ディジタル図書館研究室

研究室紹介
2017/4/25 メディアプロジェクト演習2

前田　亮

「ディジタル図書館」とは？

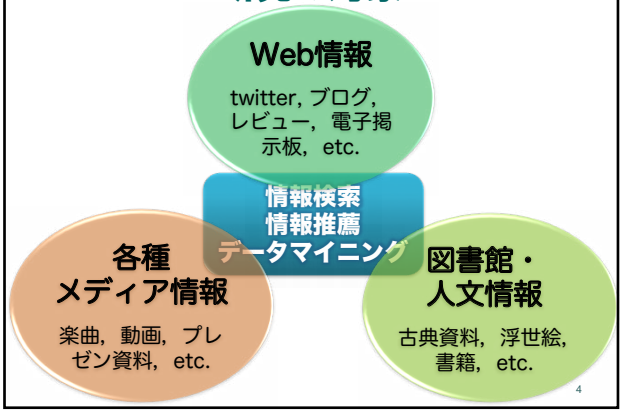
- ディジタル情報を収集・整理・保管し，提供するサービス全般
  - 対象となる情報の種類は限定しない
    - *電子書籍・新聞・雑誌，古典資料，芸術作品，Webページ，画像・映像・音源，etc.*
- Web上の情報は，従来の図書館のように体系付けられていない
- しかし，人類の知識を蓄積・提供する一種の巨大な**ディジタル図書館**と捉えることができる

ディジタル情報の利点

- 「**検索**」ができる
- 情報の内容から自動的に「**分類**」できる
- 隠れた知識を「**発見**」（マイニング）できる
- 利用者に合った情報を「**推薦**」できる
- 様々な情報源を「**統合利用**」することができる

- これらの技術を「**情報アクセス技術**」と呼ぶ

研究の対象



主な研究テーマ

- **Web情報**
  - 文章中の**ネットスラング**の正しい表現への変換
  - スポーツ中継中の**Tweet**解析によるファン視点でのイベント検出
  - **多言語Wikipedia記事**への言語横断エンティティリンキング
- **各種メディア**
  - 図形群の意味や階層構造を用いた**プレゼンスライド**検索システム
- **古典史料**
  - **古典史料テキスト**からの情報抽出
  - 異言語の**浮世絵**データベースからの同一作品の同定

Web情報

文章中のネットスラングの正しい表現への変換

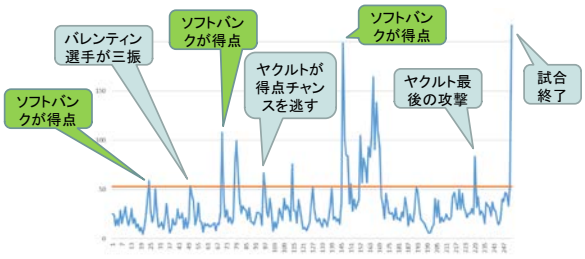
- 電子掲示板ではネットスラングが用いられる
  - 「草」「乙」「厨房」「密林」「リア充」「希ガス」…
- 形態素解析辞書と単語共起の統計情報を用いて，ネットスラングを正しい表現に変換

実際の書き込み	変換結果
ありがとう白と赤を密林に頼んできましたwktkしながら待ちますw	ありがとう白と赤をAmazonに頼んできましたワクワクテカテカしながら待ちます(笑)
本人と決めつける「自演乙！」	本人と決めつける「自演お疲れ様！」

赤字：ネットスラング　　青字：正しい日本語表現

スポーツ中継中のTweet解析によるファン視点でのイベント検出

- Tweetの盛り上がり具合（Tweet数，RT数，選手名の出現数）を基に，ファン視点でのハイライトを検出



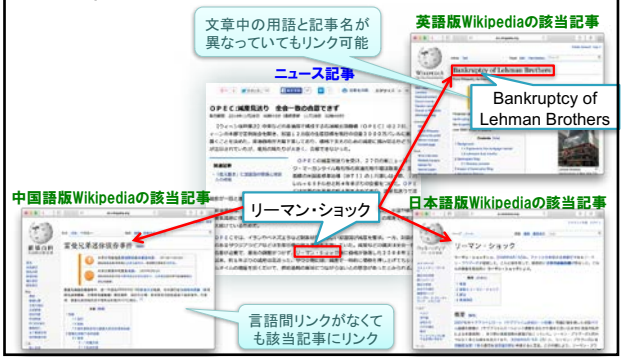
SMBC日本シリーズ2015, 東京ヤクルトスワローズvs.福岡ソフトバンクホークス

