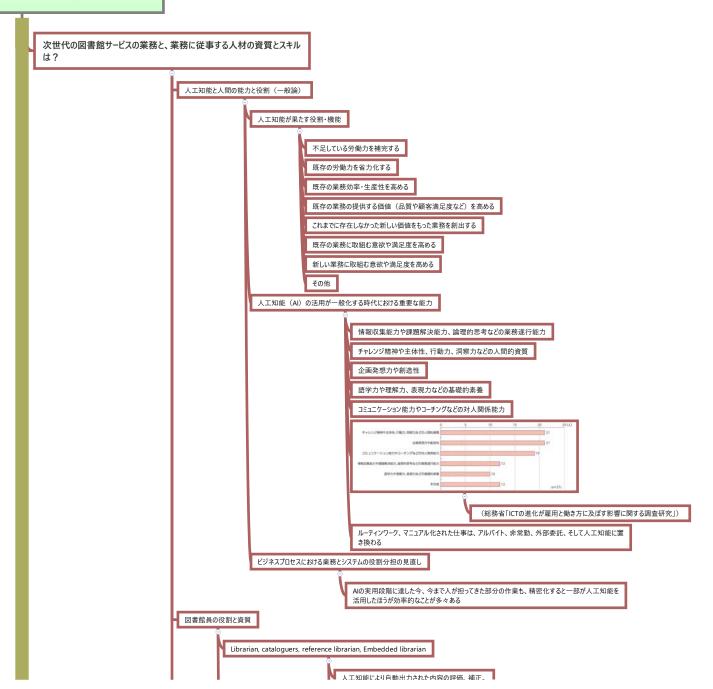
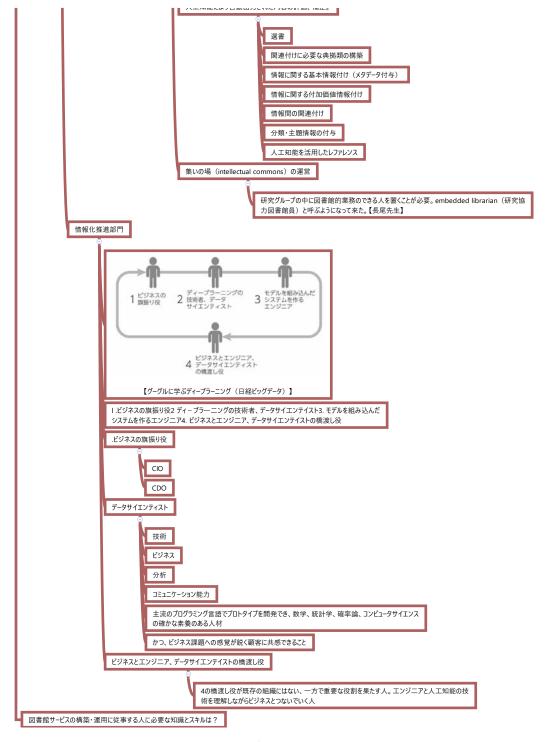
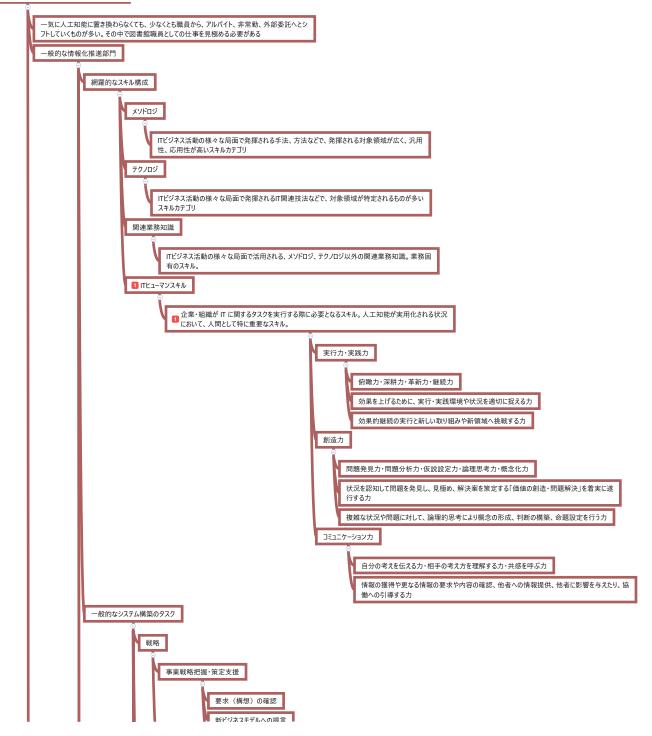
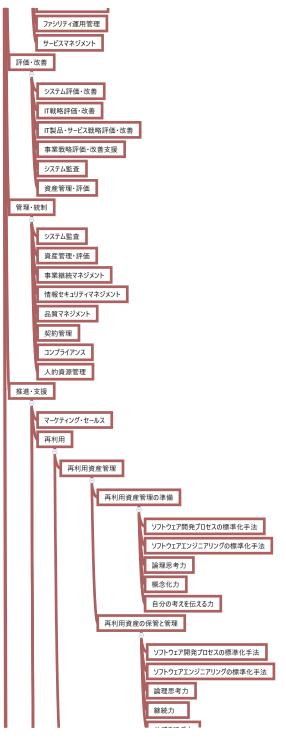
次世代図書館サービスの実現形でのタスクと必要なスキル・知識 【詳細】【2017年4月13日】



