情報科学科　成研究室の紹介

成　凱　　九州産業大学　理工学部　情報科学科
Kai　CHENG　　Department of Information Science, Faculty of Science and Engineering, Kyushu Sangyo University
http://www.is.kyusan-u.ac.jp/~chengk/

1. はじめに

本研究室はデータベース，データ工学を専門として，「データ」を軸とした教育研究を行ってきた．近年，ビッグデータ，データサイエンスや AI など，「データ」を中心とする技術の発展が目覚ましく，新しい産業革命を引き起こそうとしている．データ工学（Data Engineering）は，大量のデータを収集，加工，蓄積し，分析可能な形式に変換するためのプロセスや技術であり，産業界では一般的に「データエンジニアリング」と呼ばれ，データサイエンティストやアナリストがデータにアクセスし，有用な情報を抽出するための基盤を提供する重要な分野である．

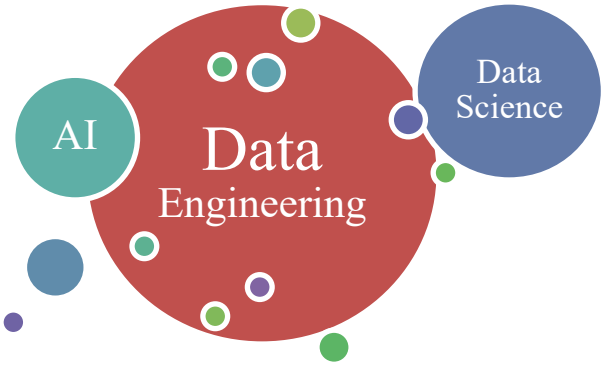

図 1　データ工学及びその関連分野

データ工学は，データベースをはじめとするデータ管理技術から発展してきた分野であり，データサイエンス，AI 等の分野と深く関連している（図 1）．データエンジニアリングは主にデータ基盤の構築とデータの操作に焦点を当てるのに対し，データサイエンスや AI はデータから価値を引き出すための解析やモデリングに焦点を当てている．

本研究室では，データ工学及びそれに関連する分野において，ビッグデータ基盤技術，自然言語処理・機械学習，情報検索及び推薦，偽情報対策，Web 情報システム等に関する研究開発を行ってきた．ここでは，いくつかをピックアップして紹介する．

2. ビッグデータ基盤技術に関する研究

近年，センサーやネットワーク技術の発達により，様々なデータが自動または半自動的に収集でき，実社会の時間的変更や傾向をデータとして把握できるようになっている．本研究室では，ビッグデータに関する基盤技術を開発している．

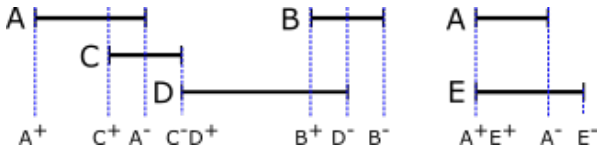

図 2　時区間データ例

2.1　時制データベースに関する研究

時制データベースは時制データ管理の代表的な技術であり，入院期間，投薬期間のような時間を表すための時間属性，及び治療費，投薬量のような非時間属性を同時に管理する．時制データベースは，データベースコア技術として，時間に伴って有効性が変化する情報，いわゆる時制データを管理するための基盤技術，時制データベースに関する研究を行っている．

時制データとは，事象の時間的制約を表し，その事象の開始と終了の時間や時間の継続を保持するようなデータであり，電子カルテ，ライフログ，環境モニタリングなどが挙げられる．時制データの有効活用は高度な意思決定を行うために重要であり，厚生医療，観光産業，金融業界など様々な分野で応用が期待されている．

2.2　シーケンスデータ解析手法の研究

様々なデータのなかで，購買履歴，DNA 系列，医療指示，移動軌跡，ライフログなど，時間の概念の有無にかかわらず，一定の順序をもつデータ，いわゆる「シーケンスデータ（sequence data）」が注目されている．シーケンスデータには興味深いパターンや特徴的動きが含まれており，それらを解析によって明らかにすることが重要である．シーケンスデータは問題領域によって多種多様であり，シーケンスデータ解析のための共通基盤が確立されていない．本研究では，問題領域にとらわれない共通の解析基盤を確立し，偽情報対策や微生物同

定等の問題に適用することを目的とする.