多言語Wikipedia記事への言語横断エンティティリンキング

- 文書中のキーワードから，該当する知識ベース記事（Wikipediaなど）に自動的にリンクを付与する手法

- 用語によっては，母国語の記事がない，あるいは他言語の記事の方が充実している場合もある
  - 「Christendom（キリスト教界）」の日本語記事は存在しない

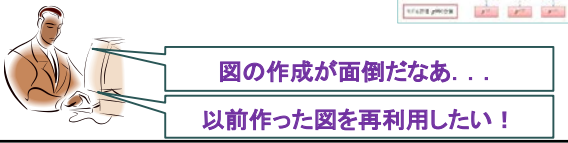
言語横断エンティティリンキングの例



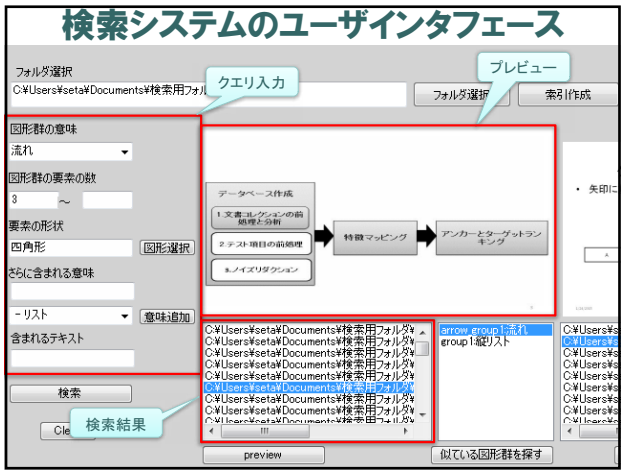
各種メディア

図形群の意味や階層構造を用いたプレゼンスライド検索システム

- プレゼンテーションスライド
  - 視覚的にわかりやすいように図が多用
- 図の作成は手間がかかる
  - 図の再利用のニーズ
- 従来のスライド検索手法では，図自体の検索が困難



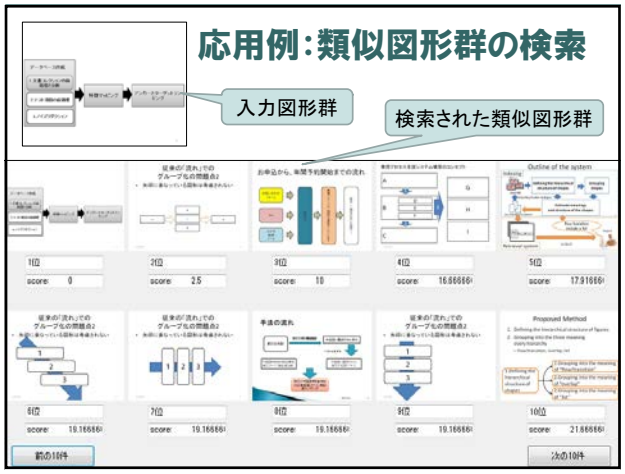
検索システムのユーザインタフェース



検索の例

入力クエリ
図形群の意味：**重なり**
図形群の要素の数：4
図形群の要素の形状：**楕円**

応用例：類似図形群の検索



古典史料

16

古典史料からの人物表現の自動抽出

平清盛

実名　別名　役職

鶏鳴清盛朝臣、義朝、義康等、軍兵都六百余騎発向白河、
六波羅亭、入道御座宰相中将被装束、次供奉行列被渡南庭、
上皇女御殿御同車、自御前渡御馬場桟敷、
前太政大臣以下歩行扈従、