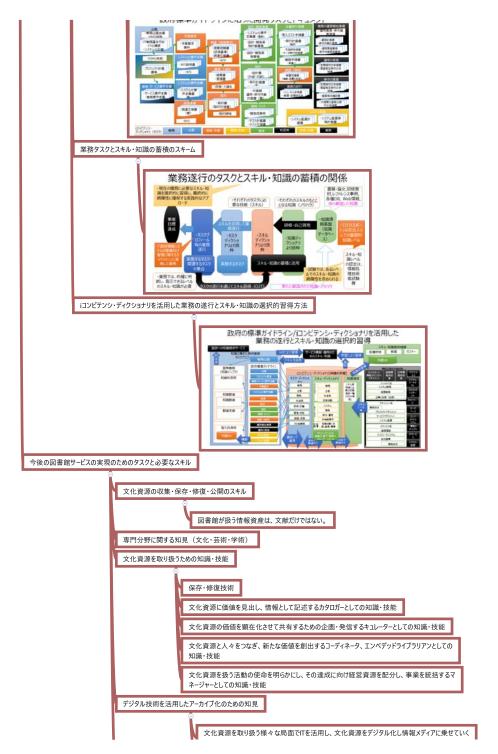




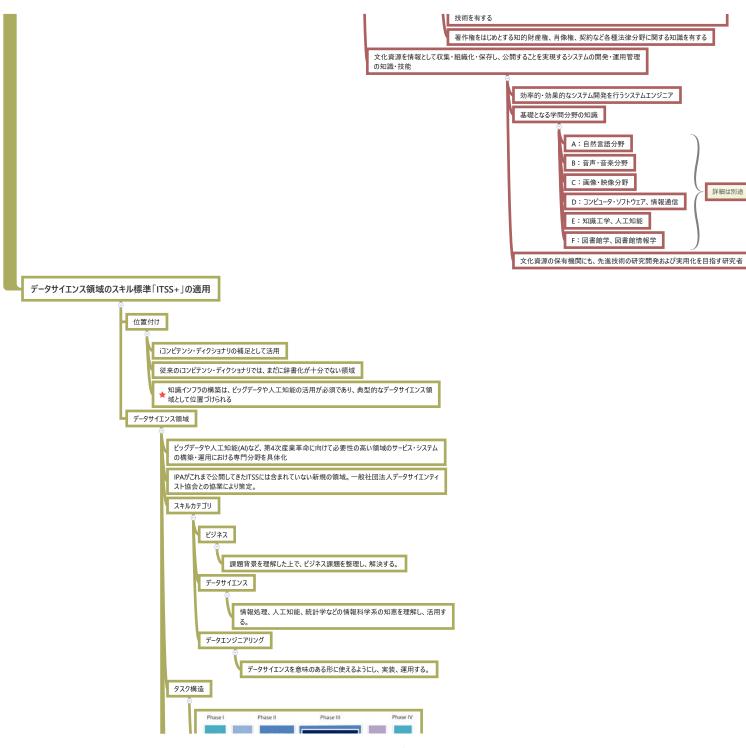




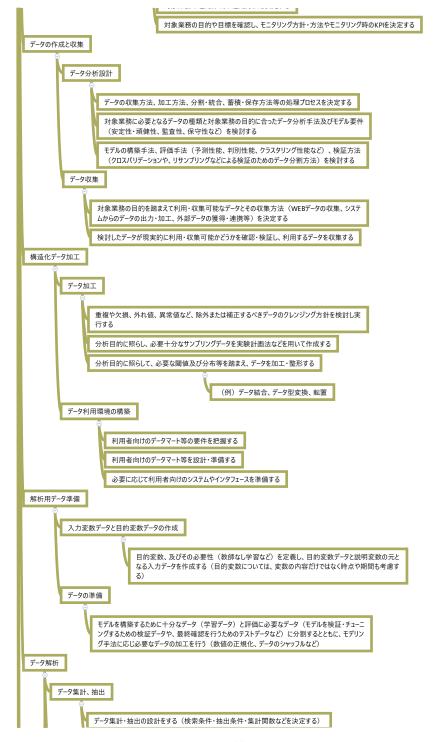




詳細は別途







データ集計・抽出を行う

性質・関係性の把握

収集したデータの基礎統計量(平均値、中央値、最頻値、標準偏差、分散など)を確認し、全 体のバラツキや傾向、異常値などを把握する

収集したデータを異なるレイヤで集計し、データが本来持つべき意味や活動目的の視点から違和感がないかどうかを見極める

収集したデータに対する基本的な分析(クロス集計、可視化など)を実施し、基本的な 2 変数間の関係性や傾向を把握する

収集した3次元以上の多変量データに対する分析(平行座標プロット、ネットワーク分析等)を 実施し、多変量におけるデータ間の関係性や傾向を把握する

データ間にどのような関係性 (例: 因果関係、ボトルネックなど) があるかを事前検討し、検証方法 (手法例:主成分分析、因子分析など、定義例: 距離定義など) を決め、関係性の有無を検証する

検定

医薬品開発などにおいて、母集団の平均の推定、2 群や多群間の平均値の差などの有意差を検定する (t 検定、クラスカル・ウォリス検定など)

品質管理などにおいて、サンプリングを行い、抽出されたサンプルから不良率などの検定を行い、評価する

A/Bテストの実施により得られた結果に対し、標本数が少ない場合、検定を行い有意差を判断する(カイ二乗検定など)

予測・判別

分析目的や選択するモデル手法の観点から、必要となる説明変数データや特徴値を作成する (手法例: p値による選択、正則化による除外・抑制、機械学習による自動作成など)

(数値予測) 売上予測、価格予測、発生確率予測など数値の予測モデルを構築する(手法例: 重回帰分析、ロジスティック回帰、ニューラルネットワークなど)

(2値の判別・予測) 与信可否判断や迷惑メール判別、顧客離脱など、二つの状態に対する判別・予測モデルを構築する(手法例:ロジスティック回帰、サポートベクターマシン (SVM) など)

(クラスの判別・予測) 新規顧客のセグメントや将来の顧客行動パターンなど、新たに発生するデータを分類するための判別・予測モデルを構築する(手法例:決定木、ニューラルネットワークなど)

(時系列予測)株価や景気変動など、過去の時系列データに基づく将来の推移・変動予測モデルを構築する(手法例:自己回帰モデル、移動平均モデル、ARIMA、SARIMAなど)

