ディジタル図書館のための情報アクセス技術

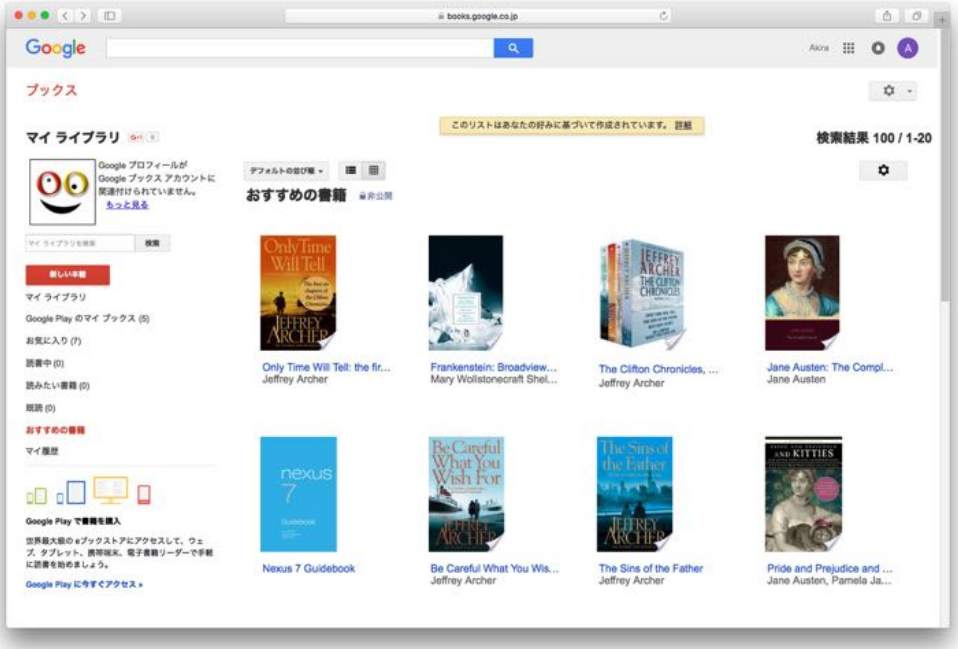
研究分野紹介

2020/10/15 先端社会デザイン概論

前田　亮

ディジタル図書館とは？

Wolf Hall


Google
ブックス
マイ ライブラリ
おすすめの書籍
検索結果 100 / 1-20
Only Time Will Tell: the fir...
Jeffrey Archer
Frankenstein: Broadview...
Mary Wollstonecraft Shel...
The Clifton Chronicles, ...
Jeffrey Archer
Jane Austen: The Compl...
Jane Austen
Nexus 7 Guidebook
Be Careful What You Wis...
Jeffrey Archer
The Sins of the Father
Jeffrey Archer
Pride and Prejudice and ...
Jane Austen, Pamela Ja...

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

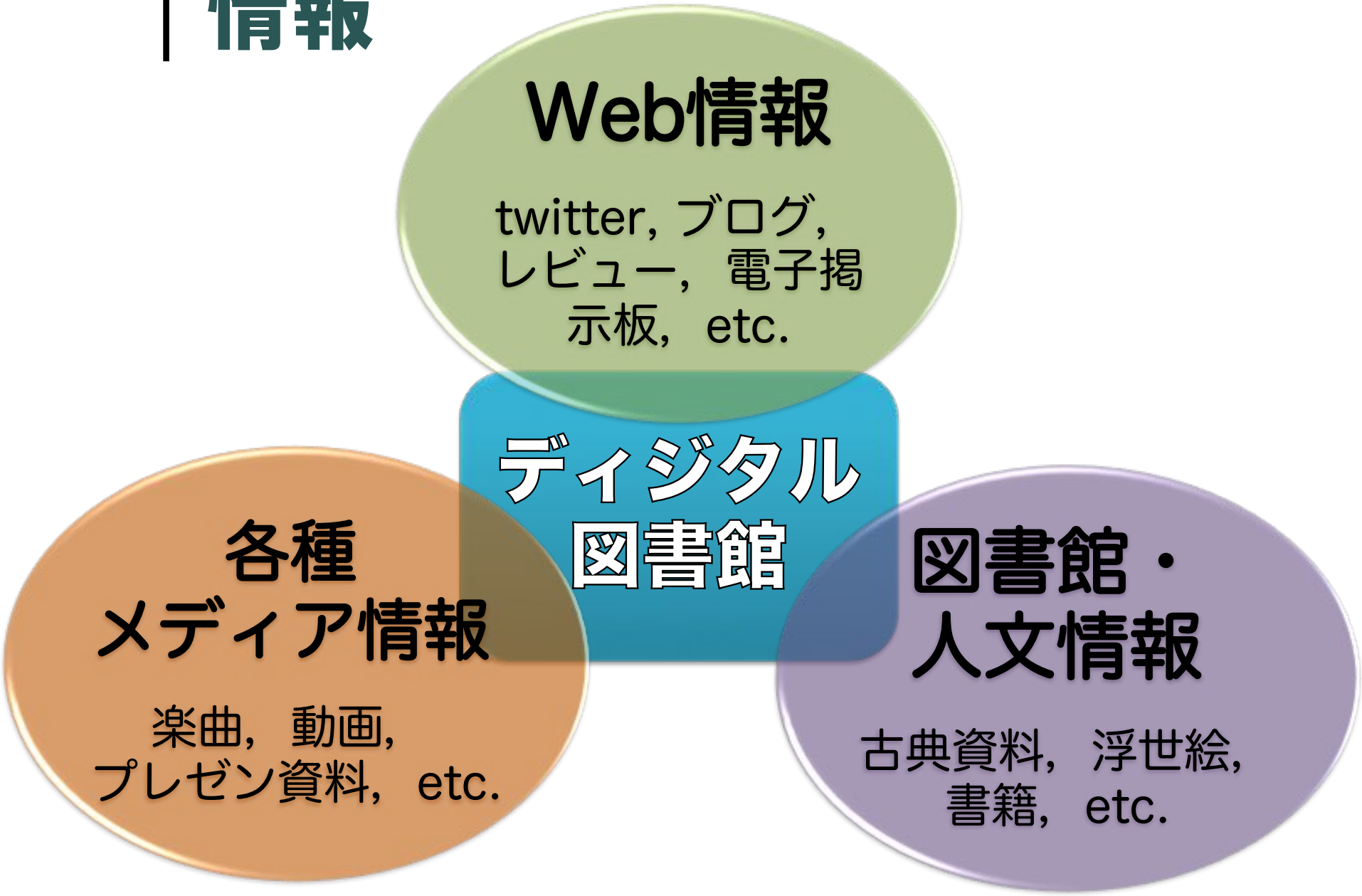
ディジタル図書館・アーカイブ（1）

- 図書館・博物館・大学・自治体など，さまざまな機関・団体が，自らが所有する貴重資料などをディジタル化して公開するサービス
  - 郷土資料・行政資料・貴重書など
  - 例：立命館大学所蔵貴重書アーカイブ
  - 例：国立国会図書館デジタルコレクション

