

デ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

1

「デジタル図書館」とは？

- デジタル情報を収集・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**デジタル図書館**と捉えることができる



2

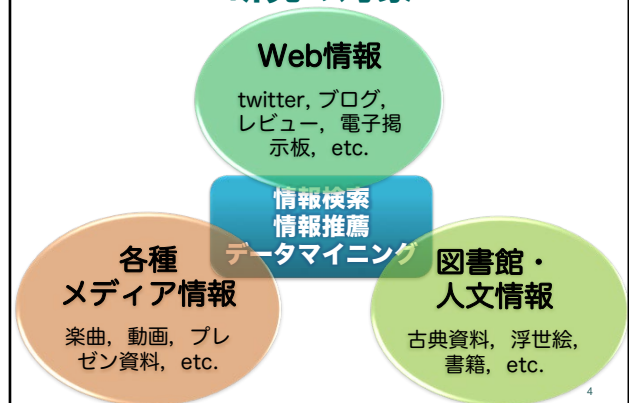
デジタル情報の利点

- 「**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



3

研究の対象



4

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ネキング
- **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⁵



Web情報

6

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
ありがとうございますを密林に頼んできましたwktkしながら待ちますw	ありがとうございますと赤を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！」

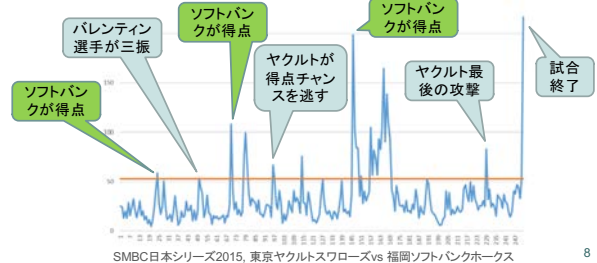
赤字: ネットスラング

青字: 正しい日本語表現

7

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



SMB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

8

多言語Wikipedia記事への言語横断エンティティリンク

- 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

日本語文書

理化学研究所発生・再生科学総合研究センターと先端医療センター(ともに神戸市)は12日、**加齢黄斑変性**(人工多能性幹細胞)から臓器の細胞を、目の細胞**加齢黄斑変性**の患者に移植する臨床手術を行ったと発表した。

リンクを付与

加齢黄斑変性?

