

4-7 データハンドリング

東京大学 数理・情報教育研究センター

2020年4月28日

2024年4月28日改訂

概要

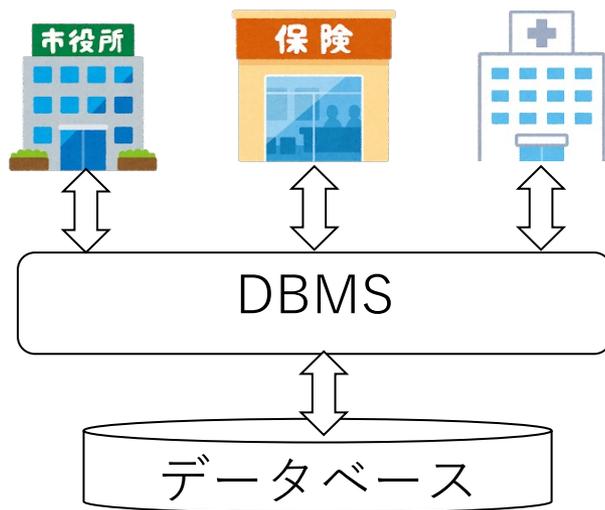
- 大規模なデータを収集，蓄積，整理し，集計・検索・抽出などに活用するためのデータベースと，その処理言語であるSQLの基礎について学びます。
- データ利活用のための簡単な前処理である名寄せなどのデータの変換，欠損値・外れ値などのデータクレンジングの基礎について学びます。
- 生成AIを利用したデータ加工の基礎について学びます。

本教材の目次

1. データベース	4
2. リレーショナルデータベース	6
3. SQL	12
4. データの変換	14
5. 欠損値	17
6. 外れ値・異常値	19
7. 生成AIを利用したデータ加工	22

データベース

- 大量のデータを蓄積，整理することで活用（集計，検索や抽出など）しやすくしたものを**データベース**と呼びます。
 - 例えば，チケット・航空券・宿泊などの予約・購入，商店の売上管理，銀行の預貯金管理，図書の貸出管理，自治体の住民登録，病院の患者情報管理，などの様々な情報システムで利用されています。
- データベースは，複数の異なる情報システムで共有し利用できます。
- 共有利用を前提としたデータベースを効率よく安全に利用できるように，データベースは**DBMS（データベース管理システム）**で管理されます。



データベース

- DBMSには主に以下の機能があります。
 - 一貫性
 - 複数の情報システムでデータを共有・一元管理し、それらが矛盾しないようにする
 - 整合性
 - データの重複や、不正なデータの登録や更新を防ぐ
 - 独立性
 - データとシステム（プログラム）を独立して管理する
 - 機密保護
 - アクセス制限や認証によりデータを保護する
 - 障害対策
 - 障害に対して復旧できるようにする
- DBMSを介することで、利用者はデータベースの仕組みを意識せず利用することができます。
- 利用者がDBMSを介してデータベースを操作するための言語にSQLがあります。

リレーショナルデータベース

- データベースでは膨大なデータを高速に効率よく処理できるように、データをあらかじめ決められた形式で蓄積します。
 - 例えば、階層型、ネットワーク型、リレーショナル型、などの形式があります。
- データをリレーショナル型の形式で整理・蓄積するデータベースをリレーショナルデータベースと呼びます。リレーショナル型は多くの実用的なデータベースで利用されています。

リレーショナルデータベース

- リレーショナルデータベースでは、データを複数のテーブル（表）の形で整理して格納します。テーブルはレコード（行）とフィールド（列）で構成されます。
 - テーブルからデータを取得する際は、レコードとフィールドを指定します。
 - 行を一意に特定可能な列の項目を主キーと呼びます。
 - 主キーは複数の列の項目を組み合わせて作ることもできます。
 - 表にインデックス（索引）を作成するとデータに効率的にアクセスすることができます。一般にインデックスは主キーや一意性がある列に定義されます。
- リレーショナルデータベースを設計する際は、正規化によりデータの構造を整理します。適切に正規化を行うことでデータに効率的にアクセスできるようになりデータベースの処理速度が向上します。
 - 正規化には第一正規化から第五正規化まであります。例えば、第一正規化ではテーブルの中で繰り返されるフィールドを元のテーブルと主キーと併せて別のテーブルに分割します。