ディジタル図書館・アーカイブ（2）

- 過去の資料をアーカイブするだけではなく，現在の資料を後世に残す必要
  - 現在膨大な量のディジタルデータが日々生産されているが，それが未来に残る保証はない
  - 「ボーンディジタル」資料などと呼ばれる
  - 例：NDL東日本大震災アーカイブ

ディジタル図書館が対象とする情報



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

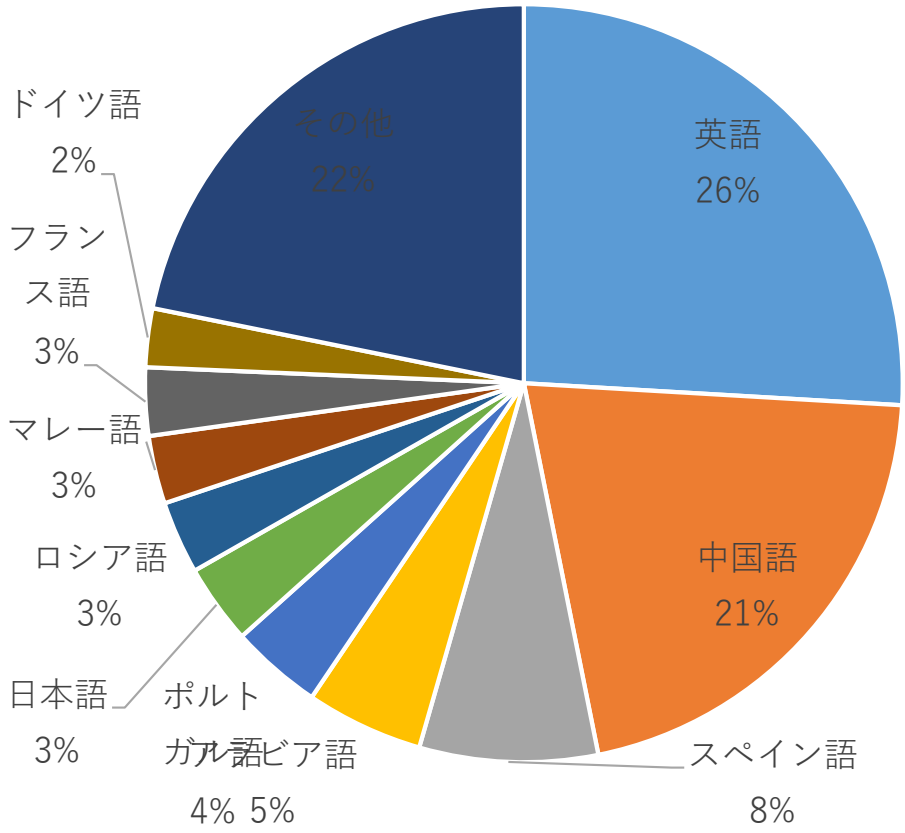
- これらの技術を「**情報アクセス技術**」と呼ぶ

ディジタル図書館のための情報アクセス技術①

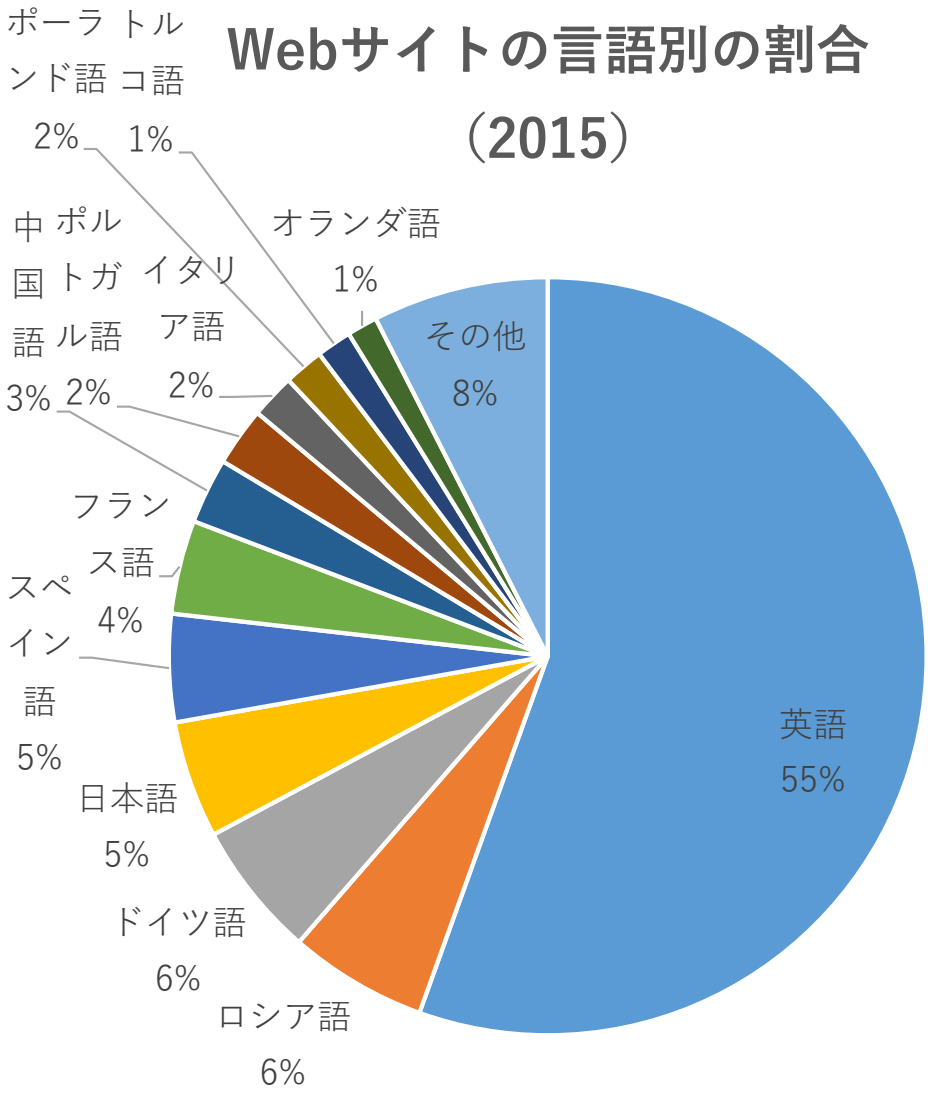
多言語情報の有効活用

インターネットで使用されている言語の統計

インターネット利用者の言語別の割合（2015）



Webサイトの言語別の割合（2015）



多言語情報アクセスの技術

- Web上には，さまざまな言語の情報が混在
  - *Wikipediaは313の言語版が存在（2020/10現在）*
- 日本語のWeb情報は，全体のわずか5%程度に過ぎない
  - *膨大なWeb情報の一部しか活用できていない*
  - *欲しい情報が日本語とは限らない*
- 情報技術により「**言語の壁**」を超える
  - **機械翻訳**サービス（Google翻訳など）
  - **多言語情報アクセス**技術