グルーピング

顧客のターゲティング/セグメンテーションや、データ傾向の把握を行うため、着目する類似度、距離を 定義し、対象をいくつかのグループに分類する(必要に応じて、分析の目的と評価指標に従って適 切なクラスタやグループ数を決定する)(手法例:階層クラスター分析、非階層クラスター分析 (k-means等)、自己組織化マップ (SOM)など)

故障検知、不正検知などを目的として、データ全体の傾向から著しく離れた対象(外れ値)や、 既知のデータとは振る舞いが異なる異常パターンを早期検知/検出する(手法例:k近傍法、 One-class SVM、反復子ニューラルネットワーク、及び検定、統計モデル、クラスター分析に基づく 手法など)

パターン発見

ヒトの行動データや製品のログデータなどにおいて頻出するパターンを抽出する(手法例:アプリオリアルゴリズムなど)

行動データや意識調査データから、コンテキストや意図/ニーズを把握する(手法例:主成分分析や因子分析など)



リアルタイム可視化 必要なリアルタイム性に応じてデータの分割転送、復元をコーディングする 非構造化データ処理 方針検討 目的を踏まえ、非構造化データの処理方針を検討する(言語:ポジネガ分析など、画像:画像 認識など、音:本人認証など) 言語処理 (データ収集) Webクローリング、DBアクセス、アンケート、文字認識、音声認識などの結果からテ キストデータを収集し、HTMLタグ等の不要な表現を除去して分析しやすいように整備する (辞書構築) 当該分野の専門用語や、同義語、類義語、対義語、人名・地名の辞書を、クラウ ドソーシングによる多数のユーザへのアンケートで収集したり、ブートストラップ法による事例の半自動生 成手法を用いて構築する (構造解析) 構築した辞書を活用してパラメータ調整を行った形態素解析器、構文解析器、固 有表現抽出器などを用いてテキストデータの構造解析を行う (特徴量変換) 低頻度語やストップワードの除去、名詞や動詞の正規化などを行い、テキストデー タを単語集合(Bag-of-words)、分散表現などの分析しやすい特徴量に変換する (教師あり分析)人手もしくは半自動で正解ラベルを収集したのち、SVMなどの教師あり分類学 習を行い、文書のジャンルや感情の判定、単語の品詞やカテゴリの推定を行う (教師なし分析)トピックモデル(Latent Dirichlet Allocation/LDA)などの統計モデルを用い て、文書をクラスタリングして類似文書にまとめたり、文書中の代表的なトピックの推定を行う (情報検索) 転置インデックスのような索引構造や潜在意味インデクシング (LSI) のような次元 圧縮手法を用いて文書の厳密・類似検索を高速に行う (文書生成) Nグラム統計モデルや再起ニューラルネットワーク (RNN) などを用いて、翻訳、要 約、対話、Q&Aなどの出力テキストを生成する 画像処理 画像データに付随するメタデータ(何の画像か、撮影条件、関心領域(Region of Interest) 等)を紐付ける 画像データを生成したセンサー特性にあった補正・修正を行う 分析しやすいデータフォーマット(チャンネルの設定、画素値の正規化・レベル変換、サイズ等)に 変換する (類似画像推定) 画像間の類似度を定義する順序や距離、画像特徴量を与える手法を決定 するとともに、類似画像を列挙するための特徴量を適切に保存・検索する手法も同時に検討する (画像認識) 画像のメタデータから認識対象のラベルデータを抽出し、画像特徴量を与える手 法、ラベルデータを対応付ける手法を決定する (画像復元) 復元精度を測る指標を選び、指標を最適化するような復元手法を決定する 音声/音楽処理 音をデジタルデータとして、wav や mp3 などにフォーマットする 音データからノイズ、雑音などの余計な音を識別して除去する 分析目的にあったデータ(音高、音量など)を抽出する (本人認証や話者識別) 音声データから抽出したスペクトラムやメル周波数ケプストラム係数 (MFCC) などの特徴量を用いて、本人認証や話者識別のためのパラメータの推定・モデル構築・ 精度評価を行う