リレーショナルデータベース

<学生テーブル>

フィールド

主キー

登録番号	学籍番号	氏名	学部コード	学部名	入学年
1	L1111	佐藤	LI	文学部	2020/4
2	E2222	鈴木	EN	工学部	2019/4
3	S3333	高橋	SC	理学部	2020/4
4	L4444	田中	LA	法学部	2019/4
5	M5555	伊藤	ME	医学部	2020/9

レコード

テーブル間の参照

<成績テーブル>

学籍番号	氏名	得点
L1111	佐藤	82
E2222	鈴木	65
S3333	高橋	77
L4444	田中	75
M5555	伊藤	70

テーブル間の参照

<学部テーブル>

学部コード	学部名	人数
LI	文学部	50
EN	工学部	100
SC	理学部	70
LA	法学部	60
ME	医学部	60

リレーショナルデータベース

- リレーショナルデータベースのテーブルに対する代表的な操作には、結合、選択、射影、があります。
 - 結合
 - 複数の表で共通する項目によりそれらの表同士を結合して新しい表を作成します.
 - 選択
 - 与えられた条件に合致する行のみを表の中から抽出して新しい表を作成します.
 - 射影
 - 表の中から指定された列のみを抽出して新しい表を作成します.

リレーショナルデータベース

- 元の表から、結合、選択、射影、などの操作で抽出した表をビューと呼びます。ビューは仮想的な表ですが、実際の表と同様に扱うことができます。
- この他にもリレーショナルデータベースに対する操作として挿入、更新、削除があります。

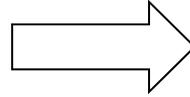
リレーショナルデータベース

テーブル1の学部コードがLIまたはLAのデータを選択

テーブル1

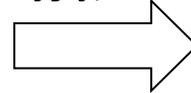
学籍番号	氏名	学部コード
L1111	佐藤	LI
E2222	鈴木	EN
S3333	高橋	SC
L4444	田中	LA
M5555	伊藤	ME

選択



学籍番号	氏名	学部コード
L1111	佐藤	LI
L4444	田中	LA

射影



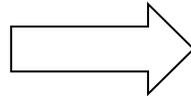
学籍番号	氏名
L1111	佐藤
E2222	鈴木
S3333	高橋
L4444	田中
M5555	伊藤

テーブル1の
学籍番号・氏名
フィールドを射影

テーブル2

学部コード	学部名
LI	文学部
EN	工学部
SC	理学部
LA	法学部
ME	医学部

結合



学籍番号	氏名	学部コード	学部名
L1111	佐藤	LI	文学部
E2222	鈴木	EN	工学部
S3333	高橋	SC	理学部
L4444	田中	LA	法学部
M5555	伊藤	ME	医学部

テーブル1と
テーブル2を
学部コード
フィールドで
結合

SQL

- **SQL**言語によりリレーショナルデータベースを操作することができます.
- データベースの操作はSQLを用いて以下のように記述することができます.

SELECT 表名.列名, 表名.列名, ...
FROM 対象となる表
WHERE 条件

- **SELECT**には対象となる表と抽出する列名を指定します.
- **FROM**には対象となる表を指定します.
- **WHERE**には行を抽出するための条件を指定します.

SQL

成績テーブル

学籍番号	氏名	得点
L1111	佐藤	82
E2222	鈴木	65
S3333	高橋	77
L4444	田中	75
M5555	伊藤	70

成績テーブルから得点が75以上の学生の氏名と得点を抽出

```
SELECT 成績.氏名, 成績.得点  
FROM 成績  
WHERE 得点 >= 75
```

```
佐藤, 82  
高橋, 77  
田中, 75
```

データの変換

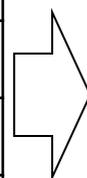
- リレーショナルデータベースでは、データをレコード（行）とフィールド（列）で構成されるテーブル（表）の形で整理して格納します。
- 表の列に入る値はさまざまなデータの型があります。データ型には数値型、文字型、日付・時刻型、バイナリ・論理型、カテゴリー型などがあります。

文字型	カテゴリー型	日付・時刻型	数値型	バイナリ・論理型
氏名	学部コード	受験日	得点	合否
佐藤	LI	2020/4/1	70	1
鈴木	EN	2020/4/1	50	0
高橋	SC	2020/4/7	80	1
田中	LA	2020/4/3	75	1
伊藤	ME	2020/4/5	60	0