研究事例

浮世絵画像データベースへの多言語情報アクセスシステム

- 国内外の多くの図書館・美術館・博物館で，浮世絵作品が公開されている
  - 同じ浮世絵作品が複数のデータベースに散在
  - データベースにより，作品情報が記述される言語が異なる

国内外に散在する美術館・博物館の浮世絵情報に対して統合的なアクセスを実現

同作者や同画題等の関連情報へのリンクを動的に生成

研究事例

多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

- 作品名の情報を用いて，異言語の浮世絵データベースから同一作品を推定

ハイパーリンクの利点と欠点

- Web上のニュース記事などで，読者が知らない専門用語が出てくることがある
  - その用語の説明へのリンクがあれば，記事の理解に有用
  - リンクは手動で付与する必要があり，手間がかかる
  - ハイパーテキストの利点が十分に活かされていない

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病院（ともに神戸市）は12日、iPS細胞（人工多能性幹細胞）から作った網膜の細胞を、目の難病「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。

人工多能性幹細胞？

加齢黄斑変性？

エンティティリンキングとは

- 文書内の重要語（キーワード）に，それを説明するWikipedia記事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
院（ともに神戸市）は12日、iPS細胞（人工多能性幹細胞）から
膜の細胞を、目の難病「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

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（キリスト教界）」の日本語記事は存在しない

研究事例

多言語Wikipedia記事への言語横断エンティティリンキング

文章中の用語と記事名が異なっていてもリンク可能

ニュース記事

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

言語間リンクがなくても該当記事にリンク

