステム60から収集し、収集したテキストデータを分類用記憶部13に記憶させる。

【0053】

前処理手段112bは、サイトにおけるページ内のテキストの情報であるテキストデータに解析処理を施して、ページ情報としての解析データを生成するものである。本実施の形態4における前処理手段112bは、分類用記憶部13に記憶されているテキストデータに形態素解析処理を施すことにより、ページ情報としての解析データを生成する。形態素解析とは、自然言語で書かれたテキストデータを、言語上の最小単位である形態素に分割し(分かち書きし)、各形態素のそれぞれの品詞や変化などを判別することである。すなわち、前処理手段112bは、テキストデータから分割した各形態素のそれぞれに、動詞、名詞、形容詞などの品詞の情報を付して整理した解析データを分類用記憶部13に記憶させる。前処理手段112bは、各形態素のそれぞれに品詞の情報を付して整理する際、ベクトル表現への変換や、予め登録された頻出語に基づく重み付けなどの処理を行ってもよい。

30

【0054】

学習処理手段112cは、分類用記憶部13に記憶された複数の解析データを用いた機械学習により、解析データへの分類設定を行うテキスト分類モデルM2を生成する。すなわち、学習処理手段112cは、過去における複数の解析データを入力とする機械学習により、解析データの性質の類似性に基づくグルーピングを行うテキスト分類モデルM2を構築する。学習処理手段112cは、上述した学習処理手段12cと同様に、テキスト分類モデルM2の更新処理を実行する。学習処理手段112cは、実施の形態1の学習処理手段12cと同様、DNNを用いた学習などの種々の機械学習の手法により、もしくは複数の機械学習を組み合わせた手法により、テキスト分類モデルM2を生成することができる。

40

【0055】

50

分類処理手段 1 1 2 d は、解析データをテキスト分類モデル M 2 の入力とし、入力とした解析データを性質の類似性に基づく複数のグループのうちの何れかに分類するものである。すなわち、分類処理手段 1 1 2 d は、分類対象である複数の解析データを順次テキスト分類モデル M 2 の入力とする。そして、分類処理手段 1 1 2 d は、入力とした解析データを複数のグループのうちの何れかに分類すると共に、分類先のグループの分類タグを付して分類データを生成する。本実施の形態 2 の分類データは「テキスト分類データ」とも称する。

【 0 0 5 6 】

本実施の形態 2 の分類用記憶部 1 3 は、分類処理プログラム P 2 を含む制御部 1 1 2 の動作プログラムの他、ページ情報としての解析データの分類処理に関連する種々のデータを記憶する。また、分類用記憶部 1 3 には、解析データを分類するためのテキスト分類モデル M 2 が格納される。分類処理プログラム P 2 は、コンピュータとしての制御部 1 1 2 及び分類用記憶部 1 3 を、収集処理手段 1 1 2 a、前処理手段 1 1 2 b、学習処理手段 1 1 2 c、及び分類処理手段 1 1 2 d として機能させるためのプログラムである。他の構成及び代替構成については、前述した実施の形態 1 と同様である。また、分類装置 1 1 0 による分類処理方法のうち、学習処理手段 1 1 2 c が行う学習処理工程、及び分類処理手段 1 1 2 d が行う分類処理工程は、図 5 に基づく説明等と同様である。

【 0 0 5 7 】

以上のように、本実施の形態 2 の分類装置 1 1 0 は、ページ情報としての解析データをテキスト分類モデル M 2 の入力とし、入力とした解析データを複数のグループのうちの何れかに分類する分類処理手段 1 1 2 d を有している。よって、分析対象のページに紐づくページ情報のグルーピングを自動的に効率よく行うことができるため、処理の迅速化及び効率化を図ることができる。また、本実施の形態 2 のページ情報は、ウェブページ内のテキストの情報であるテキストデータであり、テキストデータからは、アドレス情報からよりも、ウェブページのグルーピングに有用な情報が多く得られる。そのため、ウェブページのグルーピングをより高精度に行うことができる。他の効果等については、前述した実施の形態 1 と同様である。

【 0 0 5 8 】

ところで、図 7 に基づく上記の説明では、制御部 1 1 2 が前処理手段 1 1 2 b を有する例を示したが、制御部 1 1 2 は、前処理手段 1 1 2 b を有しない構成としてもよい。学習処理手段 1 1 2 c は、過去における複数のテキストデータを入力とする機械学習によりテキスト分類モデル M 2 を構築するようにしてもよい。この場合、学習処理手段 1 1 2 c は、例えばキャラクタレベル C N N を用いた機械学習により、テキストデータの性質の類似性に基づくグルーピングを行うテキスト分類モデル M 2 を生成する。したがって、分類処理手段 1 1 2 d は、分類対象である複数のテキストデータを順次テキスト分類モデル M 2 の入力とする。そして、分類処理手段 1 1 2 d は、入力としたテキストデータを複数のグループのうちの何れかに分類すると共に、分類先のグループの分類タグを付して分類データを生成する。

【 0 0 5 9 】

また、前処理手段 1 1 2 b は、テキストデータ中から不要データを削除して解析データを生成するものであってもよい。不要データとは、例えば数値又は改行などの、テキスト分類モデル M 2 の訓練及び該モデルでの予測に必要な文字等である。この場合、学習処理手段 1 1 2 c は、過去における複数の解析データを入力とする機械学習によりテキスト分類モデル M 2 を構築するものとなる。分類処理手段 1 1 2 d は、分類対象である複数の解析データを順次テキスト分類モデル M 2 の入力とし、入力とした解析データを複数のグループのうちの何れかに分類して分類先のグループの分類タグを付して分類データを生成する。

【 0 0 6 0 】

実施の形態 3 .

図 6、図 8～図 1 0 を参照して、本実施の形態 3 におけるデータ分析システム 3 0 0 及

10

20

30

40

50

びその周辺機器等の構成例について説明する。上述した実施の形態 1 及び 2 と同等の構成及び構成部材については同一の符号を付して説明は省略する。

【0061】

本実施の形態 3 のクローリングシステム 60 は、サイト内のページをクロールすることにより、該ページ内の各オブジェクトの情報を含む表示関連データを取得する機能を有している。クローリングシステム 60 は、サイトが複数のページで構成されている場合、各ページそれぞれの表示関連データを取得する。ここで、オブジェクトとは、コンピュータ上で操作や処理の対象となる何らかの実体である。オブジェクトには、表示画面上にアイコンなどの形で表示されるデータ集合や操作要素、及び表示画面上に配置される個々の図形などが含まれる。例えば、オブジェクトは、Document Object Model (DOM) の要素であってもよい。

10

【0062】

図 8 に示すように、本実施の形態 3 における分類装置 210 の制御部 212 は、収集処理手段 212a と、前処理手段 212b と、学習処理手段 212c と、分類処理手段 212d と、を有している。収集処理手段 212a は、サイトにおけるページの内容物の情報であって、該ページの各オブジェクトの情報を含む表示関連データを、逐次又は定期的にクローリングシステム 60 から収集し、収集した表示関連データを分類用記憶部 13 に記憶させる。

【0063】

前処理手段 212b は、サイトにおけるページの各オブジェクトの情報を含む表示関連データから、オブジェクトごとに、その種別を示す種別情報と、その位置及びサイズを示す配置情報との組み合わせによる識別情報を生成する。すなわち、前処理手段 212b は、表示関連データが有する各オブジェクトのそれぞれについて、種別情報と配置情報とを組み合わせで識別情報を生成する。例えば、ウェブページのオブジェクトに関し、オブジェクトの種別が 26 種類存在し、位置及びサイズを示す配置情報が 1024 種類存在する場合、オブジェクトの識別情報は、26624 (26×1024) 種類存在することになる。もっとも、オブジェクトの種別は、27 種類以上もしくは 25 種類以下に設定されてもよく、配置情報は、1025 種類以上もしくは 1023 種類以下に設定されてもよい。