データの変換

- データ型が文字型や日付・時刻型の場合，行ごとに文字列の表記が異なる**表記揺れ**が生じることがあります．データに表記揺れがある際は**名寄せ**を行うことで表記を統一するようにします．

社名	郵便番号	登録日
株式会社 東京	1231234	2020-4-1
(株) 東京	123-1234	2020/4/1
(株) トウキョウ	1 2 3 2 3 4	2020年4月1日
株式会社 Osaka	567-5678	1/5/2020
株式会社 大阪	5675678	1 st May 2020



名寄せ

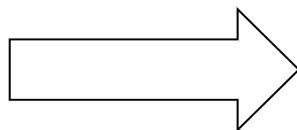
社名	郵便番号	登録日
株式会社 東京	123-1234	2020年4月1日
株式会社 大阪	567-5678	2020年5月1日

- データ型が数値型の場合，行ごとに数値の単位が異なることがあります．このような場合は数値の変換と単位の統一をするようにします．

データの変換

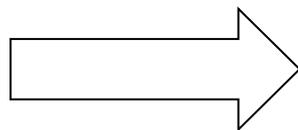
- データ型がカテゴリー型の場合，数値型に変換するか，または各カテゴリーに対応した新たなバイナリ型の列を作ることがあります。

氏名	学部コード
佐藤	LI
鈴木	EN
高橋	SC
田中	LA
伊藤	ME



カテゴリを数値に変換

氏名	学部コード
佐藤	1
鈴木	2
高橋	3
田中	4
伊藤	5



各カテゴリをバイナリに変換

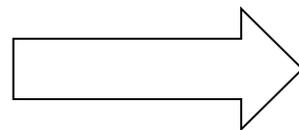
氏名	LI	EN	SC	LA	ME
佐藤	1	0	0	0	0
鈴木	0	1	0	0	0
高橋	0	0	1	0	0
田中	0	0	0	1	0
伊藤	0	0	0	0	1

欠損値

- 何からの原因により欠損しているデータの値を欠損値と呼びます。
 - 例：機器の故障による計測データの欠損値，未回答項目によるアンケートデータの欠損値など
- データのクレンジングでは，分析対象のデータに対して欠損値の検出を行い，必要であれば欠損値の補完や除去を行います。
- 欠損値の補完や除去などの処理においては，欠損値の発生パターンをよく理解してから行う必要があります。
 - 欠損値はランダムに発生したものか？規則性はあるか？
- 欠損値の補完
 - 欠損値を含むデータの非欠損値の代表値（平均や中央値）で補完
 - 欠損値を含むデータの非欠損値から確率分布の母数を推定し，推定した確率分布に基づいて補完
 - 欠損値を含むデータを他のデータから回帰するモデルを推定し，推定した回帰モデルに基づいて補完 など

欠損値

時	気温 (°C)
0	9.5
1	- 欠損値
2	8.7
3	8.8
4	7.4
5	7.6
6	9.5
7	- 欠損値
8	12.9
9	14.3



非欠損値の平均値
で欠損値を補完

時	気温 (°C)
0	9.5
1	9.8
2	8.7
3	8.8
4	7.4
5	7.6
6	9.5
7	9.8
8	12.9
9	14.3

外れ値・異常値

- データの一部の値が他の値から大きく外れた値を**外れ値**と呼びます。
 - 例：機器の異常による計測データの外れ，集計の誤りによるアンケートデータの外れ値など
- データの**クレンジング**では，分析対象のデータに対して外れ値の検出を行い，必要であれば外れ値の除去を行います。
- 外れ値の除去においては，その性質（偶発的なものか傾向を持ったものかなど）をよく理解してから行う必要があります。

外れ値

- 外れ値の検出
 - データの四分位点の第1四分位点 (Q1) , 第3四分位点 (Q3) , 四分位範囲(IQR=Q3-Q1)をもとに,
 - データの上限値を $Q3+1.5 \times IQR$, 下限値を $Q1-1.5 \times IQR$ と定め, 上限値以上または下限値以下の値を外れ値と定めることができます.
 - データの分布が既知の場合, 平均 μ と標準偏差 σ を元に, 例えば $\mu+2\sigma$ を上限値, $\mu-2\sigma$ を下限値として外れ値を検出することができます.