加齢黄斑変性

日本語Wikipedia記事

- 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)」の日本語記事は存在しない

9

言語横断エンティティリンクの例

文章中の用語と記事名が異なってもリンク可能

英語版Wikipediaの該当記事

Bankruptcy of Lehman Brothers

中国語版Wikipediaの該当記事

リーマン・ショック

日本語版Wikipediaの該当記事

リーマン・ショック

言語間リンクがなくても該当記事にリンク

各種メディア

11

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

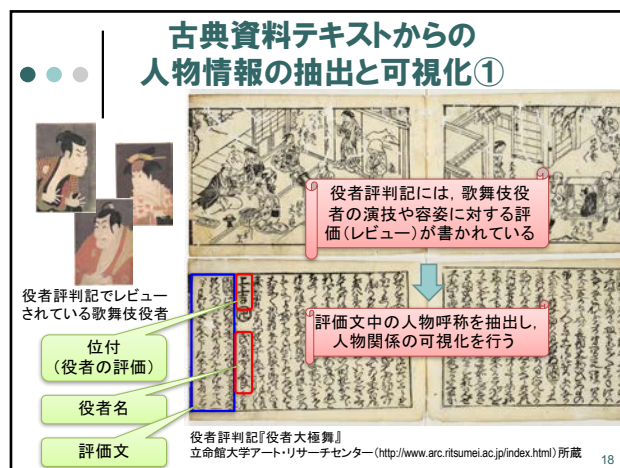
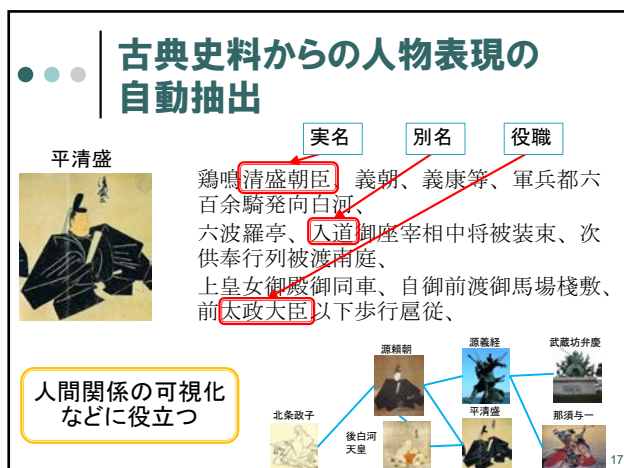
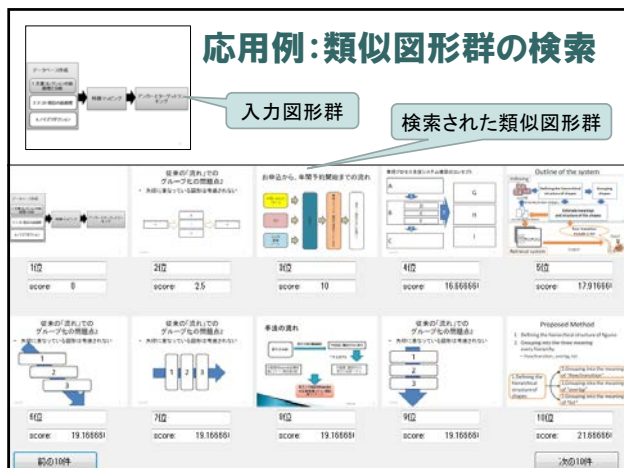
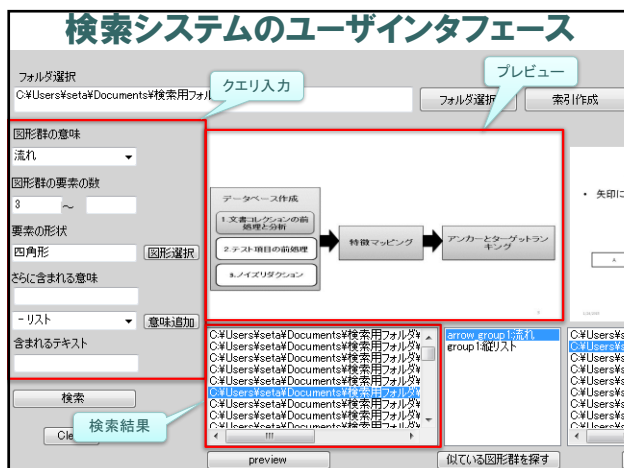
- プレゼンテーションスライド
 - 視覚的にわかりやすいように図が多用
- 図の作成は手間がかかる
 - 図の再利用のニーズ
- 従来のスライド検索手法では、図自体の検索が困難

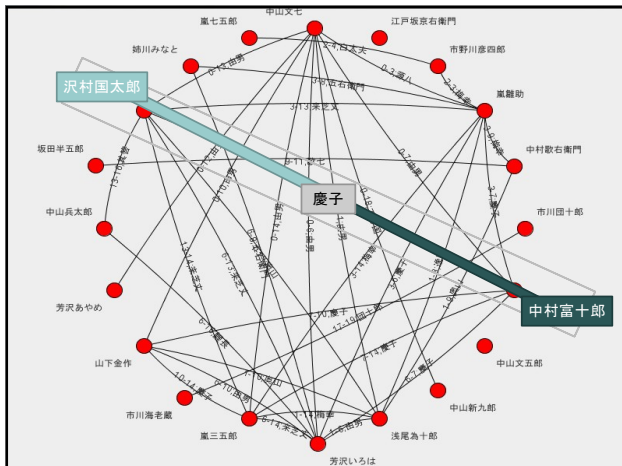


図の作成が面倒だなあ...

以前作った図を再利用したい!