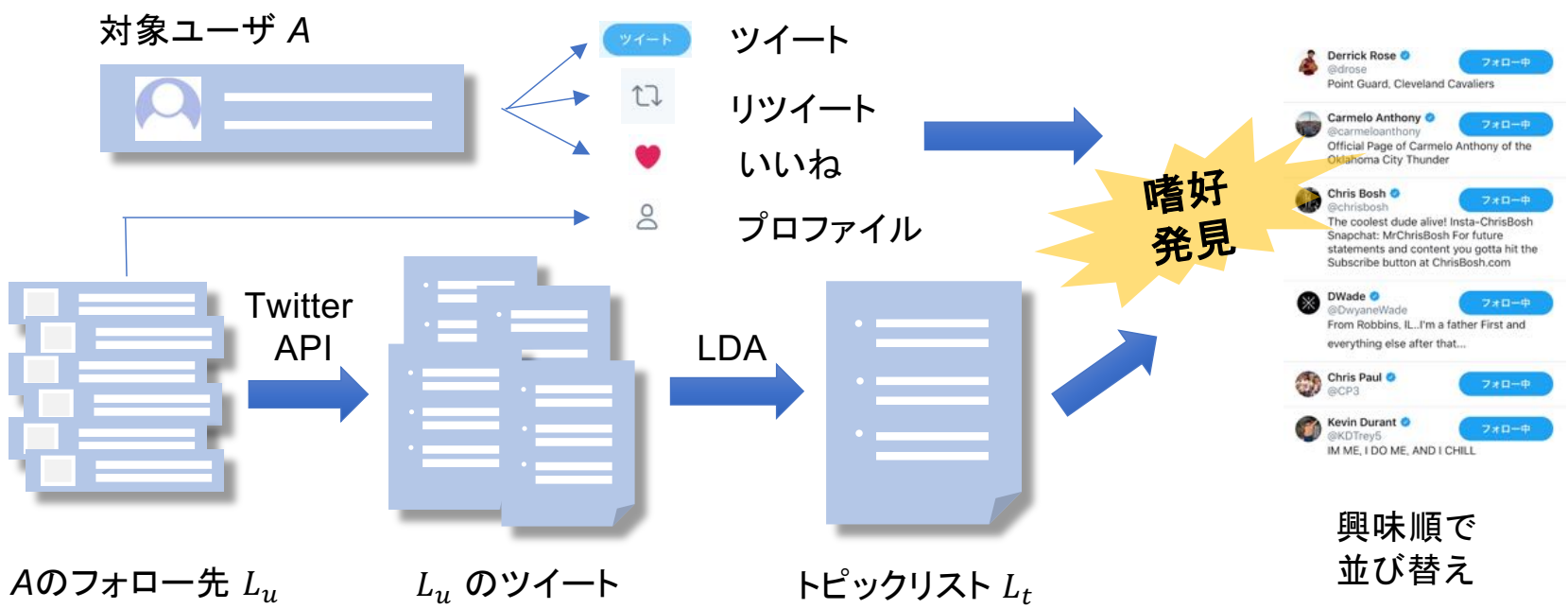
ディジタル図書館のための情報アクセス技術②

ソーシャルメディアのための情報技術

研究事例

Twitterのフォロー情報に基づくユーザ嗜好の推定

- 対象ユーザのフォロー情報および行動履歴から，ユーザの嗜好を予測し，タイムラインを並び替える

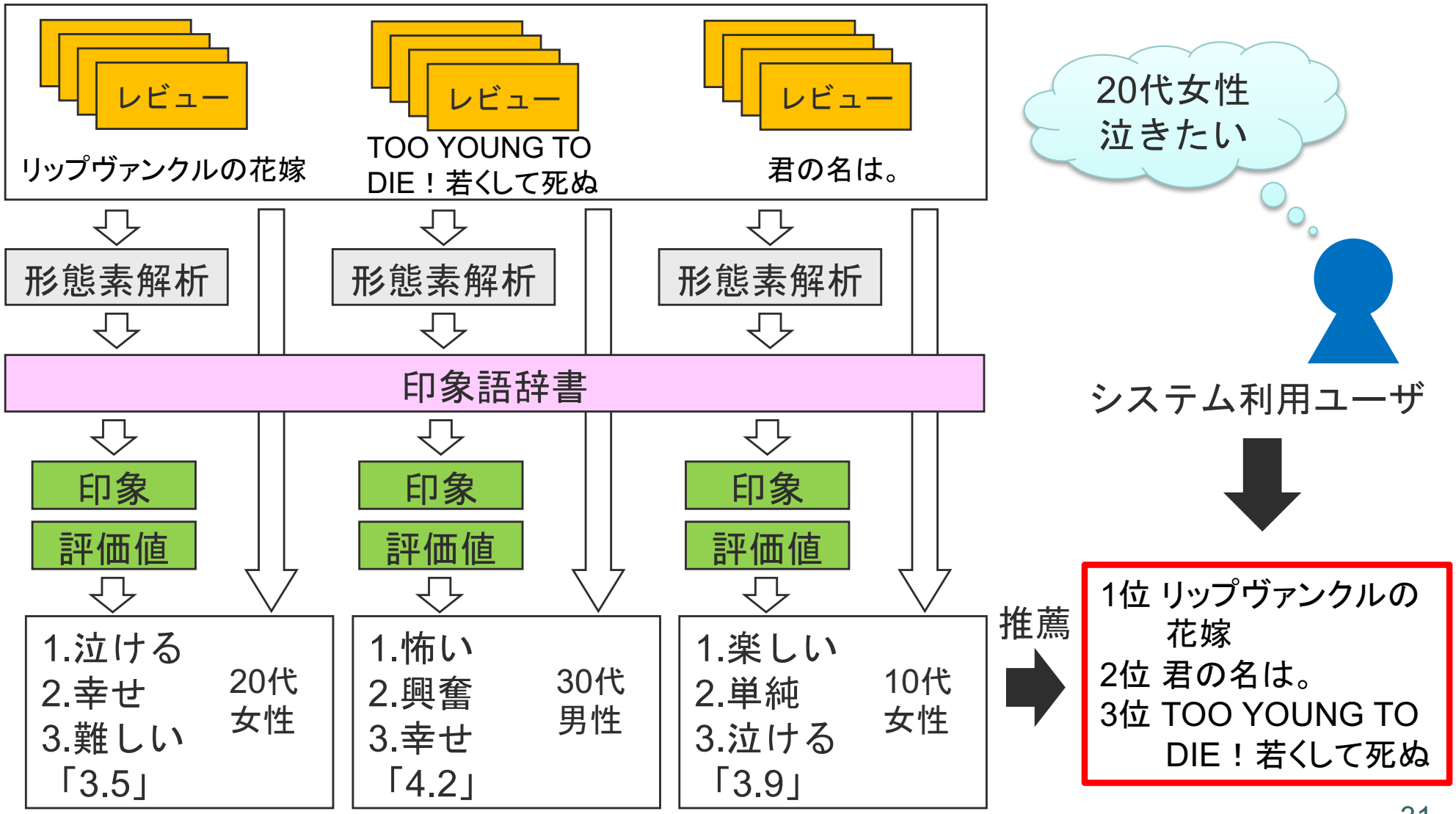


研究事例

Twitterを利用した漫画作品の評判分析

研究事例

映画レビューからの印象推定に基づく映画推薦システム



研究事例

プロ野球の試合ダイジェストの自動生成

- 試合のデータからダイジェストの文章を自動生成
- 利用者の目的に応じて，動的に内容や分量を変更可能
  - 自分が応援している球団向けのダイジェスト
  - 媒体に応じて文章長を変更
    - スマートフォン，Webページ，新聞など

ディジタル図書館のための情報アクセス技術③

人文情報への情報技術の適用

ディジタルヒューマニティーズとは？

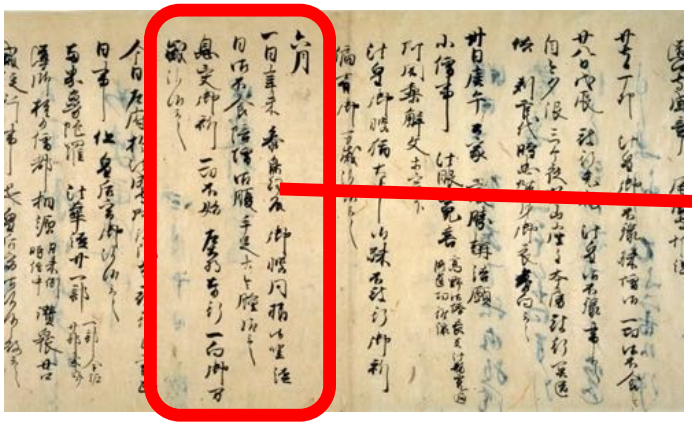
- 近年，**ディジタルヒューマニティーズ**と呼ばれる研究分野が盛んになってきている
  - 人文系の研究で扱う情報のディジタル化が進んできていることが背景
- コンピュータを活用した人文学の研究全般を指す学際的な学問分野
  - 日本では「**人文情報学**」とも呼ばれる
  - 人文学，情報学，図書館学，博物館学など様々な分野の研究者が関わっている

ディジタルヒューマニティーズの研究分野

- 研究に用いる資料のディジタル化
  - 古文書，絵画，遺跡，遺物，その他有形・無形の文化遺産，ファッション，ゲーム，etc.
- ディジタル化した資料の研究での活用
  - **資料の保存・整理**
    - ディジタルアーカイブの構築
  - **統計的な分析**
    - データマイニング，テキストマイニング
  - 従来の人文学研究の方法論にとらわれない新たな研究手法の構築

古典史料のテキストマイニング

- 古典史料のディジタルアーカイブは，ディジタル画像での提供が一般的
- 一部の史料は，電子テキスト化されつつある
  - 電子テキスト化することで統計的分析が可能
- 「**テキストマイニング**」とは，テキスト情報を統計的に分析し，隠れた知識を発見する技術