20

【0064】

そして、前処理手段 212b は、生成したオブジェクトごとの識別情報に基づき、ページ情報としてのフレーム特徴量（ワイヤフレーム特徴量）を生成する。一例として、前処理手段 212b は、生成した各識別情報を BoW (Bag-of-Words) などのベクトル表現に変換し、例えばトピックモデルの一種である LDA (Latent Dirichlet Allocation) を用いて変換後のデータの次元削減を行うことにより、フレーム特徴量を生成する。もっとも、前処理手段 212b は、上記に限らず、ベクトル表現に変換したデータの次元削減の手法として、例えば、オートエンコーダ (Autoencoder) などの教師なし学習を採り入れたものであってもよい。

30

【0065】

ここで、図 9 及び図 10 を用いて、フレーム特徴量の生成処理を概念的に説明する。例えば、図 9 の表示画面 85x では、背景としての写真のオブジェクト 91a と、複数の写真がスライドショー形式で切り替わる設定のオブジェクト 91b と、リンクが貼られている写真のオブジェクト 91c と、を例示している。また、表示画面 85x では、単純文字のオブジェクト 92 と、リンクが貼られている文字又はロゴのオブジェクト 93 と、リンクが貼られている文字が複数並べられたオブジェクト 94 と、単純文字が複数並べられたオブジェクト 95 と、を例示している。加えて、表示画面 85x では、複数の文字などを表形式で整理したオブジェクト 96 と、文字を含むバーのオブジェクト 97 と、を例示している。

40

【0066】

図 9 では、写真のオブジェクトに斜線を施し、リンクが貼られているオブジェクトを矩形で示し、リンクが貼られているオブジェクトを丸みを帯びた矩形で示している。オブジ

50

エクト 94 は、いわゆるメニューバーなどである。図 9 では、オブジェクト 96 の中で色彩が付されているものをオブジェクト 96 a としている。オブジェクトには、上記以外にも様々な種別が存在し、オブジェクトの種別は、さらに細かく設定してもよく、もっと大まかに設定してもよい。

【0067】

図 9 のようなウェブページについて、各オブジェクトの種別に対応するパーツ情報だけを画像として抽出すると、図 10 に例示するようなワイヤフレーム画像となる。パーツ情報は、ウェブページのソースコードに含まれる、少なくとも 1 つのタグを含んだ情報であり、オブジェクトの種別を示す種別情報に対応する。すなわち、前処理手段 212 b は、ウェブページをワイヤフレーム画像に変換した後、該ワイヤフレーム画像から、固定長ベクトルの埋め込み表現（ベクトル表現）であるフレーム特徴量をページ情報として生成するものである。

【0068】

学習処理手段 212 c は、分類用記憶部 13 に記憶された複数のフレーム特徴量を用いた機械学習により、フレーム特徴量への分類設定を行うページ分類モデル M3 を生成する。すなわち、学習処理手段 212 c は、過去における複数のフレーム特徴量を入力とする機械学習により、フレーム特徴量の性質の類似性に基づくグルーピングを行うページ分類モデル M3 を構築する。学習処理手段 212 c は、上述した学習処理手段 12 c と同様に、ページ分類モデル M3 の更新処理を実行する。学習処理手段 212 c は、実施の形態 1 の学習処理手段 12 c と同様、DNN を用いた学習などの種々の機械学習の手法により、もしくは複数の機械学習を組み合わせた手法により、ページ分類モデル M3 を生成することができる。

【0069】

分類処理手段 212 d は、フレーム特徴量を、過去における複数のフレーム特徴量を用いた機械学習により生成されたページ分類モデル M3 の入力とし、入力としたフレーム特徴量を、性質の類似性に基づく複数のグループのうちの何れかに分類するものである。すなわち、分類処理手段 212 d は、分類対象である複数のフレーム特徴量を順次ページ分類モデル M3 の入力とする。そして、分類処理手段 212 d は、入力としたフレーム特徴量を複数のグループのうちの何れかに分類すると共に、分類先のグループの分類タグを付して分類データを生成する。本実施の形態 3 の分類データは「ページ分類データ」とも称する。

【0070】

本実施の形態 3 の分類用記憶部 13 は、分類処理プログラム P3 を含む制御部 212 の動作プログラムの他、ページ情報としてのフレーム特徴量の分類処理に関連する種々のデータを記憶する。また、分類用記憶部 13 には、フレーム特徴量を分類するためのページ分類モデル M3 が格納される。分類処理プログラム P3 は、コンピュータとしての制御部 212 及び分類用記憶部 13 を、収集処理手段 212 a、前処理手段 212 b、学習処理手段 212 c、及び分類処理手段 212 d として機能させるためのプログラムである。他の構成及び代替構成については、上述した実施の形態 1 及び 2 と同様である。また、分類装置 210 による分類処理方法のうち、学習処理手段 212 c が行う学習処理工程、及び分類処理手段 212 d が行う分類処理工程は、図 5 に基づく説明等と同様である。

【0071】

以上のように、本実施の形態 3 の分類装置 210 は、ページ情報としてのフレーム特徴量をページ分類モデル M3 の入力とし、入力としたフレーム特徴量を複数のグループのうちの何れかに分類する分類処理手段 212 d を有している。よって、分析対象のページに紐付くページ情報のグルーピングを自動的に効率よく行うことができるため、処理の迅速化及び効率化を図ることができる。

【0072】

ところで、ウェブページの内容物を示す表示関連データからは、アドレス情報からよりも、ウェブページのグルーピングに有用な情報が多く得られる。しかしながら、一般にウ

10

20

30

40

50

ウェブページは可変長であることから、表示関連データをそのままグルーピングに用いるのは困難である。なぜなら、画像を入力とする機械学習手法は、一般に、入力画像サイズが固定されており、可変長の画像に対応しているものは少ないためである。併せて、画像を無理に固定のサイズに縮小すると、分類処理において本来的に必要な情報を失い、分類処理の性能が悪化する恐れがあるためでもある。つまり、こうした事情から、これまでのページグルーピングには、表示関連データが用いられていない。

【0073】

この点、本実施の形態3の前処理手段212bは、分析対象である可変長のウェブページに対応する表示関連データから、オブジェクトごとの、種別情報と配置情報との組み合わせによる識別情報を生成する。そして、前処理手段212bは、生成したオブジェクトごとの識別情報に基づいてフレーム特徴量を生成する。すなわち、前処理手段212bは、機械学習の入力データとして使用できるように、表示関連データを、固定長のベクトル表現に変換する機能を有している。つまり、前処理手段212bは、可変長のウェブページに対応する表示関連データに対し、ベクトル表現への次元圧縮（次元削減）を施すことにより、フレーム特徴量を生成するようになっている。そのため、ウェブページのグルーピングの更なる高精度化を図ることができる。

【0074】

実施の形態4.

図11を参照して、本実施の形態4におけるデータ分析システム400及びその周辺機器等の構成例について説明する。上述した実施の形態1～3と同等の構成及び構成部材については同一の符号を付して説明は省略する。

【0075】

図11に示すように、本実施の形態4における分類装置310の制御部312は、収集処理手段312aと、前処理手段312bと、学習処理手段312cと、分類処理手段312dと、を有している。収集処理手段312aは、ページ情報としてのアドレス情報と、テキストデータと、表示関連データとを、逐次又は定期的に収集し、収集した各情報を分類用記憶部13に記憶させる。

【0076】

前処理手段312bは、テキスト前処理手段B2と、ページ前処理手段B3と、を有している。テキスト前処理手段B2は、上述した実施の形態2の前処理手段112bと同様に機能する。すなわち、テキスト前処理手段B2は、テキストデータに解析処理を施して、ページ情報としての解析データを生成するものである。ページ前処理手段B3は、上述した実施の形態3の前処理手段212bと同様に機能する。すなわち、ページ前処理手段B3は、分類対象のページに対応する表示関連データから、オブジェクトごとの、種別情報と配置情報との組み合わせによる識別情報を生成し、生成したオブジェクトごとの識別情報に基づいて、ページ情報としてのフレーム特徴量を生成する。

【0077】

学習処理手段312cは、アドレス学習手段C1と、テキスト学習手段C2と、ページ学習手段C3と、統括学習手段Cxと、を有している。アドレス学習手段C1は、上述した実施の形態1の学習処理手段12cと同様に学習済みモデルを生成し、更新する。すなわち、アドレス学習手段C1は、過去における複数のアドレス情報を入力とする機械学習により、入力としたアドレス情報からアドレス特徴量を抽出して出力するアドレス分類モデルM10を構築する。アドレス特徴量は、アドレス情報に分類タグを付したアドレス分類データであってもよく、アドレス分類データとは異なる態様の情報であってもよい。