12





古典資料テキストからの人物情報の抽出と可視化②

地名

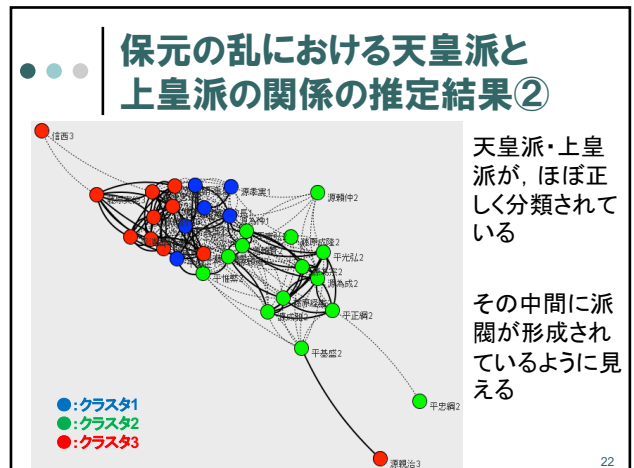
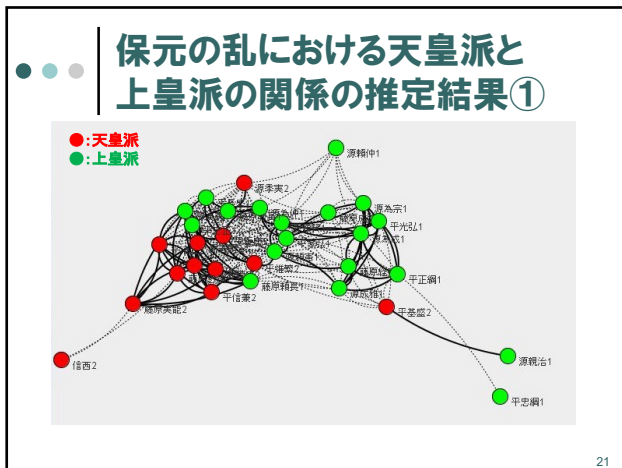
九日南前使并元日隆待従等、参議家通執筆、次令下管内覧奏聞、次被
 荷前日時、去五日陸陽家依例、申務出、々々々々々々々々々々、付
 外記、々々副次書也、入夜種申結言資長御書付、下官等仰、仰東
 寺准頂日時僧名可定申由、結言連仰下官等、下官召、右少史佐伯
 久孝（佐伯久孝）>献之、次仰史召問文、即待参、
 次被定僧名、<参議家通（藤原家通）>執筆、次令下管内覧奏聞、
 <殿下（藤原基房）>令候朝餽給、即返給、次被下々官、々々下史了、
 次<上卿（藤原資長）>率下官并外、

人名

人名と地名が
同段落に出現

段落

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



『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

機械学習による注釈候補の自動抽出

学習データ

注釈付けされた文書

原るに夫れ東大寺は、平城に御宇せし勝宝感神皇帝の御願、天下第一大伽藍なり。後述を蓮花蔵に表して、舎那の大像を鎮、寶宇を平城宮に移して、庵尼の室殿を構ふ。

既存の注釈

東大寺：「東大寺」の初見は天平……

勝宝感神皇帝：聖武天皇の尊号、……

娑婆：忍耐を意味するサンスクリット……

舎那：毗盧舎那・盧舎那の略称、……

学習処理

注釈の出現規則

入力データ

注釈付けされていない文書

当伽藍の本願、勝宝感神聖武皇帝、俗に天帝と号するは是れなり。藤原宮に御宇せし文武天皇の太子なり。其の母は太皇太后藤原宮子。是れ藤原正一位太政大臣藤原朝臣不比等の女なり。

抽出処理

注釈候補を抽出した文書

当伽藍の本願、勝宝感神聖武皇帝、俗に天帝と号するは是れなり。藤原宮に御宇せし文武天皇の太子なり。其の母は太皇太后藤原宮子。是れ藤原正一位太政大臣藤原朝臣不比等の女なり。

出力データ

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

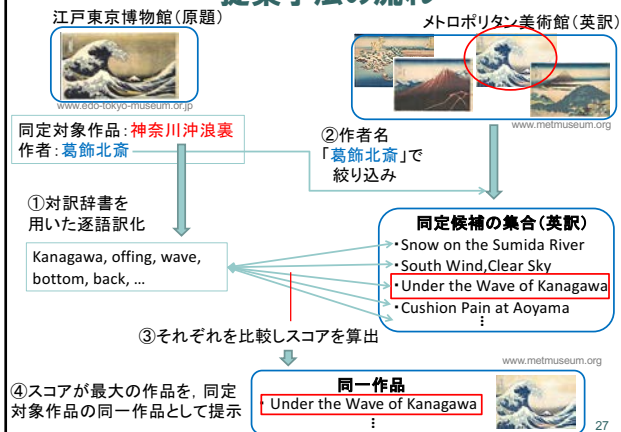
- データベース間で散在している**同一作品を見つけたい**
- 同一作品のメタデータを比較することにより、誤りの発見・修正、より詳細な情報の取得などが可能となる



- 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.ritsumeai.ac.jp/>
- メールアドレス
 - amaeda@is.ritsumeai.ac.jp

32