六月
一日　辛未　参鳥羽殿、御悩同様御坐、経日御不食陪増、御腹手足等令腫給云々、息災御祈一切不始、左大将奉行、一向御万歳沙汰云々、

研究事例

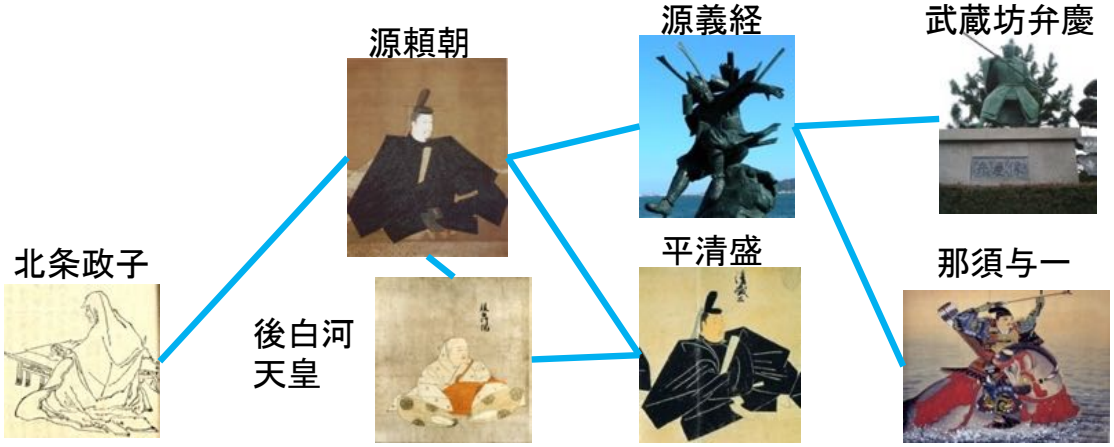
古典史料からの人物表現の自動抽出

平清盛

実名　別名　役職

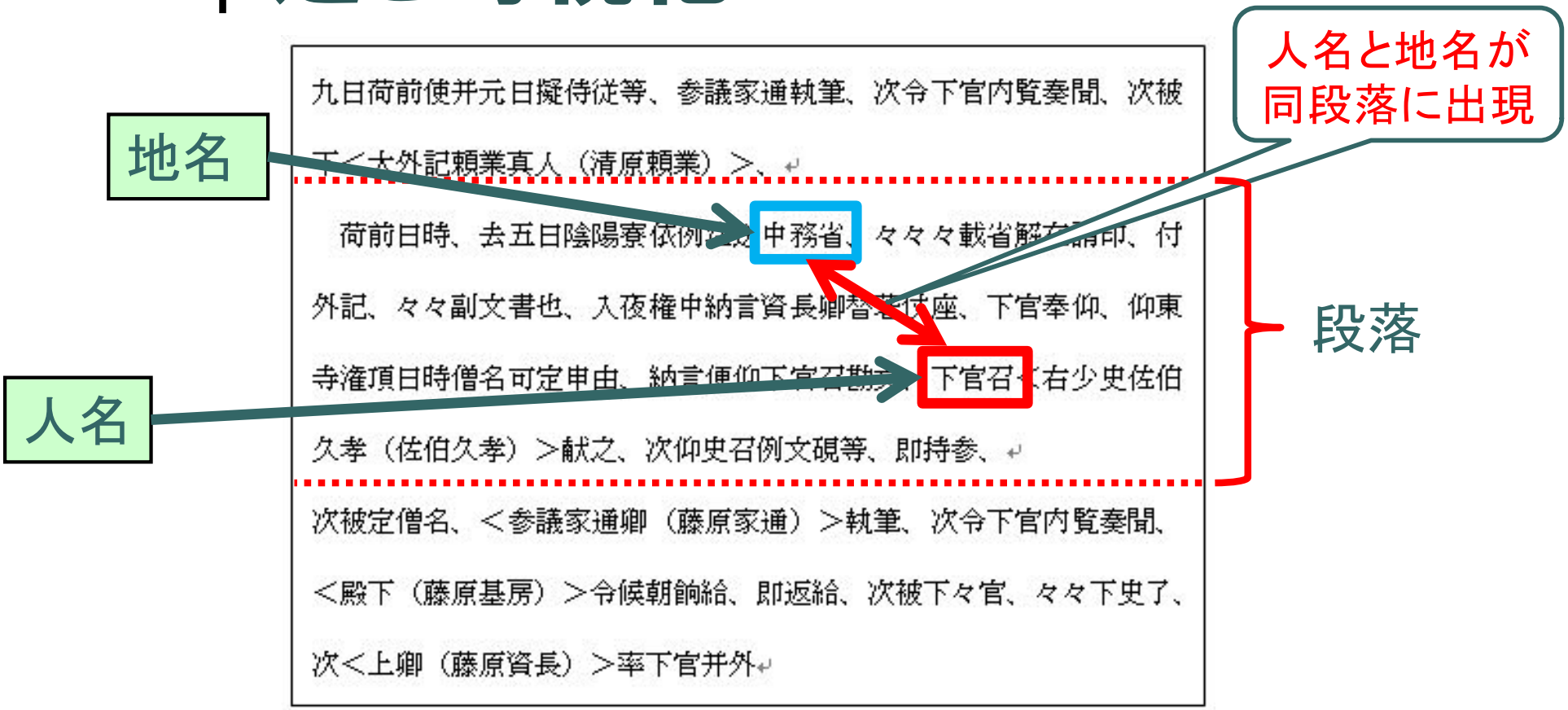
鶏鳴清盛朝臣、義朝、義康等、軍兵都六百余騎発向白河、
六波羅亭、入道御座宰相中将被装束、次供奉行列被渡南庭、
上皇女御殿御同車、自御前渡御馬場桟敷、
前太政大臣以下歩行扈従、