【0078】

アドレス学習手段C1は、例えば、DNNの隠れ層（中間層）の出力ベクトルを埋め込み表現としたものを、アドレス分類データとは異なる態様のアドレス特徴量として抽出し、これを統括学習手段Cxに利用させてもよい。この場合、アドレス学習手段C1は、アドレス分類モデルM10におけるDNNの隠れ層の、出力層の1つ手前の層などの出力ベクトルを埋め込み表現としたアドレス特徴量を抽出する。また、アドレス分類モデルM1

0におけるDNNの出力層に、活性化関数の一種であるSoftMax関数を採用している場合は、SoftMax関数によってグループごとの確率が出力される。よって、この場合、アドレス学習手段C1は、各グループそれぞれの確率ベクトルを、アドレス分類データとは異なる態様のアドレス特徴量として抽出してもよい。

【0079】

さらに、アドレス学習手段C1は、キャラクタレベルCNNを用いる構成の場合、例えば、キャラクタレベルCNNの隠れ層の出力ベクトルを埋め込み表現としたものを、アドレス分類データとは異なる態様のアドレス特徴量として抽出してもよい。また、アドレス分類モデルM10におけるキャラクタレベルCNNの出力層にSoftMax関数を採用している場合、アドレス学習手段C1は、各グループそれぞれの確率ベクトルを、アドレス分類データとは異なる態様のアドレス特徴量として抽出してもよい。

10

【0080】

テキスト学習手段C2は、上述した実施の形態2の学習処理手段112cと同様に学習済みモデルを生成し、更新する。すなわち、テキスト学習手段C2は、過去における複数の解析データを入力とする機械学習により、入力とした解析データからテキスト特徴量を抽出して出力するテキスト分類モデルM20を構築する。テキスト特徴量は、解析データに分類タグを付したテキスト分類データであってもよく、テキスト分類データが生成される途中の埋め込み表現（ベクトル表現）などの中間的な表現の情報であってもよい。

【0081】

ページ学習手段C3は、上述した実施の形態3の学習処理手段212cと同様に学習済みモデルを生成し、更新する。すなわち、ページ学習手段C3は、過去における複数のフレーム特徴量を入力とする機械学習により、入力としたフレーム特徴量からページ特徴量を抽出して出力するページ分類モデルM30を構築する。ページ特徴量は、フレーム特徴量に分類タグを付したページ分類データであってもよく、ページ分類データが生成される途中の埋め込み表現（ベクトル表現）などの中間的な表現の情報であってもよい。

20

【0082】

統括学習手段Cxは、アドレス分類手段D1により生成されたアドレス特徴量と、テキスト分類手段D2により生成されたテキスト特徴量と、ページ分類手段D3により生成されたページ特徴量と、を入力とする機械学習により、入力とした各情報に対応するウェブページを分類する統括分類モデルM4を生成するものである。本実施の形態4において、統括学習手段Cxは、GBDTを用いた機械学習により統括分類モデルM4を生成するようになっている。

30

【0083】

分類処理手段312dは、アドレス分類手段D1と、テキスト分類手段D2と、ページ分類手段D3と、統括分類手段Dxと、を有している。アドレス分類手段D1は、上述した実施の形態1の分類処理手段12dと同様に分類処理を実行する。すなわち、アドレス分類手段D1は、分類対象のページに対応するアドレス情報をアドレス分類モデルM10の入力として、アドレス特徴量を抽出する。

【0084】

テキスト分類手段D2は、上述した実施の形態2の分類処理手段112dと同様に分類処理を実行する。すなわち、テキスト分類手段D2は、分類対象のページに対応する解析データをテキスト分類モデルM20の入力として、テキスト特徴量を抽出する。ページ分類手段D3は、上述した実施の形態3の分類処理手段212dと同様に分類処理を実行する。すなわち、ページ分類手段D3は、分類対象のページに対応するフレーム特徴量をページ分類モデルM30の入力としてページ特徴量を抽出する。

40

【0085】

統括分類手段Dxは、ウェブページごとのアドレス特徴量、テキスト特徴量、及びページ特徴量を用いて、各ウェブページのグルーピングを行うものである。より具体的に、統括分類手段Dxは、分類対象であるウェブページに紐付けられたアドレス特徴量、テキスト特徴量、及びページ特徴量を統括分類モデルM4の入力とし、該ウェブページを、性質

50

の類似性に基づく複数のグループのうちの何れかに分類するものである。すなわち、統括分類手段D xは、分類対象である複数のウェブページに紐付けられた各特徴量を順次統括分類モデルM 4の入力とする。そして、統括分類手段D xは、入力とした各特徴量に紐付くウェブページを複数のグループのうちの何れかに分類すると共に、分類先のグループの分類タグを該ウェブページを示す情報に付して分類データを生成する。本実施の形態4の分類データは「統括分類データ」とも称する。

【0086】

ところで、テキスト分類手段D 2及びページ分類手段D 3は、クローリング及び前処理を経たデータを用いるため、テキスト分類手段D 2又はページ分類手段D 3による1ページ当たりの処理時間は、アドレス分類手段D 1の1ページ当たりの処理時間よりも長くなる。また、アドレス分類手段D 1とテキスト分類手段D 2とページ分類手段D 3とが行う処理内容は、それぞれ異なっているため、同一のページに対する分類処理に要する時間もそれぞれ異なる。1ページ当たりの分類処理に要する時間は、概ね、アドレス分類手段D 1、テキスト分類手段D 2、ページ分類手段D 3の順に長くなる。したがって、アドレス分類手段D 1の処理速度に合わせて全体的な分類処理を行うと、テキスト分類手段D 2及びページ分類手段D 3での遅延分のデータが、未処理の欠損データ（欠損値）となる。

【0087】

そこで、本実施の形態4では、GBDTに基づく統括分類モデルM 4が弱学習器として内包する複数の決定木の中に、「欠損データであるか否か？」という条件分岐を設定している。そのため、統括分類手段D xは、テキスト分類手段D 2及びページ分類手段D 3における欠損データも、統括分類モデルM 4の入力とすることができる。

【0088】

本実施の形態4の分類用記憶部13は、分類処理プログラムP 4を含む制御部312の動作プログラムの他、ページ情報としてのフレーム特徴量の分類処理に関連する種々のデータを記憶する。また、分類用記憶部13には、アドレス情報からアドレス特徴量を抽出するためのアドレス分類モデルM 10と、解析データ等からテキスト特徴量を抽出するためのテキスト分類モデルM 20と、フレーム特徴量からページ特徴量を抽出するためのページ分類モデルM 30と、が格納される。さらに、分類用記憶部13には、アドレス特徴量、テキスト特徴量、及びページ特徴量を入力として分類データを生成する統括分類モデルM 4が格納される。分類処理プログラムP 4は、コンピュータとしての制御部312及び分類用記憶部13を、収集処理手段312a、前処理手段312b、学習処理手段312c、及び分類処理手段312dとして機能させるためのプログラムである。他の構成及び代替構成については、上述した実施の形態1～3と同様である。また、分類装置310による分類処理方法のうち、学習処理手段312cが行う学習処理工程、及び分類処理手段312dが行う分類処理工程の基本的な流れは、図5に基づく説明等と同様である。

【0089】

次に、図12を参照して、本実施の形態4における分類処理方法の各工程について説明する。ここでは、学習処理手段312cによって各分類モデルを生成する学習処理工程については省略する。なお、図12における横並びの各工程は、並行して行われる。

【0090】

〔収集処理工程〕

図12に示すように、収集処理手段312aは、外部から逐次又は定期的にアドレス情報、テキストデータ、及び表示関連データを収集し、収集した各データを分類用記憶部13に記憶させる。その際、収集処理手段312aは、同じウェブページに対応するアドレス情報とテキストデータと表示関連データとを紐付けて分類用記憶部13に記憶させる。

【0091】

〔前処理工程〕

前処理手段312bは、テキストデータに対する形態素解析をもとに解析データを生成する。あるいは前処理手段312bは、テキストデータ中から不要データを削除して解析データを生成する（テキスト前処理工程）。