人間関係の可視化などに役立つ

17

古典資料テキストからの人物情報の抽出と可視化①

役者評判記には，歌舞伎役者の演技や容姿に対する評価（レビュー）が書かれている

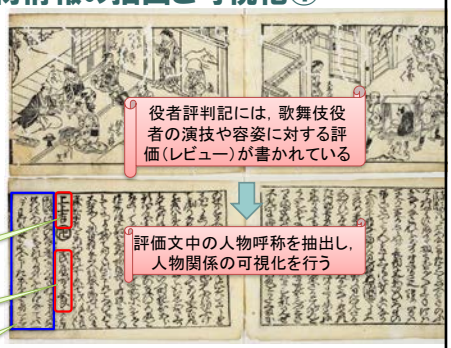
役者評判記でレビューされている歌舞伎役者

評価文中の人物呼称を抽出し，人物関係の可視化を行う

位付（役者の評価）

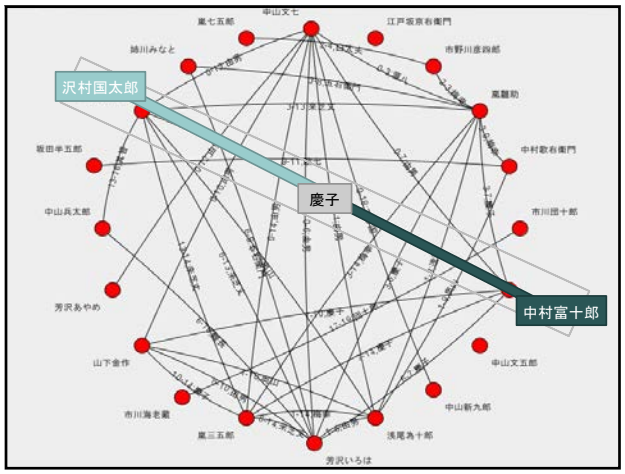
役者名

評価文

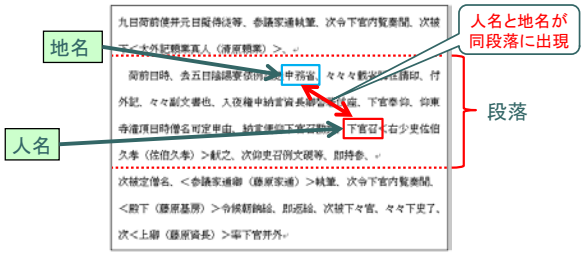


役者評判記『役者大極舞』
立命館大学アート・リサーチセンター（http://www.arc.ritsumei.ac.jp/index.html）所蔵

18



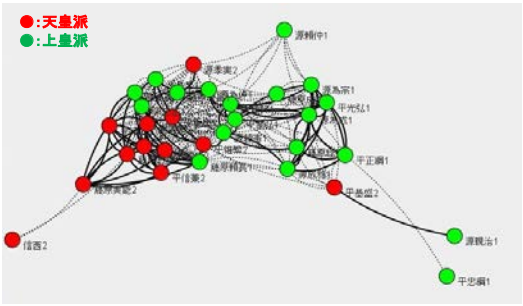
古典資料テキストからの人物情報の抽出と可視化②



- ある人名と地名の組が，同じ段落の中に現れた回数を集計することで，人名と場所の関連を推定

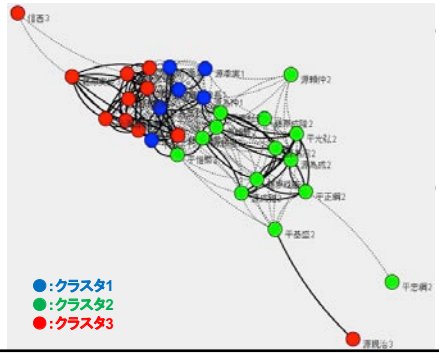
20

保元の乱における天皇派と上皇派の関係の推定結果①



21

保元の乱における天皇派と上皇派の関係の推定結果②



天皇派・上皇派が，ほぼ正しく分類されている

その中間に派閥が形成されているように見える

22

『東大寺要録』への注釈作業支援システム

- 12世紀に成立したとされる歴史資料
  - 東大寺の歴史と当時の状況などが記されている
  - 全十巻のうち一巻（約2万字）が電子テキスト化
- 東大寺要録研究会が書き下しと注釈作業中