人間関係の可視化などに役立つ



研究事例

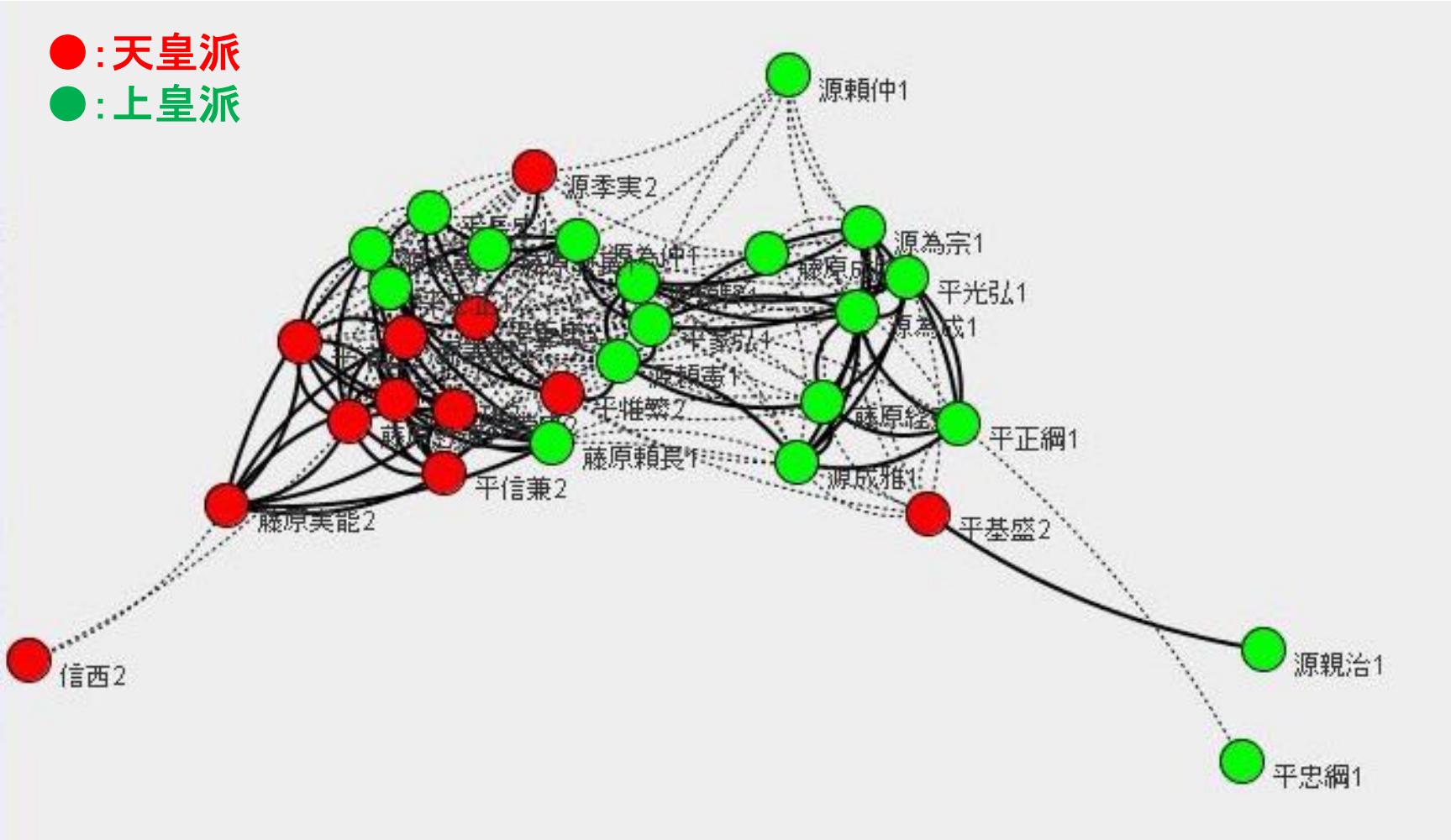
古典史料からの人物関係の推定と可視化



- 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

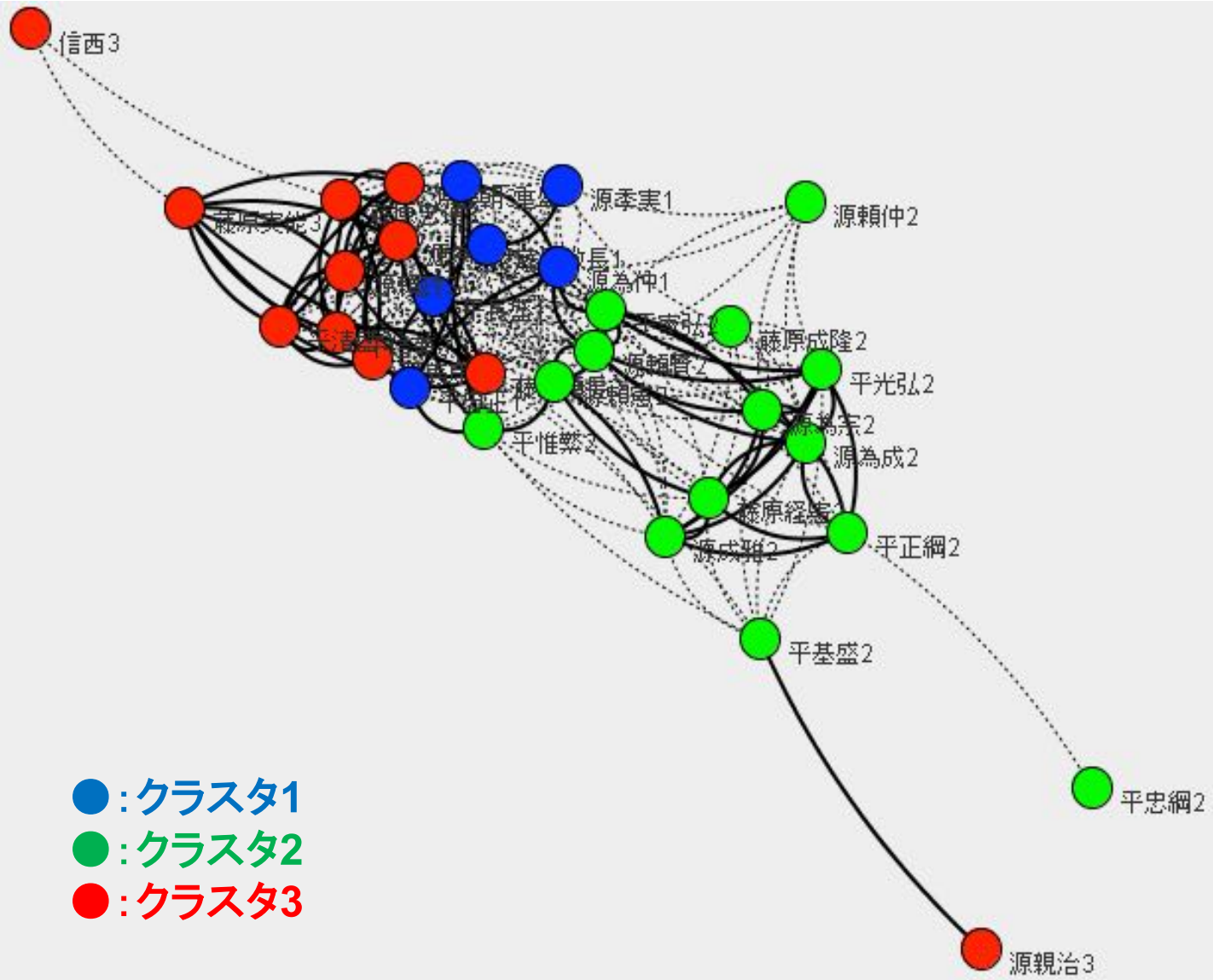
研究事例

保元の乱における天皇派と上皇派の人物関係の推定結果①



研究事例

保元の乱における天皇派と上皇派の人物関係の推定結果②

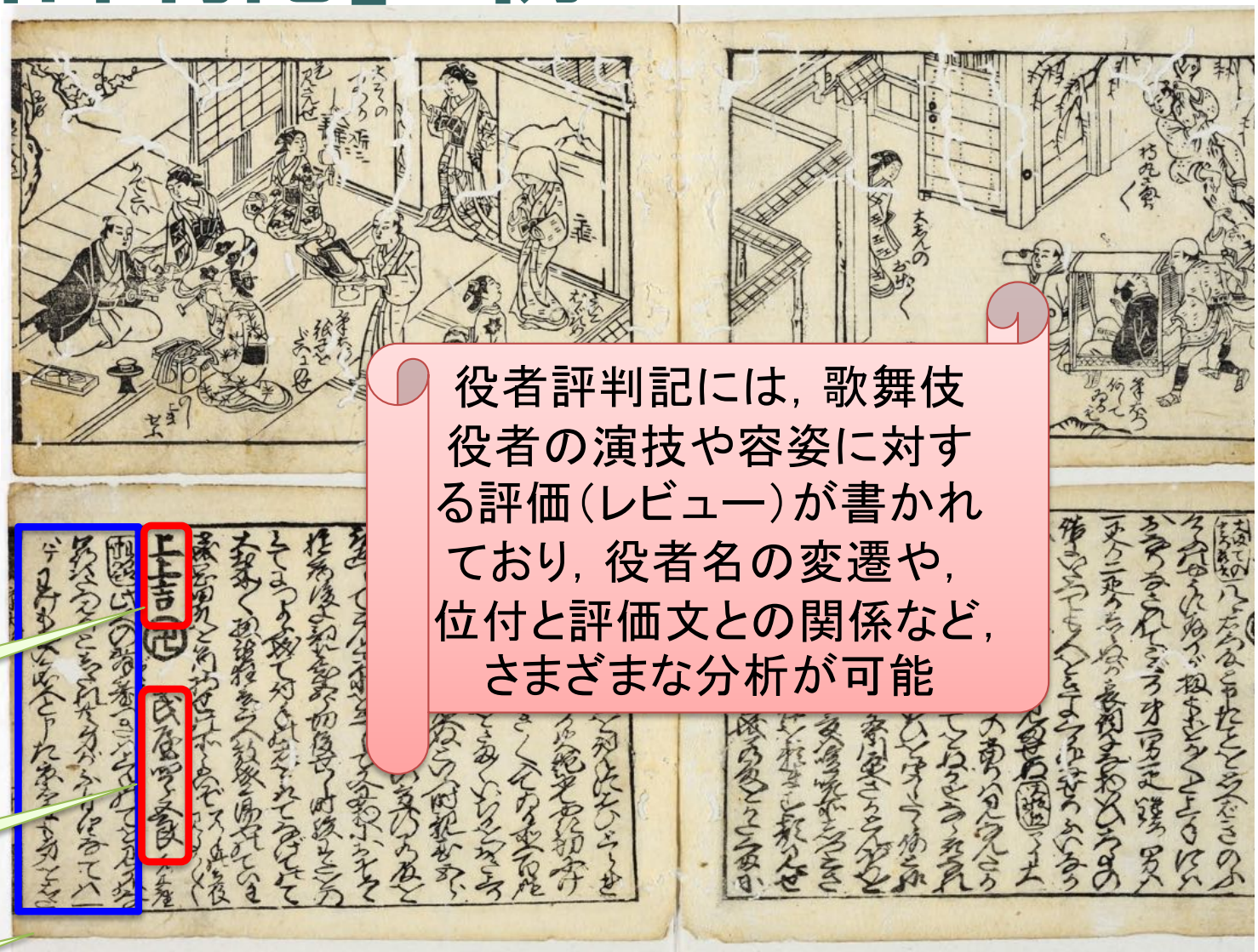


- 天皇派・上皇派が，ほぼ正しく分類されている