前処理手段 3 1 2 b は、分類対象のページに対応する表示関連データから、オブジェクトごとの、種別情報と配置情報との組み合わせによる識別情報を生成し、生成したオブジェクトごとの識別情報に基づいてフレーム特徴量を生成する（ページ前処理工程）。

【0092】

〔分類処理工程〕

分類処理手段 3 1 2 d は、アドレス分類手段 D 1 により、分類対象のページに対応するアドレス情報をアドレス分類モデル M 1 0 の入力としてアドレス特徴量を抽出する（アドレス分類工程）。

分類処理手段 3 1 2 d は、テキスト分類手段 D 2 により、分類対象のページに対応する解析データをテキスト分類モデル M 2 0 の入力としてテキスト特徴量を抽出する（テキスト分類工程）。

10

分類処理手段 3 1 2 d は、ページ分類手段 D 3 により、分類対象のページに対応するフレーム特徴量をページ分類モデル M 3 0 の入力としてページ特徴量を抽出する（ページ分類工程）。

【0093】

分類処理手段 3 1 2 d は、分類対象である複数のウェブページに紐付けられたアドレス特徴量、テキスト特徴量、及びページ特徴量を、順次統括分類モデル M 4 の入力とする。そして、分類処理手段 3 1 2 d は、入力とした各特徴量に紐付くウェブページを複数のグループのうちの何れかに分類すると共に、分類先のグループの分類タグを該ウェブページを示す情報に付して分類データを生成する（統括分類工程）。

20

【0094】

以上のように、本実施の形態 4 の分類装置 3 1 0 によっても、分析対象のウェブページのグルーピングを自動的に効率よく行うことができるため、処理の迅速化及び効率化を図ることができる。また、分類装置 3 1 0 は、アドレス情報をアドレス分類モデル M 1 0 の入力としてアドレス特徴量を抽出するアドレス分類手段 D 1 と、解析データをテキスト分類モデル M 2 0 の入力としてテキスト特徴量を抽出するテキスト分類手段 D 2 と、フレーム特徴量をページ分類モデル M 3 0 の入力としてページ特徴量を抽出するページ分類手段 D 3 と、を有している。また、分類装置 3 1 0 は、ウェブページごとのアドレス特徴量、テキスト特徴量、及びページ特徴量を用いて、各ウェブページのグルーピングを行う統括分類手段 D x を有している。すなわち、分類装置 3 1 0 は、アドレス情報、テキストデータ、及び表示関連データという 3 つの情報を用いた 2 段階の分類処理により、1 つの閲覧データを用いる構成では得られない情報を機械学習及び分類処理に反映させることができる。そのため、ウェブページのグルーピングの精度向上を図ることができる。

30

【0095】

上記の説明では、統括学習手段 C x が G B D T を用いた機械学習により統括分類モデル M 4 を生成する例を示したが、これに限定されない。統括学習手段 C x は、D N N を用いた学習などの種々の機械学習の手法により、もしくは複数の機械学習を組み合わせた手法により、統括分類モデル M 4 を生成してもよい。

【0096】

ところで、分類装置 3 1 0 は、テキスト前処理手段 B 2 を有しない構成としてもよい。この場合、テキスト学習手段 C 2 は、過去における複数のテキストデータを入力とする機械学習により、入力としたテキストデータからテキスト特徴量を抽出するテキスト分類モデル M 2 0 を構築する。すなわち、テキスト分類手段 D 2 は、分類対象のページに対応するテキストデータをテキスト分類モデル M 2 0 の入力としてテキスト特徴量を生成する。

40

【0097】

<変形例 4 a>

図 1 3 を参照して、実施の形態 4 の変形例 4 a におけるデータ分析システム 4 0 0 A の構成例について説明する。本変形例 4 a における制御部 3 1 2 は、収集処理手段 3 1 2 a と、テキスト前処理手段 B 2 と、学習処理手段 3 1 2 c と、分類処理手段 3 1 2 d と、を有している。本変形例 4 a の学習処理手段 3 1 2 c は、アドレス学習手段 C 1 と、テキス

50

ト学習手段C 2と、統括学習手段C x aと、を有している。本変形例4 aの分類処理手段3 1 2 dは、アドレス分類手段D 1と、テキスト分類手段D 2と、統括分類手段D x aと、を有している。本変形例4 aの収集処理手段3 1 2 aは、ページ情報としてのアドレス情報と、テキストデータとを、逐次又は定期的に収集し、収集した各情報を分類用記憶部1 3に記憶させる。

【0 0 9 8】

統括学習手段C x aは、アドレス分類手段D 1により生成されたアドレス特徴量と、テキスト分類手段D 2により生成されたテキスト特徴量とを入力とする機械学習により、入力とした各特徴量に対応するウェブページを分類する統括分類モデルM 4 1を生成する。統括分類手段D x aは、分類対象である複数のウェブページに紐付けられたアドレス特徴量及びテキスト特徴量を、順次統括分類モデルM 4 1の入力とする。そして、統括分類手段D x aは、入力とした各特徴量に紐付くウェブページを複数のグループのうちの何れかに分類すると共に、分類先のグループの分類タグを該ウェブページを示す情報に付して分類データを生成する。

【0 0 9 9】

以上のように、本変形例4 aの分類装置3 1 0は、アドレス情報及びテキストデータという2つの情報を用いた2段階の分類処理により、1つの閲覧データを用いる構成では得られない情報を機械学習及び分類処理に反映させることができる。したがって、ウェブページのグルーピングの精度向上を図ることができる。ところで、分類装置3 1 0は、テキスト前処理手段B 2を有しない構成としてもよい。この場合、テキスト学習手段C 2は、過去における複数のテキストデータを入力とする機械学習により、入力としたテキストデータからテキスト特徴量を抽出するテキスト分類モデルM 2 0を構築する。

【0 1 0 0】

<変形例4 b>

図1 4を参照して、実施の形態4の変形例4 bにおけるデータ分析システム4 0 0 Bの構成例について説明する。本変形例4 bにおける制御部3 1 2は、収集処理手段3 1 2 aと、ページ前処理手段B 3と、学習処理手段3 1 2 cと、分類処理手段3 1 2 dと、を有している。本変形例4 bの学習処理手段3 1 2 cは、アドレス学習手段C 1と、ページ学習手段C 3と、統括学習手段C x bと、を有している。本変形例4 bの分類処理手段3 1 2 dは、アドレス分類手段D 1と、ページ分類手段D 3と、統括分類手段D x bと、を有している。本変形例4 bの収集処理手段3 1 2 aは、ページ情報としてのアドレス情報と、表示関連データとを、逐次又は定期的に収集し、収集した各情報を分類用記憶部1 3に記憶させる。

【0 1 0 1】

統括学習手段C x bは、アドレス分類手段D 1により生成されたアドレス特徴量と、ページ分類手段D 3により生成されたページ特徴量とを入力とする機械学習により、入力とした各特徴量に対応するウェブページを分類する統括分類モデルM 4 2を生成する。統括分類手段D x bは、分類対象である複数のウェブページに紐付けられたアドレス特徴量及びページ特徴量を、順次統括分類モデルM 4 2の入力とする。そして、統括分類手段D x bは、入力とした各特徴量に紐付くウェブページを複数のグループのうちの何れかに分類すると共に、分類先のグループの分類タグを該ウェブページを示す情報に付して分類データを生成する。

【0 1 0 2】

以上のように、本変形例4 bの分類装置3 1 0は、アドレス情報及び表示関連データという2つの情報を用いた2段階の分類処理により、1つの閲覧データを用いる構成では得られない情報を機械学習及び分類処理に反映させることができる。したがって、ウェブページのグルーピングの精度向上を図ることができる。

【0 1 0 3】

<変形例4 c>

図1 5を参照して、実施の形態4の変形例4 cにおけるデータ分析システム4 0 0 Cの

構成例について説明する。本変形例 4 c における制御部 3 1 2 は、収集処理手段 3 1 2 a と、前処理手段 3 1 2 b と、学習処理手段 3 1 2 c と、分類処理手段 3 1 2 d と、を有している。本変形例 4 c の学習処理手段 3 1 2 c は、テキスト学習手段 C 2 と、ページ学習手段 C 3 と、統括学習手段 C x c と、を有している。本変形例 4 b の分類処理手段 3 1 2 d は、テキスト分類手段 D 2 と、ページ分類手段 D 3 と、統括分類手段 D x c と、を有している。本変形例 4 c の収集処理手段 3 1 2 a は、テキストデータと表示関連データとを逐次又は定期的に収集し、収集した各情報を分類用記憶部 1 3 に記憶させる。

