第11回　同志社大学「新ビジネスフォーラム」　3月19日（水）13:55～14:40

学術データの大規模分析
～共同研究のパートナー探索に向けて～

同志社大学理工学部
桂井麻里衣

Katsurai Laboratory
Doshisha University
Intelligent Mechanism Laboratory

桂井 麻里衣

同志社大学理工学部インテリジェント情報工学科　准教授
サイエンスコミュニケーション研究センター　センター長

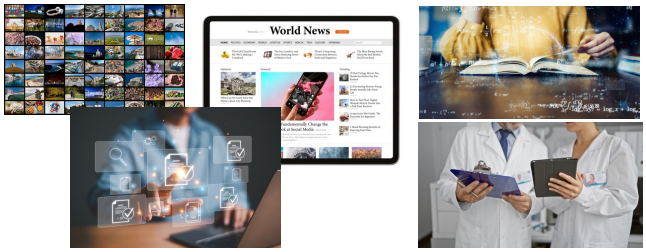
博士（情報科学）
北海道札幌市出身・京都市在住
京田辺キャンパス勤務

https://researchmap.jp/katsurai

知的機構研究室（理工学部 桂井研）

知的創造活動の分析・支援

～画像・テキストコンテンツ、コミュニケーション履歴～



Katsurai Laboratory
Doshisha University
Intelligent Mechanism Laboratory

教員1名　事務補佐員1名
大学院生12名　学部生8名
※2024年度データ

情報検索や推薦への応用を見据え、**企業が所有するデータ**を分析

例：ウェブサービス関連企業

Katsurai Laboratory
Doshisha University
Intelligent Mechanism Laboratory

教員1名　事務補佐員1名
大学院生12名　学部生8名
※2024年度データ

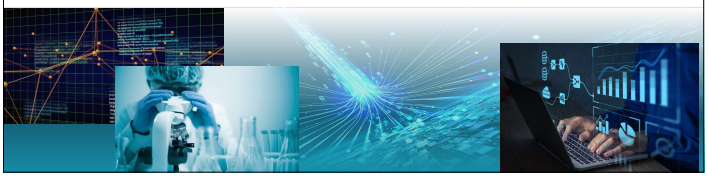
高度な知的創造活動である**学術研究**の分析については
公的予算を継続的に獲得

2017年度～2022年度	科研費　若手研究（B）
2018年度～2019年度	JST ACT-I
2019年度～2021年度	JST ACT-X
2020年度～2024年度	科研費　基盤研究（B）
2025年度～2028年度	科研費　基盤研究（B）

本講演では上記公的予算の成果による**ビジネス応用**の可能性を提案します

Science of Science　～科学の科学～

- 自然科学、計算科学、社会科学の知見や理論を融合させた学際的な新興分野
- 論文、特許、科研費など**大規模な学術データ**に基づく学術研究活動の分析
  …研究者の能力向上、科学全体を高めるシステムや政策の開発に生かす
- 分析したい内容にあわせて情報源を選定



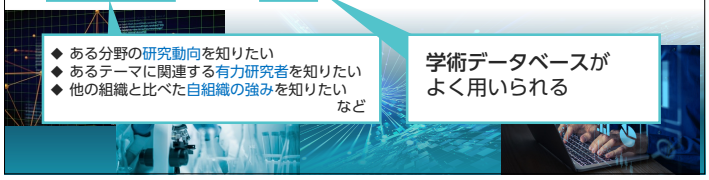
Science of Science　～科学の科学～

6

- 自然科学、計算科学、社会科学の知見や理論を融合させた学際的な新興分野
- 論文、特許、科研費など大規模な学術データに基づく学術研究活動の分析
  …研究者の能力向上、科学全体を高めるシステムや政策の開発に生かす
- 分析したい内容にあわせて情報源を選定