外れ値

ID	観測値
1	13.4
2	13.8
3	15.2
4	11.4
5	15.2
6	6.2
7	12.7
8	16
9	18.2
10	10

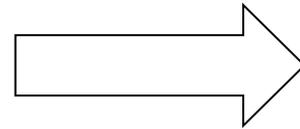
Q1 : 11.725

Q3 : 15.2

IQR : 3.474

データの上限値

$Q3 + 1.5IQR = 20.4$



データの下限値

$Q1 - 1.5IQR = 6.5$

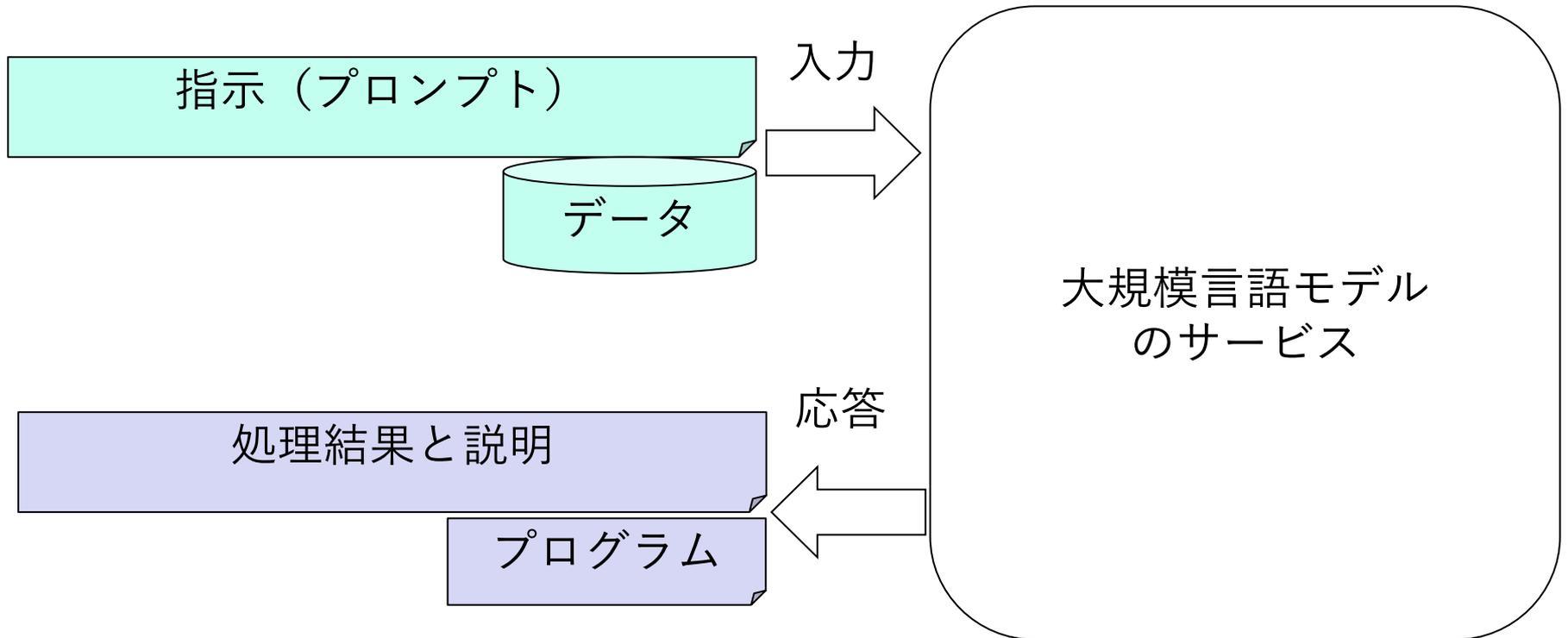
ID	観測値
1	13.4
2	13.8
3	15.2
4	11.4
5	15.2
6	6.2 外れ値
7	12.7
8	16
9	18.2
10	10

生成AIを利用したデータ加工

- 主に基盤モデルに基づく生成モデルを用いてテキスト、画像、その他のコンテンツを生成する人工知能（AI）の技術を生成AIと呼びます。
 - 基盤モデルは、大規模なデータをもとに事前学習された機械学習のモデルです。
 - 例えば大規模言語モデル（Large Language Model, LLM）は、大規模なテキストデータをもとに学習された基盤モデルです。
 - 大規模言語モデルは、与えられた文脈（単語の接続）から次の単語を予測するという事前学習により学習されます。
 - 加えて、指示に対して適切な応答を返すことや、倫理的・社会的に適切な応答を返すことを学習させることで、大規模言語モデルはさまざまなタスク（問題）を解くことができるようになります。
- 大規模言語モデルの応用として、プログラミングの支援とそれを活用したデータの処理があります。