【0104】

統括学習手段 C x c は、テキスト分類手段 D 2 により生成されたテキスト特徴量と、ページ分類手段 D 3 により生成されたページ特徴量とを入力とする機械学習により、入力とした各特徴量に対応するウェブページを分類する統括分類モデル M 4 3 を生成する。統括分類手段 D x c は、分類対象である複数のウェブページに紐付けられたテキスト特徴量及びページ特徴量を、順次統括分類モデル M 4 3 の入力とする。そして、統括分類手段 D x c は、入力とした各特徴量に紐付くウェブページを複数のグループのうちの何れかに分類すると共に、分類先のグループの分類タグを該ウェブページを示す情報に付して分類データを生成する。

【0105】

以上のように、本変形例 4 c の分類装置 3 1 0 は、テキストデータ及び表示関連データという 2 つの情報をを用いた 2 段階の分類処理により、1 つの閲覧データを用いる構成では得られない情報を機械学習及び分類処理に反映させることができる。したがって、ウェブページのグルーピングの精度向上を図ることができる。ところで、分類装置 3 1 0 は、テキスト前処理手段 B 2 を有しない構成としてもよい。この場合、テキスト学習手段 C 2 は、過去における複数のテキストデータを入力とする機械学習により、入力としたテキストデータからテキスト特徴量を抽出するテキスト分類モデル M 2 0 を構築する。

【0106】

実施の形態 5 .

図 1 6 を参照して、本実施の形態 5 におけるデータ分析システム 5 0 0 及びその周辺機器等の構成例について説明する。上述した実施の形態 1 ~ 4 と同等の構成及び構成部材については同一の符号を付して説明は省略する。

【0107】

図 1 6 に示すように、本実施の形態 5 における分類装置 4 1 0 の制御部 4 1 2 は、収集処理手段 3 1 2 a と、前処理手段 3 1 2 b と、学習処理手段 4 1 2 c と、分類処理手段 4 1 2 d と、を有している。

【0108】

学習処理手段 4 1 2 c は、過去における複数組の、ウェブページの位置を示すアドレス情報、該ウェブページに対応する解析データ、及び該ウェブページに対応するフレーム特徴量を入力とする機械学習により、入力とした各情報に紐付くウェブページを分類して分類データを生成する分類モデル M 5 を構築する。学習処理手段 4 1 2 c は、実施の形態 1 の学習処理手段 1 2 c と同様、DNN を用いた学習などの種々の機械学習の手法により、もしくは複数の機械学習を組み合わせた手法により、分類モデル M 5 を生成することができる。

【0109】

分類処理手段 4 1 2 d は、ウェブページの位置を示すアドレス情報、該ウェブページに対応する解析データ、及び該ウェブページに対応するフレーム特徴量を、ページ情報として分類モデル M 5 の入力とする。そして、分類処理手段 4 1 2 d は、分類モデル M 5 の入力としたページ情報に紐付くウェブページを複数のグループのうちの何れかに分類すると共に、分類先のグループの分類タグを該ウェブページを示す情報に付して分類データを生成する。

【0110】

本実施の形態 5 の分類用記憶部 1 3 は、分類処理プログラム P 5 を含む制御部 4 1 2 の

動作プログラムの他、ページ情報の分類処理に関連する種々のデータを記憶する。また、分類用記憶部 13 には、ページ情報を分類するための分類モデル M5 が格納される。分類処理プログラム P5 は、コンピュータとしての制御部 412 及び分類用記憶部 13 を、収集処理手段 312a、前処理手段 312b、学習処理手段 412c、及び分類処理手段 412d として機能させるためのプログラムである。他の構成及び代替構成については、上述した実施の形態 1～4 と同様である。また、分類装置 410 による分類処理方法のうち、学習処理手段 412c が行う学習処理工程、及び分類処理手段 412d が行う分類処理工程の基本的な流れは、図 5 に基づく説明等と同様である。

【0111】

以上のように、本実施の形態 5 の分類装置 410 によっても、分析対象のウェブページのグルーピングを自動的に効率よく行うことができるため、処理の迅速化及び効率化を図ることができる。また、分類処理手段 412d は、ウェブページごとのアドレス情報、解析データ、及びフレーム特徴量を分類モデル M5 の入力とし、入力とした各情報に紐付くウェブページを複数のグループのうちの何れかに分類する。すなわち、分類装置 410 によれば、アドレス情報、テキストデータ、及び表示関連データという 3 つの情報を用いた分類処理により、1 つの閲覧データを用いる構成では得られない情報を機械学習及び分類処理に反映させることができる。そのため、ウェブページのグルーピングの精度向上を図ることができる。

【0112】

ところで、分類装置 410 は、テキスト前処理手段 B2 を有しない構成としてもよい。この場合、学習処理手段 412c は、過去におけるウェブページごとのアドレス情報、テキストデータ、及びフレーム特徴量を入力とする機械学習により、入力とした各情報に紐付くウェブページを分類する分類モデル M5 を構築する。分類処理手段 412d は、ウェブページごとのアドレス情報、テキストデータ、及びフレーム特徴量を、ページ情報として分類モデル M5 の入力とし、入力とした各情報に紐付くウェブページを複数のグループのうちの何れかに分類する。

【0113】

また、収集処理手段 312a は、変形例 4c と同様、テキストデータと表示関連データとを逐次又は定期的に収集するものであってもよい。

かかる構成において、分類装置 410 がテキスト前処理手段 B2 を有する場合、学習処理手段 412c は、過去におけるウェブページごとの解析データ及びフレーム特徴量を入力とする機械学習により、入力とした各情報に紐付くウェブページを分類する分類モデル M5 を構築する。そして、分類処理手段 412d は、ウェブページごとの解析データ及びフレーム特徴量を、ページ情報として分類モデル M5 の入力とし、入力とした各情報に紐付くウェブページを複数のグループのうちの何れかに分類する。

かかる構成において、分類装置 410 がテキスト前処理手段 B2 を有しない場合、学習処理手段 412c は、過去におけるウェブページごとのテキストデータ及びフレーム特徴量を入力とする機械学習により、入力とした各情報に紐付くウェブページを分類する分類モデル M5 を構築する。分類処理手段 412d は、ウェブページごとのテキストデータ及びフレーム特徴量を、ページ情報として分類モデル M5 の入力とし、入力とした各情報に紐付くウェブページを複数のグループのうちの何れかに分類する。

【0114】

<変形例 5a>

図 17 を参照して、本実施の形態 5 の変形例 5a におけるデータ分析システム 500A の構成例について説明する。本変形例 5a の収集処理手段 312a は、変形例 4a と同様、アドレス情報とテキストデータとを逐次又は定期的に収集する。したがって、学習処理手段 412c は、過去における複数組の、ウェブページの位置を示すアドレス情報、及び該ウェブページに対応する解析データを入力とする機械学習により、入力とした各情報に紐付くウェブページを分類する分類モデル M5 を構築する。分類処理手段 412d は、ウェブページの位置を示すアドレス情報、及び該ウェブページに対応する解析データを、ペ

10

20

30

40

50

ージ情報として分類モデルM5の入力とし、入力としたページ情報に紐付くウェブページを複数のグループのうちの何れかに分類する。

【0115】

以上のように、本変形例5aの分類装置410は、アドレス情報及びテキストデータという2つの情報を用いた分類処理により、1つの閲覧データを用いる構成では得られない情報を機械学習及び分類処理に反映させることができる。したがって、ウェブページのグルーピングの精度向上を図ることができる。ところで、分類装置410は、テキスト前処理手段B2を有しない構成としてもよい。この場合、学習処理手段412cは、過去におけるウェブページごとのアドレス情報及びテキストデータを入力とする機械学習により、入力とした各情報に紐付くウェブページを分類する分類モデルM5を構築する。分類処理手段412dは、ウェブページごとのアドレス情報及びテキストデータを、ページ情報として分類モデルM5の入力とし、入力としたページ情報に紐付くウェブページを複数のグループのうちの何れかに分類する。