◆ ある分野の研究動向を知りたい
◆ あるテーマに関連する有力研究者を知りたい
◆ 他の組織と比べた自組織の強みを知りたい
など

学術データベースが
よく用いられる



7

STEP1　情報源とする学術データベースを選択

日本語論文の例：

https://cir.nii.ac.jp/

科研費：

https://kaken.nii.ac.jp/

英語論文の例：

https://www.scopus.com/

STEP2　分析対象データを抽出

STEP3　分析手法を適用

STEP4　結果を可視化

8

STEP1　情報源とする学術データベースを選択

日本語論文の例：

https://cir.nii.ac.jp/

科研費：

https://kaken.nii.ac.jp/

英語論文の例：

https://www.scopus.com/

STEP2　分析対象データを抽出

ある分野の国際的な研究トレンドを分析する場合

Scopus　学会誌名をいくつか選択し、以下の情報を大量に収集

・論文タイトル　・キーワード　・抄録

STEP3　分析手法を適用

STEP4　結果を可視化

9

STEP1　情報源とする学術データベースを選択

日本語論文の例：

https://cir.nii.ac.jp/

科研費：

https://kaken.nii.ac.jp/

英語論文の例：

https://www.scopus.com/

STEP2　分析対象データを抽出

STEP3　分析手法を適用

テキスト処理　統計分析・機械学習など

STEP4　結果を可視化

10

STEP1　情報源とする学術データベースを選択

日本語論文の例：

https://cir.nii.ac.jp/

科研費：

https://kaken.nii.ac.jp/

英語論文の例：

https://www.scopus.com/

STEP2　分析対象データを抽出

STEP3　分析手法を適用

STEP4　結果を可視化

情報検索インタフェースなどを構築、
非専門家でも解釈しやすい形で提示

嘉本＆桂井，JSAI2025より抜粋

11

現状の問題点と、当研究室の強み

既存の学術データベースは全ての研究成果を網羅しているとはいえない

例えば…

著者の所属情報や
抄録の抜けが多い

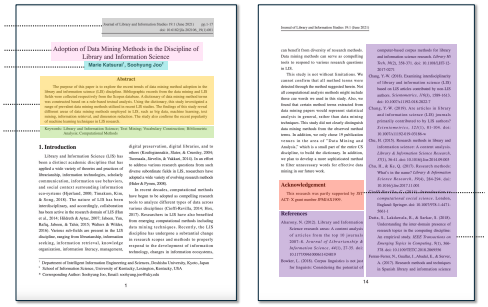
収録論文誌を厳選
（国際会議は抜けが多い）

そこで当研究室では、**ウェブ上の論文PDFからの情報抽出に着手**

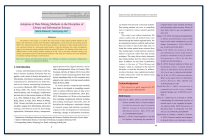
…既存のデータベースにない情報の分析・可視化を実現

論文PDFの要素

論文誌名
タイトル
著者名
抄録
キーワード
本文
謝辞
参考文献



Marie Katsurai and Soohyung Joo, "**Adoption of Data Mining Methods in the Discipline of Library and Information Science,**" Journal of Library and Information Studies, vol. 19, no. 1, pp. 1–17, June 2021.



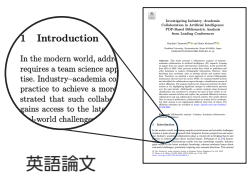
論文誌名 ： Journal of Library and Information Studies
タイトル： Adoption of Data Mining Methods in the Discipline of Library and Information Science
著者名：Marie Katsurai and Soohyung Joo
キーワード：Library and Information Science; Text Mining; Vocabulary Construction; ...
本文：1. Introduction Library and Information Science (LIS) has been a distinct academic discipline ..

PDFから要素ごとにテキストを抽出することが重要

成果紹介①

日本語論文PDFからの情報抽出
～機械学習モデルの構築～

オープンな論文PDF解析器は英語モデルのみ



英語論文

日本語論文

問題点

既存の英語モデルを日本語論文PDFに適用すると要素認識に失敗してしまう

当研究室での取り組み

- 論文PDF解析器を日本語に対応させるための機械学習体制を確立

機械学習用データ構築チーム

<front>学術情報推薦における</front>
<front>概要</front>
<front>研究者への</front>
<body>1 はじめに </body>
<body>インターネット上では</body>
<page>- 3322 -</page>
人手で注釈を付与


データベース化


モデル訓練

当研究室で蓄積したノウハウを利用

- 日本語論文PDFに対し、**従来モデルよりも遥かに高い抽出性能を達成**
  - 論文PDFが公開される国内学会について、いちはやく発表内容を分析することが可能に
- 本成果は言語処理学会第31回年次大会（NLP2025）で発表
  - 「日本語論文に特化したPDF文書解析器の構築と性能評価」（嘉本，梅澤，長尾，桂井）

本研究成果のビジネス面での応用可能性

- 組織に蓄積されている大量のPDFファイル…データベース化のコストが問題
- 機械的にテキストを抽出できると様々な分析に役立つ

PDF
企画書、報告書
仕様書など

テキスト抽出

分析手順（後述の成果③も応用）


文書作成者

・これまでの仕事内容
・保有スキル
などの特徴を機械的に表現

スキル面で類似

経歴類似　同部署経験

社員間の関連性算出へ

成果紹介②

研究トレンド分析
～人工知能（AI）分野の産学連携状況を解明～

分析の背景

- 組織間の協力関係を知ることは戦略上重要
- 昨今競争の激しいAI分野では民間企業の研究が台頭

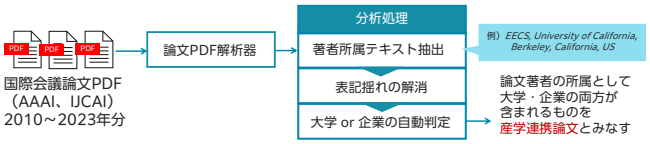
大学
利益にとらわれない研究
学生や教員との自由な連携

企業
データや計算機が豊富
アカデミアでできない
実験が可能

問い AI分野では国際的にどのような産学連携が行われているか？

国際会議論文PDFから著者情報を抽出…動向調査へ

分析フレームワーク



抽出した産学連携論文に基づき、以下に着目して分析を実施

- 産学連携論文を多く発表している組織はどこか？
  →論文数トップ10の機関を表示
- どのような連携関係が強力となっているか？
  →組織間の共著関係をネットワークとして可視化