- その中間に派閥が形成されているように見える
- デモ

研究事例

古典史料テキストからの情報抽出「役者評判記」の例

役者評判記でレビューされている歌舞伎役者



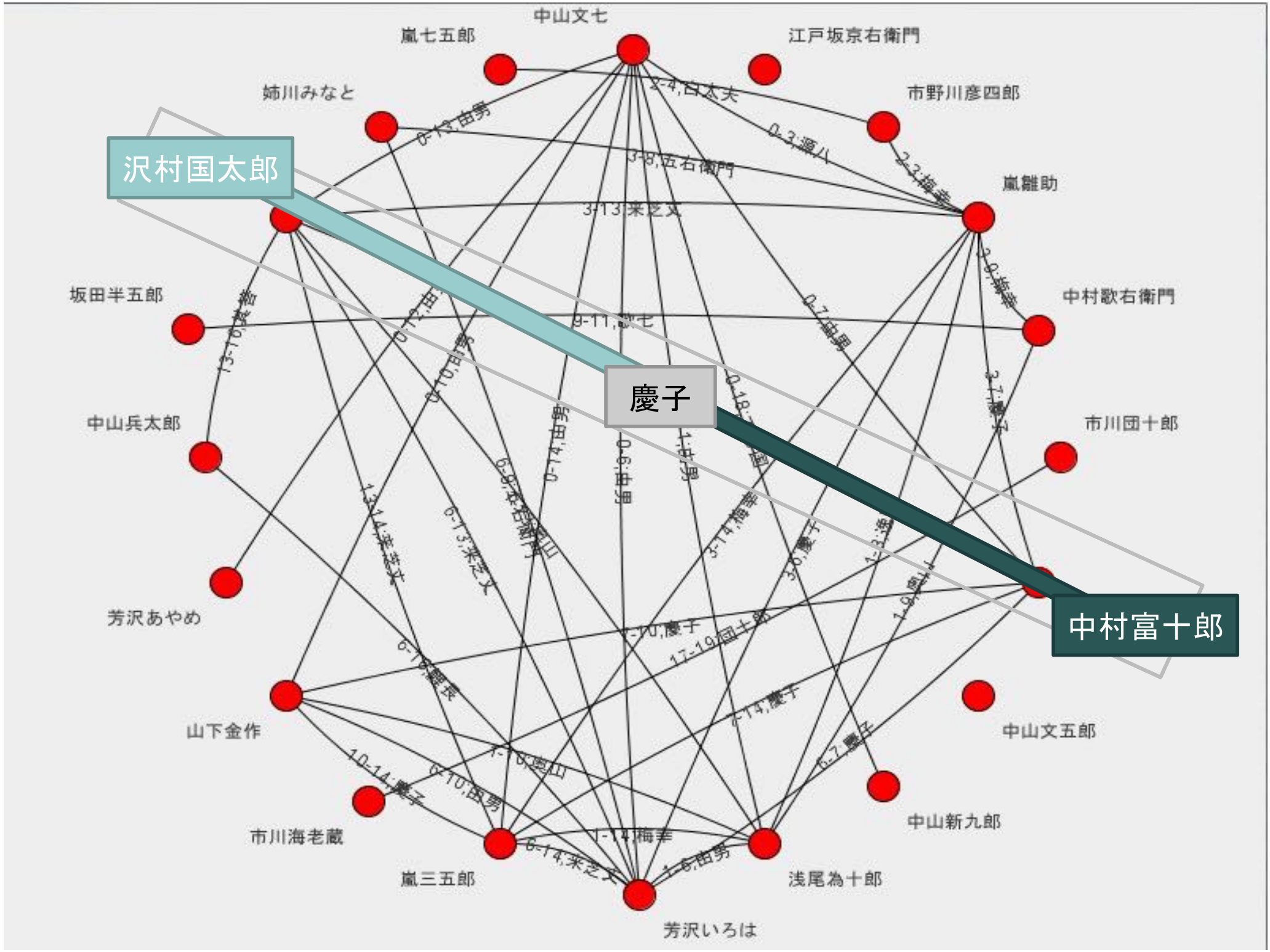
役者評判記には，歌舞伎役者の演技や容姿に対する評価（レビュー）が書かれており，役者名の変遷や，位付と評価文との関係など，さまざまな分析が可能

位付（役者の評価）

役者名

評価文

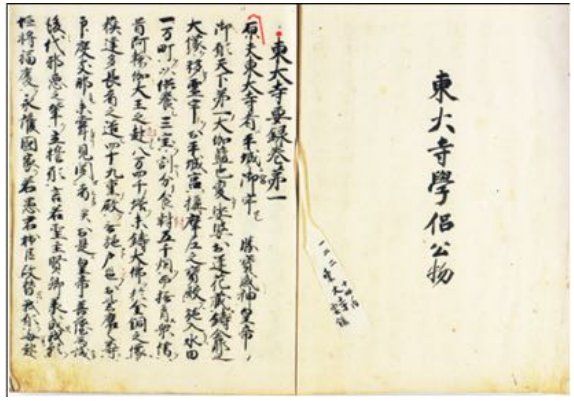
役者評判記『役者大極舞』
立命館大学アート・リサーチセンター（http://www.arc.ritsumei.ac.jp/index.html）所蔵