10

【0116】

<変形例5b>

図18を参照して、本実施の形態5の変形例5bにおけるデータ分析システム500Bの構成例について説明する。本変形例5bの収集処理手段312aは、変形例4bと同様、アドレス情報と表示関連データとを逐次又は定期的に収集する。したがって、学習処理手段412cは、過去における複数組の、ウェブページの位置を示すアドレス情報、及び該ウェブページに対応するフレーム特徴量を入力とする機械学習により、入力とした各情報に紐付くウェブページを分類する分類モデルM5を構築する。分類処理手段412dは、ウェブページごとのアドレス情報及びフレーム特徴量を、ページ情報として分類モデルM5の入力とし、入力としたページ情報に紐付くウェブページを複数のグループのうちの何れかに分類する。

20

【0117】

以上のように、本変形例5bの分類装置410は、アドレス情報及び表示関連データという2つの情報を用いた分類処理により、1つの閲覧データを用いる構成では得られない情報を機械学習及び分類処理に反映させることができる。したがって、ウェブページのグルーピングの精度向上を図ることができる。

【0118】

上述した各実施の形態は、分類装置、データ分析システム、分類処理プログラム、及び分類処理方法における好適な具体例であり、本発明の技術的範囲は、これらの態様に限定されるものではない。例えば、上記の説明では、収集処理手段がクロールシステム60からテキストデータ及び表示関連データを収集する例を示したが、これに限定されない。収集処理手段は、サイト内のページをクロールすることにより、該ページのテキストデータ及び表示関連データのうちの少なくとも一方を取得する機能を有していてもよい。

30

【符号の説明】

【0119】

10、110、210、310、410 分類装置、11 通信部、12、112、212、312、412 制御部、12a、112a、212a、312a 収集処理手段、12c、112c、212c、312c、412c 学習処理手段、12d、112d、212d、312d、412d 分類処理手段、13 分類用記憶部、20 分析装置、22 分析処理部、23 分析用記憶部、23p 分析処理プログラム、30 管理装置、31 管理通信部、32 管理制御部、32a 情報処理手段、32b 表示処理手段、33 管理記憶部、34、84 入力部、35、85 表示部、40 サービス提供システム、50 解析システム、60 クローリングシステム、80 端末装置、81 端末通信部、82 端末制御部、82a 情報処理手段、82b 表示処理手段、83 端末記憶部、85x 表示画面、91a、91b、91c、92、93、94、95、96、96a、97 オブジェクト、100、200、300、400、400A~400C、500、500A、500B データ分析システム、112b、212b、312b、412b 前処理手段

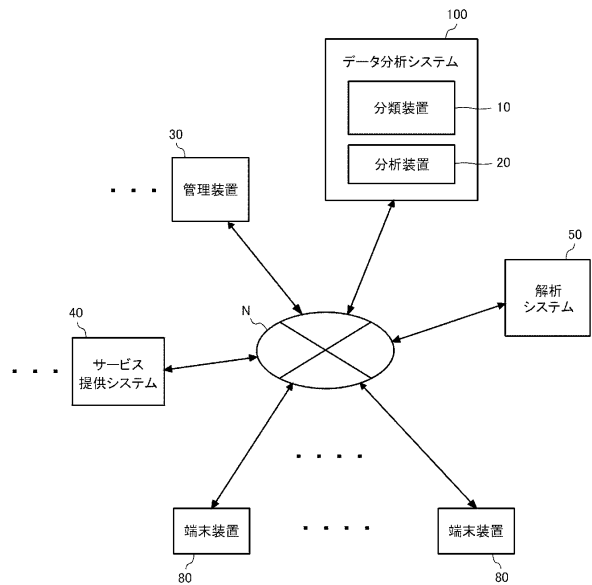
40

50

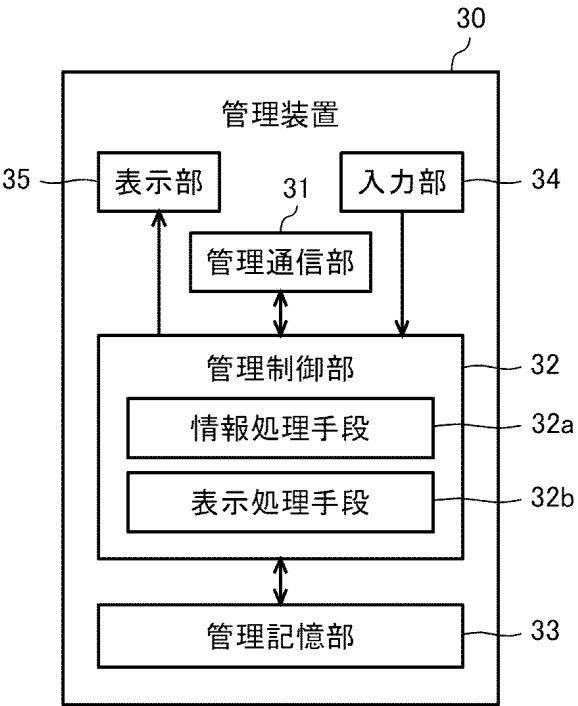
、B 2 テキスト前処理手段、B 3 ページ前処理手段、C 1 アドレス学習手段、C 2 テキスト学習手段、C 3 ページ学習手段、C x、C x a～C x c 統括学習手段、D 1 アドレス分類手段、D 2 テキスト分類手段、D 3 ページ分類手段、D x、D x a～D x c 統括分類手段、M 1、M 1 0 アドレス分類モデル、M 2、M 2 0 テキスト分類モデル、M 3、M 3 0 ページ分類モデル、M 4、M 4 1、M 4 2、M 4 3 統括分類モデル、M 5 分類モデル、N ネットワーク、P 1～P 5 分類処理プログラム。