電子テキスト化

原夫東大寺者。平城宮御宇勝実威神皇帝御願。天下第一大伽藍也。

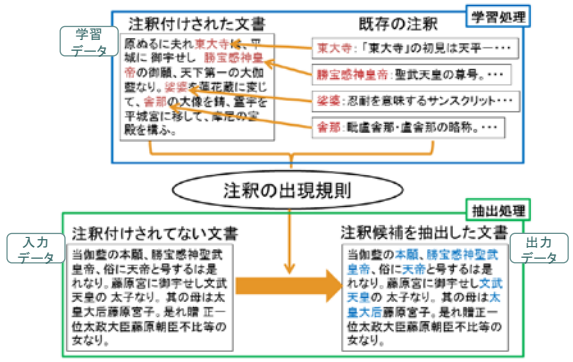
注釈付き書き下し文

東大寺要録巻第一原ぬるに夫れ東大寺は、平城に**御宇**せし**勝宝感神皇帝**の御願、天下第一の大伽藍なり。

- 注釈作業には，複数の分野の専門知識が必要
  - 専門分野によって知識に差がある

23

機械学習による注釈候補の自動抽出



24

異言語の浮世絵データベースからの同一作品の同定

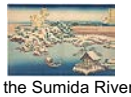
- 浮世絵は，国内外の美術館・博物館で個々にディジタル化され公開
  - 同じ浮世絵作品が複数のデータベースに散在
  - データベースによってメタデータの言語・表記が異なる

25

浮世絵研究者からの要望

- データベース間で散在している同一作品を見つけたい
- 同一作品のメタデータを比較することにより，誤りの発見・修正，より詳細な情報の取得などが可能となる

「雪月花隅田」
作家/制作者：葛飾北斎/画
制作年：天保3年
www.edo-tokyo-museum.or.jp

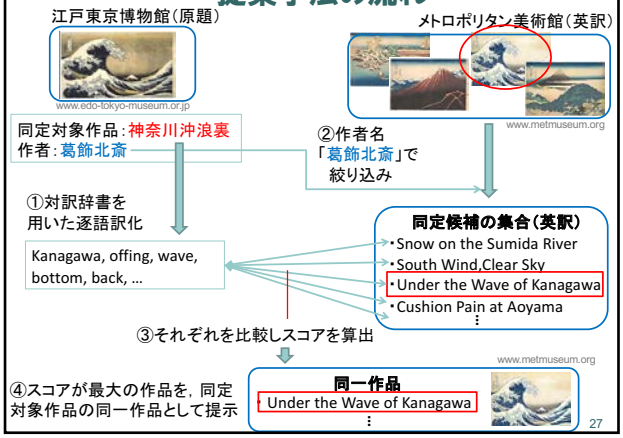


「Snow on the Sumida River, From the series, Snow, Moon…」
Artist：Katsushika Hokusai
Date：ca. 1833
www.metmuseum.org

- 本研究では原題と英訳を用いた同一作品の同定手法を提案

26

提案手法の流れ



27

メディアプロジェクト演習2 作品制作のヒント（1）

- **「調査研究」**
  - **新しいWeb技術**にはどのようなものがあり，それによって何が実現できるか
  - **ディジタル図書館システム**としてどのようなものが開発され，技術的課題は何か
- **「未来創造」**
  - **従来の図書館とディジタル図書館**は，将来どのように共存していけばよいか
  - ディジタル情報を**長期にわたって保存**していくには，どのようにすればよいか

28

メディアプロジェクト演習2 作品制作のヒント（2）

- **「ソフトウェア」**
  - 簡単な**情報検索システム**の作成
    - メディア情報学実験2「情報検索」の発展
    - 検索結果のランキング・分類・可視化など
  - 簡単な**ディジタル図書館システム**の作成
    - 各種WebサービスAPIの利用
      - 国立国会図書館サーチAPI
      - 図書館API（カーリル）
      - 楽天ブックス書籍検索API
      - 版元ドットコム・書誌情報API

29

ディジタル図書館研究室について

- **現在のメンバー**
  - 前田亮 教授
  - ビルゲサイハン・バトジャルガル 専門研究員
  - 大学院 博士課程1名，修士課程7名
  - 学部生11名
- **共同研究**
  - 本学文学部，MOT研究科など
- **研究成果（2016年度）**
  - ジャーナル1件，国際会議6件，国内会議5件，受賞1件

30

おわりに

- 当研究室では，**ディジタル図書館・情報アクセス技術**に関わる研究を幅広く行っている
- 研究室ホームページ
  - **http://www.dl.is.ritsumei.ac.jp/**
- メールアドレス
  - **amaeda@is.ritsumei.ac.jp**

31