中山文七
嵐七五郎
江戸坂京右衛門
姉川みなと
市野川彦四郎
沢村国太郎
嵐雛助
坂田半五郎
中村歌右衛門
慶子
中山兵太郎
市川団十郎
芳沢あやめ
中村富十郎
山下金作
中山文五郎
市川海老蔵
中山新九郎
嵐三五郎
浅尾為十郎
芳沢いろは

研究事例

『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。

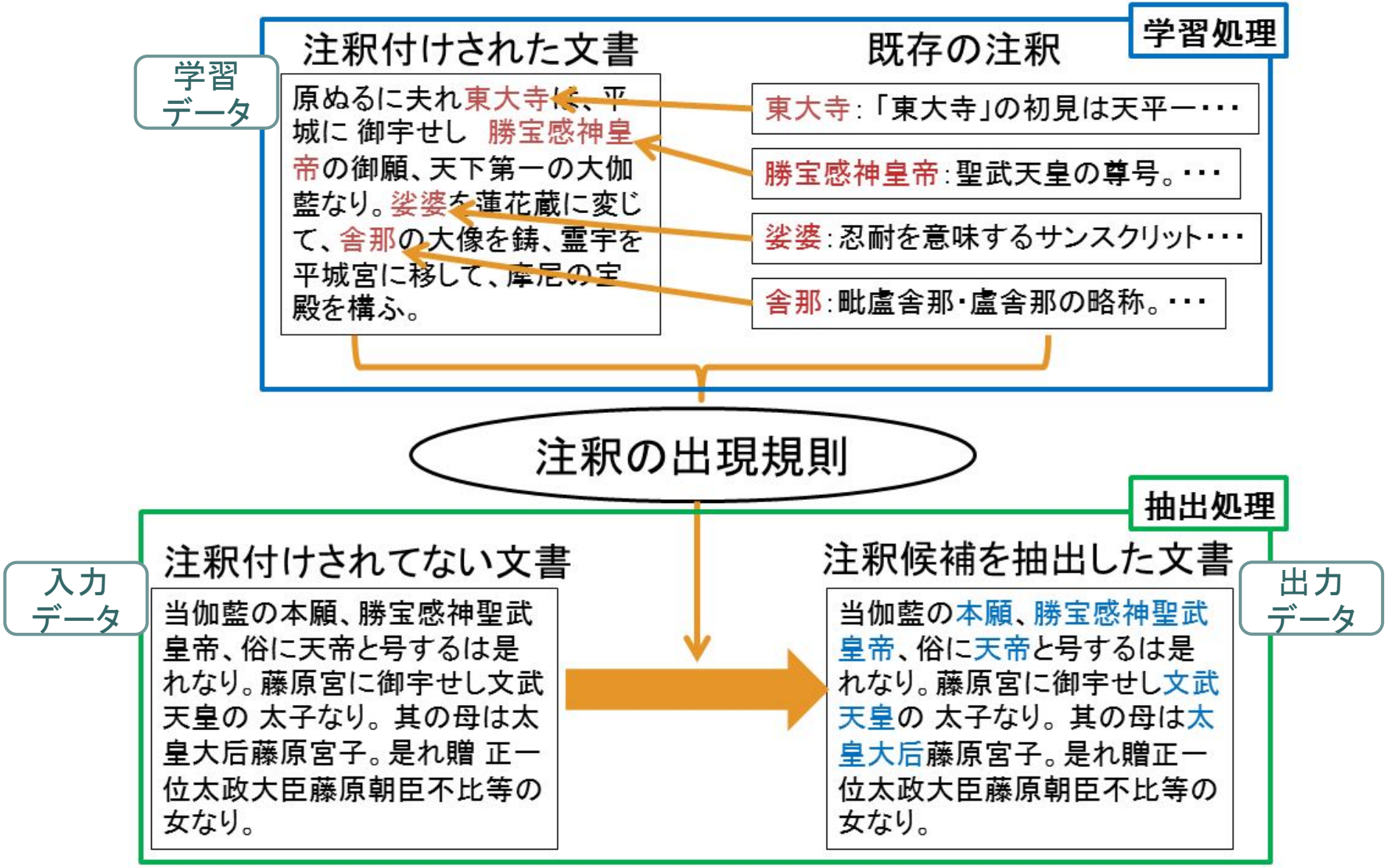
注釈付き書き下し文

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

研究事例

機械学習を用いた注釈候補の自動抽出



研究事例

篆書体による蔵書印の文字認識および検索

- **篆書体**：紀元前200年頃に作られた漢字の書体
  - 現代でも印鑑などに用いられる
- **蔵書印**：所蔵を明示するために蔵書に捺した印影
  - 篆書体が多く使われる

- 当研究室で開発した「**白川フォント**」を学習データとし，機械学習による文字認識を行う

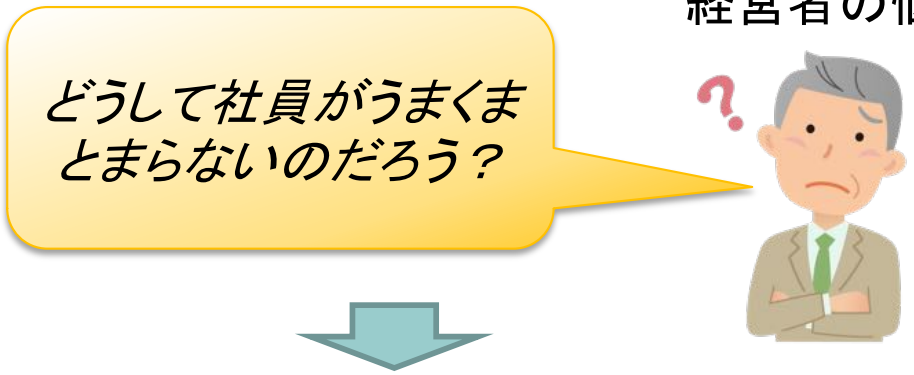
研究事例

経営哲学に関する講演録を用いた質問応答システム

講演録をテキスト化し，
アーカイブとして保存

経営哲学
講演録アーカイブ

経営者の悩み



講演録アーカイブから，質問の答えになり得る箇所を抽出し，利用者に提示

・・・同時に会社に住んでいる社員の考え方によって社風が決まってくるわけです。ですから、いい考え方をする社員がいる場合には素晴らしい社風になっていきます。・・・

先端社会デザイン概論
調査研究のヒント

- **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
- **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

まとめ

- ディジタル図書館のこれまでの発展の歴史と最近の事例を紹介した
- ディジタル図書館への活用が見込まれる最新の情報技術をいくつか紹介した

- ディジタル情報は爆発的に増大しているが，その有効活用・長期保存の技術は発展途上
- 過去から現在の知識を未来に伝えるために，ディジタル図書館は重要な役割を担っている