【図面】

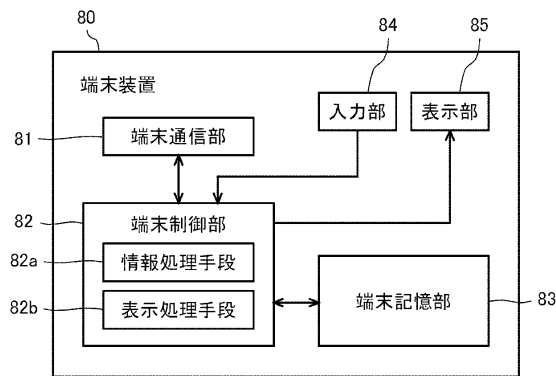
【図 1】



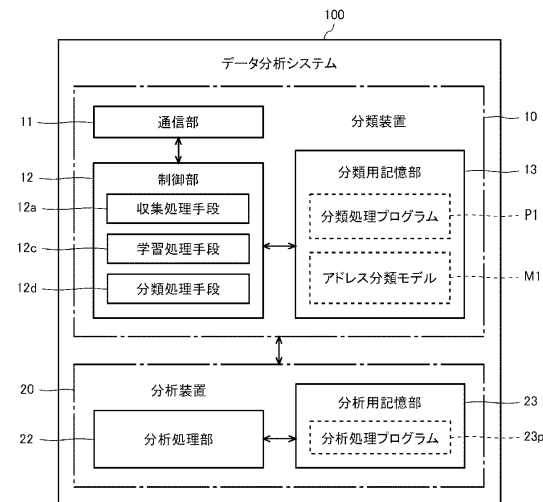
【図 2】



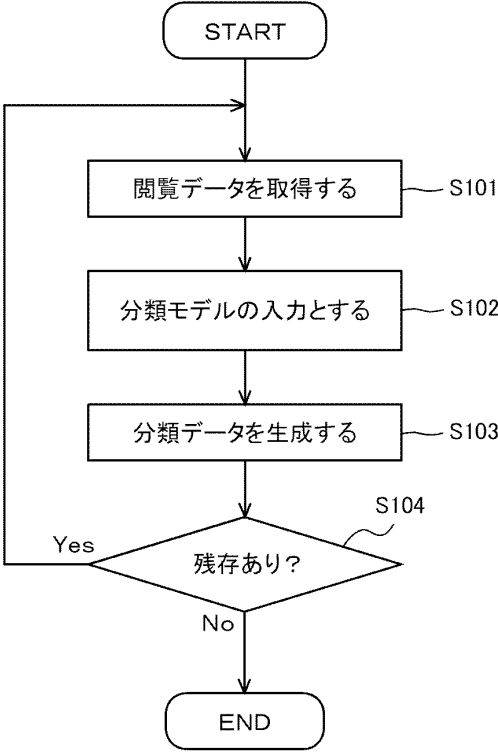
【図 3】



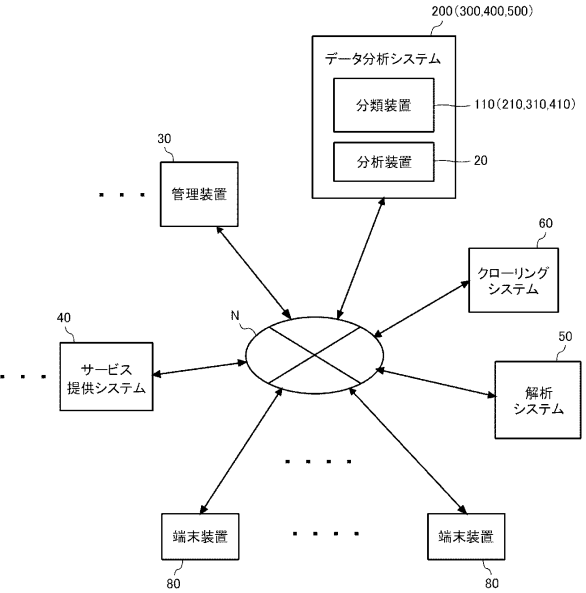
【図 4】



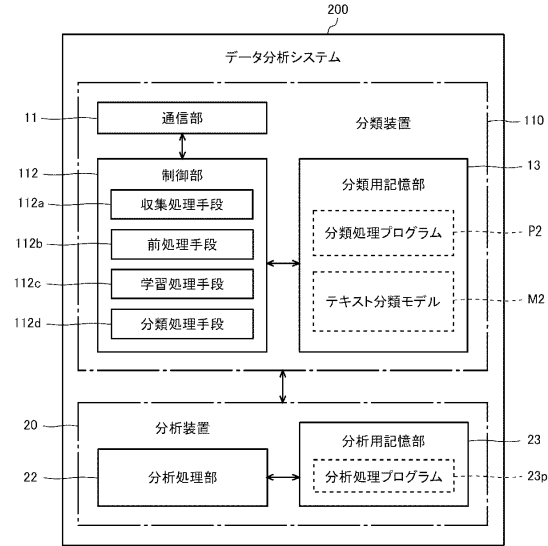
【図 5】



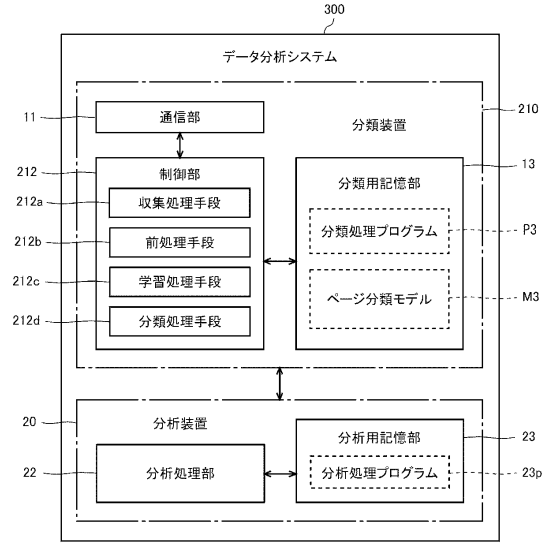
【図 6】



【図 7】



【図 8】



10

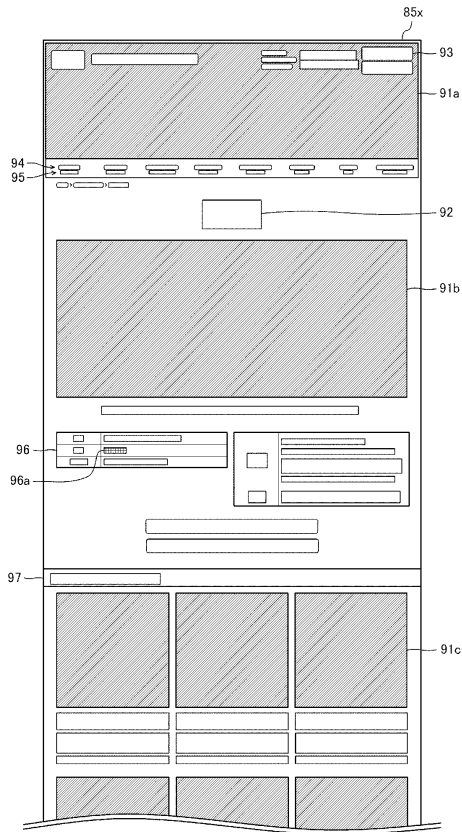
20

30

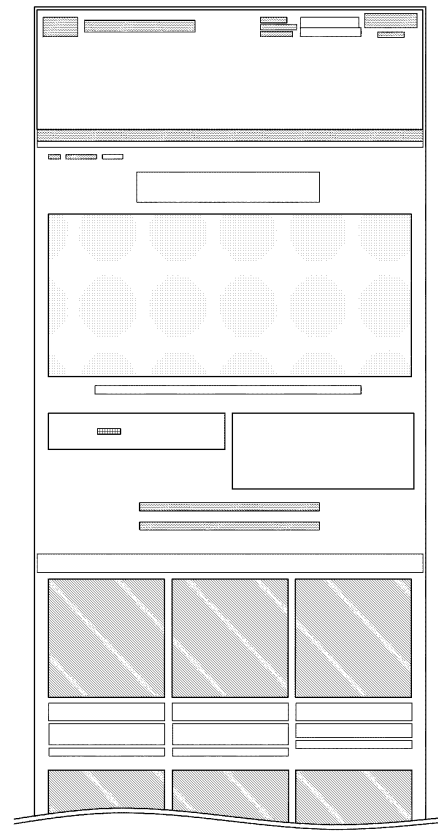
40

50

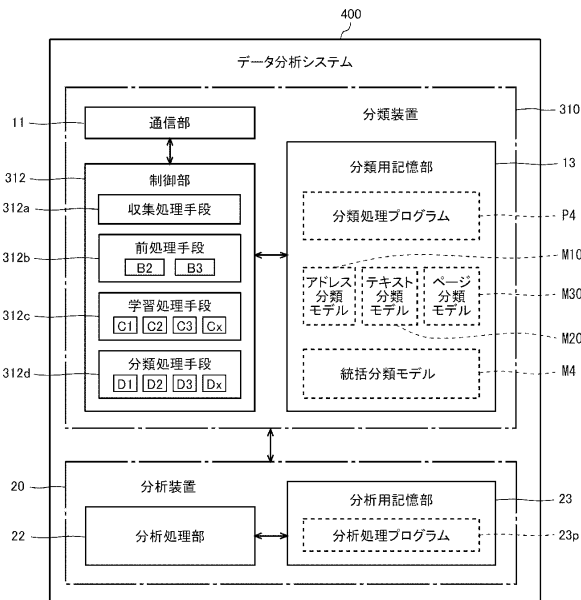
【図 9】



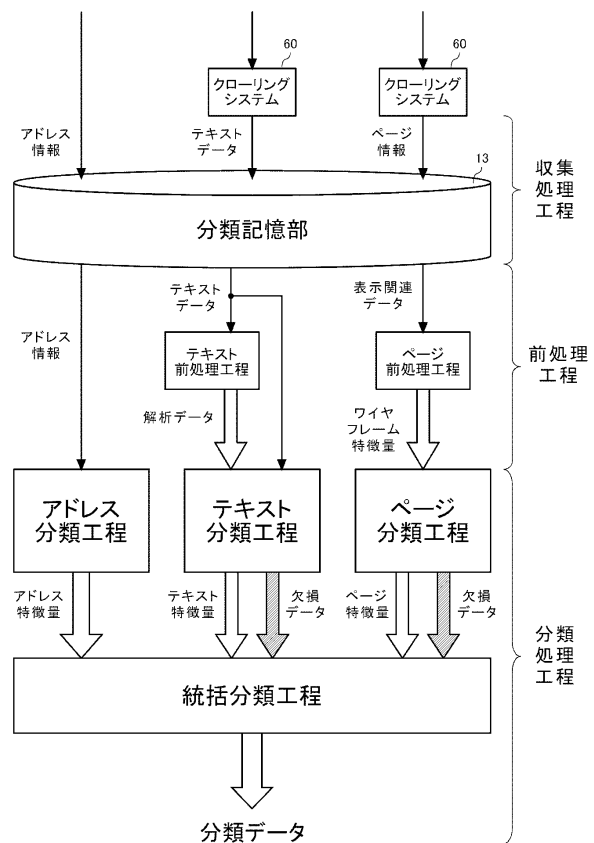
【図 10】



【図 11】