生成AIを利用したデータ加工

- 大規模言語モデルに適切な指示（プロンプト）とデータを与えることで、モデルが生成したプログラムをもとにデータの集計・検索・抽出や前処理の作業を対話的に行うことができます。



生成AIを利用したデータ加工

大規模言語モデルのサービスとの対話の例

- 添付のファイルを読み込んでください
(教育用標準データセット*の基本素材 (SSDSE-E) のCSVファイルをアップロード)

* <https://www.nstac.go.jp/use/literacy/ssdse/>

- CSVファイルを正常に読み込むことができました。ファイルは日本の各種統計データを含んでおり、都道府県ごとの人口、出生率、医療施設数、消費支出など多岐にわたる情報が含まれています。具体的なデータ処理や分析が必要な場合は、どのように進めたいか教えてください。
- 総人口の列について北海道から沖縄県までの47都道府県の平均値を計算してください
- 47都道府県の総人口の平均値は約2,658,426人です。

生成AIを利用したデータ加工

大規模言語モデルのサービスとの対話の例（つづき）

- この平均値以上の総人口の都道府県を抽出してください

- 47都道府県の平均値以上の総人口を持つ都道府県は以下の通りです：
 - 北海道：5,140,000人
 - 茨城県：2,840,000人
 - 埼玉県：7,337,000人
 - 千葉県：6,266,000人
 - 東京都：14,038,000人
 - 神奈川県：9,232,000人
 - 静岡県：3,582,000人
 - 愛知県：7,495,000人
 - 大阪府：8,782,000人
 - 兵庫県：5,402,000人
 - 広島県：2,760,000人
 - 福岡県：5,116,000人

生成AIを利用したデータ加工

大規模言語モデルのサービスとの対話の例（つづき）

- 47都道府県の総人口の第1四分位点，第3四分位点，四分位範囲を求めてください
- 47都道府県の総人口について，以下の統計値が求められました：
 - 第1四分位点（Q1）：1,046,500人
 - 第3四分位点（Q3）：2,655,000人
 - 四分位範囲（IQR）：1,608,500人

生成AIを利用したデータ加工

大規模言語モデルのサービスとの対話の例（つづき）

- これらの値から外れ値の基準をつくれますか

- はい、これらの四分位数を使って外れ値の基準を作成することができます。外れ値の基準は、以下のように計算されます：
 - 下限の外れ値の基準: $Q1 - 1.5 \times IQR$
 - 上限の外れ値の基準: $Q3 + 1.5 \times IQR$
- これらの値を計算してみましょう。47都道府県の総人口に基づいた外れ値の基準は次のようになります：
 - 下限の外れ値の基準：-1,366,250人（実際の応用では、人口が負の値になることはありませんので、この下限は実用的ではありません）
 - 上限の外れ値の基準：5,067,750人 この上限値を超える人口を持つ都道府県は外れ値と考えられます。この基準により、特に人口の多い都道府県が識別されます。

生成AIを利用したデータ加工

大規模言語モデルのサービスとの対話の例（つづき）

- 具体的にどの都道府県の値が外れ値ですか

- 人口が外れ値として識別された都道府県は以下の通りです：

- 北海道：5,140,000人
- 埼玉県：7,337,000人
- 千葉県：6,266,000人
- 東京都：14,038,000人
- 神奈川県：9,232,000人
- 愛知県：7,495,000人
- 大阪府：8,782,000人
- 兵庫県：5,402,000人
- 福岡県：5,116,000人

これらの都道府県は、47都道府県の総人口の上限外れ値の基準（5,067,750人）を超える人口を持っています。これは、主に大都市圏が集中している地域であり、日本の人口分布の特徴を反映しています。

生成AIを利用したデータ加工

- 生成AIとして大規模言語モデルを利用したこれらの処理は、実際はモデルが自ら生成したプログラムによって行われています。
- 大規模言語モデルには、実際は存在しない情報を生成してしまうハルシネーション（幻覚）という問題が存在します。
- 大規模言語モデルが生成したプログラムとその処理結果についても、手続きや結果に誤りが無いか十分に留意して利用する必要があります。
- また、生成AIのサービス提供者が入力した情報をどのように取り扱っているのかを十分に確認した上で利用する必要があります。
 - 特に、機密情報や個人情報の扱いには十分に留意が必要です。