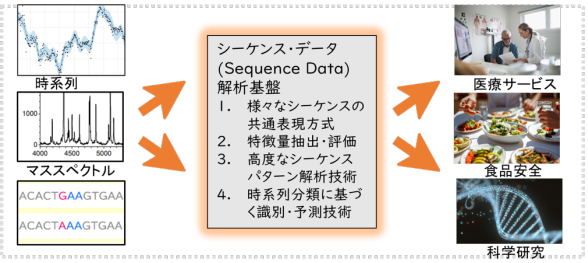

図 3 シーケンスデータ解析

3. 自然言語処理・機械学習に関する研究

これまで述べてきた研究テーマは, 主に数値等の構造化データを中心としている. インターネットの普及に伴い, SNS や口コミサイト等において, 自然言語で書かれた文章を解析し, 新しい知見を得ることが重要になっている. 本研究室では, 主に大学院生の研究テーマとして, 自然言語処理, 機械学習に関する研究を行ってきた.

3.1 偽情報対策に関する研究

a. 統計的特徴量に基づくフェイクレビュー検出

インターネットを介して商品を購入する際に, 実際に商品を手に取ってみることはできないため他の購入者の口コミであるレビューを参考にすることが多い. しかし, レビューの中に実際に商品を購入していない者によって書かれるフェイクレビューも存在する. レビューというものは個人の主観に基づいて書かれたものであり, レビューテキストからそのレビューが本物であるかどうかということを見分けるのが難しい. 本研究では, 「他の大多数のレビューとは異なる振る舞いを持つレビュー」をフェイクレビューとして考え, 異常検出という手法でフェイクレビュー検出を試みた.

b. 深層感情分析に基づくデマ検知

SNS 上で発信された情報の中に, デマを含む偽情報も拡散され, 社会問題となっている. デマ検知のためにはユーザ属性, 投稿コンテンツと拡散ネットワーク等複数の側面からデマの特徴を抽出し, 検知を行うことが一般的である. また, 単一側面ではデマ検知の精度が低く, 複数の側面を用いたマルチタスク学習が必要とされている. しかし, 先行研究では一部の側面しか用いられなかった. 本研究ではユーザ信頼性を考慮したスタンス分類と極性分類を補助タスクとする分類モデル提案し, Twitter データセットを利用し, デマ検知の精度を評価した.

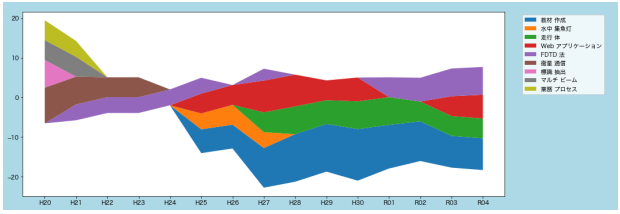

図 4 ThemeRiver による研究分野変化の可視化

3.2 自然言語解析・可視化

学術情報の電子化・大規模化が急速に進み, 学術データに隠れたパターンを解析し, 視覚的に確認することが研究者にとって重要になっている. 本研究では, 卒業研究のテーマを解析し, その中で隠れた技術分野の変遷を可視化し, 学生, 教員に提示することで, 学生の研究室選択や教員の研究指導に役立つ事を目的とする. 情報科学科卒業生約 20 年にわたる卒業研究のテーマに対して形態素解析を行い, 出現頻度の高い単語をキーワードとしてワードクラウドを使って可視化し, また, 年度別, 研究室別に集計を行い, それぞれのキーワードの変化を ThemeRiver で可視化した. 図 4 のように, 研究室単位での研究テーマの特色や, 時代の流れに伴う研究テーマの変化を確認できた.

4. Web 情報システム開発

本研究室では, 実社会に役立つ Web アプリケーション, Web 情報システムの開発も行っている. システム開発に関連する授業科目「データベース」や「Web プログラミング演習」等を担当している関係で, 演習の題材として, または, 卒業研究のテーマとして学生に課すことで, Web情報システムを設計から開発までの作業を経験してもらっている. また, 地域連携プロジェクトとして, 施設予約システムの開発を行い, 時制データベースの研究成果を実社会に還元している.

5. まとめ

本研究室これまで取り組んできた研究テーマを紹介した. 近年, 科学研究費, KSU 基盤研究費, 受託研究研究費等, 学内外から予算を獲得して, これまでの研究を深めながら, 新しい領域(例えば, 微生物同定)の開拓を行っている. 共同研究等にご興味のある方はご相談下さい. また大学院進学や卒業研究で本研究室をご希望する方にはさらに詳しく説明するのでご連絡下さい.