【図 12】



10

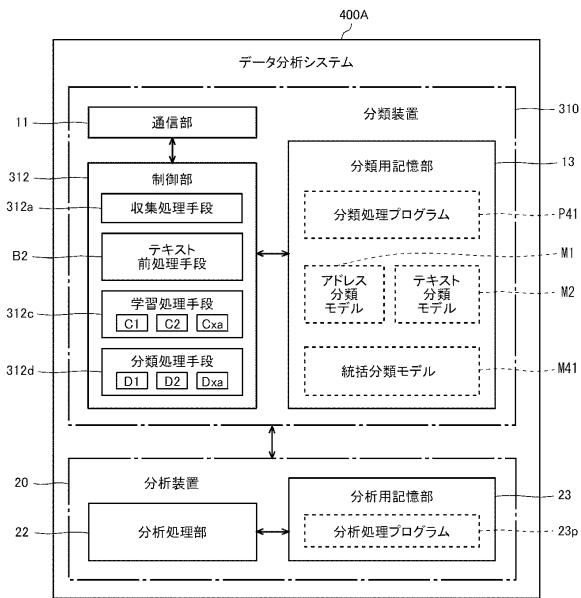
20

30

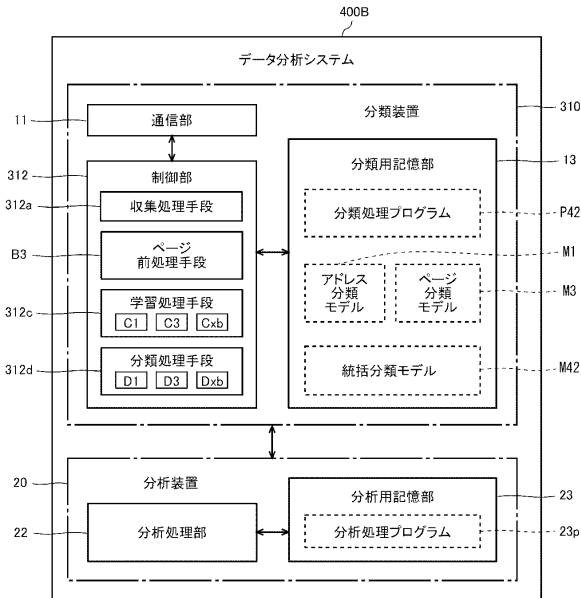
40

50

【図 1 3】

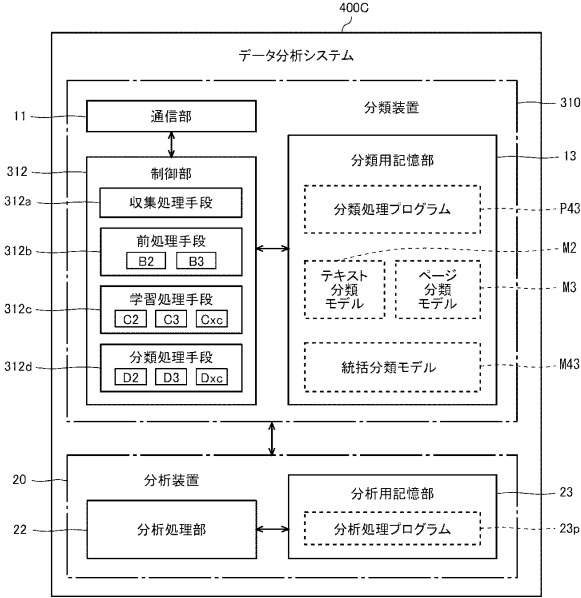


【図 1 4】

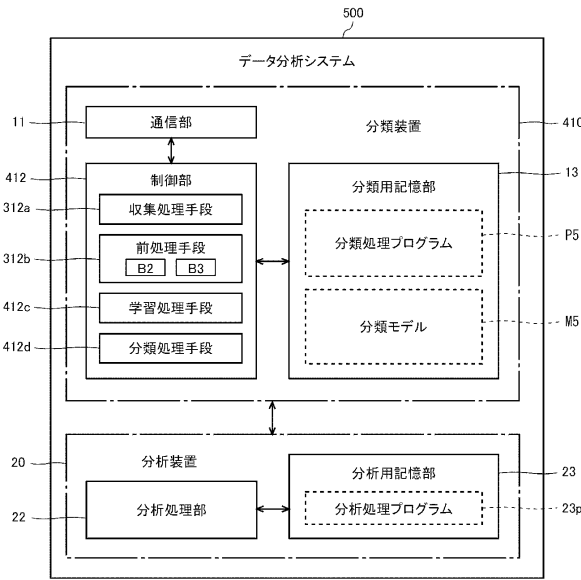


10

【図 1 5】



【図 1 6】



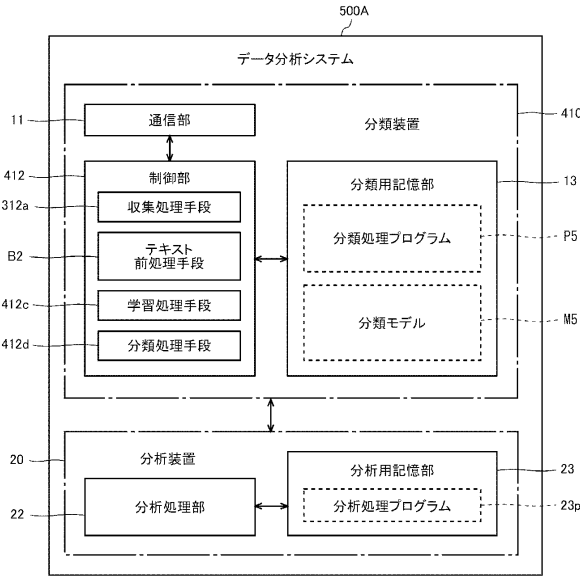
20

30

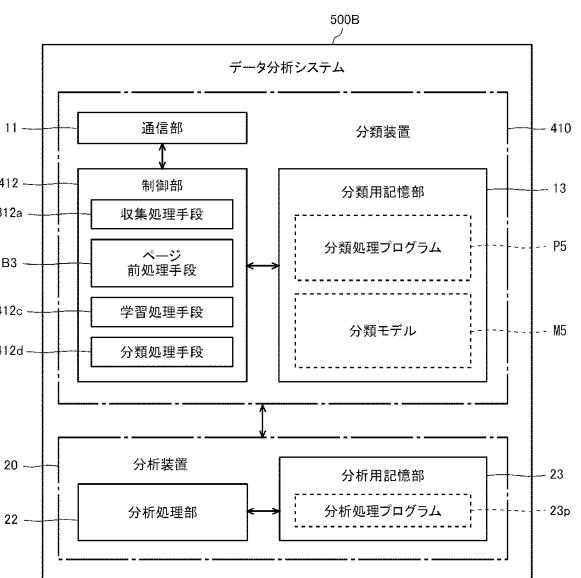
40

50

【図 17】



【図 18】



10

20

30

40

50

フロントページの続き

- (56)参考文献 特開 2 0 0 8 - 1 7 6 5 1 1 (J P , A)
米国特許出願公開第 2 0 1 4 / 0 0 4 0 7 2 3 (U S , A 1)
- (58)調査した分野 (Int.Cl., D B 名)
G 0 6 Q 1 0 / 0 0 - 9 9 / 0 0