産学連携論文数トップ10（アカデミア）

- 20,549本の論文分析結果
- 上位9位までが中国の大学

	大学名	論文数
1	浙江大学	321
2	清華大学	303
3	北京大学	265
4	中国科学技術大学	234
5	中国科学院大学	207
6	上海交通大学	179
7	南洋理工大学	143
8	中国科学院	126
9	北京航空航天大学	124
10	Carnegie Mellon University	119

データ出典：Investigating Industry–Academia Collaboration in Artificial Intelligence: PDF-Based Bibliometric Analysis from Leading Conferences（Yamauchi & Katsurai、ICADL2024）

産学連携論文数トップ10（企業）

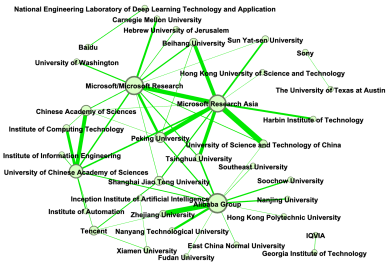
- ビッグテックが多くランクイン
- やはり中国の企業が目立つ

	企業名	論文数
1	Microsoft Research Asia (MSRA)	612
2	Alibaba Group	544
3	Microsoft Research/ Microsoft	407
4	Tencent	203
5	Meta	98
6	Huawei Technologies	57
7	Google	57
8	Baidu	44
9	Jingdong	39
10	Amazon	21

データ出典：Investigating Industry–Academia Collaboration in Artificial Intelligence: PDF-Based Bibliometric Analysis from Leading Conferences（Yamauchi & Katsurai、ICADL2024）

組織間の共著ネットワーク

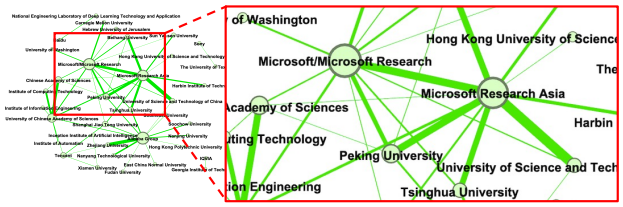
- 共著論文数が6本以上ある関係を可視化



データ出典：Investigating Industry–Academia Collaboration in Artificial Intelligence: PDF-Based Bibliometric Analysis from Leading Conferences（Yamauchi & Katsurai、ICADL2024）

組織間の共著ネットワーク

- 中国の組織間で活発に産学連携論文を発表していることがわかる



データ出典：Investigating Industry–Academia Collaboration in Artificial Intelligence: PDF-Based Bibliometric Analysis from Leading Conferences（Yamauchi & Katsurai，ICADL2024）

まとめと成果③への接続

- AI分野の有名国際会議では中国の産学連携に勢いがあることを示した
- これらの成果は国際会議ICADL2024フルペーパーで発表
  - Investigating Industry–Academia Collaboration in Artificial Intelligence: PDF-Based Bibliometric Analysis from Leading Conferences（Yamauchi & Katsurai）
- 本研究と同様の分析をAI分野以外でも行うことは政策的にも重要と考えられる

日本国内の産学連携を活性化させるにはどうしたらよいか？

→ 国内研究者の情報整備＋**分野横断的に研究者を発見**する仕組みづくりへ（成果③）

成果紹介③

研究者の専門内容解析
～研究者検索・推薦への応用～

学術情報システムによる研究者検索

- 膨大な学術情報から分野の専門家を見つけるには**キーワード検索**が必要
  - 検索対象を明確に言語化しなければならない
  - ある程度知識がなければキーワードの選定が困難

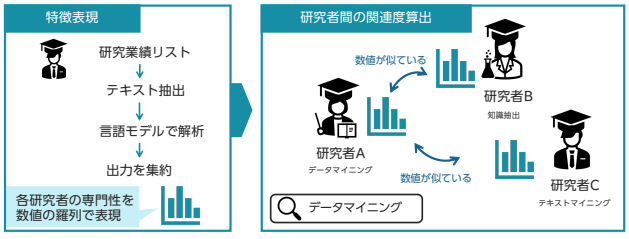
- キーワード検索によって分野の有名研究者にたどりつけたとしても…

すでに多数のプロジェクトに関係しており共同研究を打診しにくい可能性がある

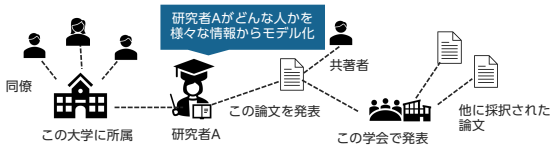
- 研究者Aと専門が近いにも関わらず、**そのキーワードを普段使っていない研究者は検索結果に表示されない**という問題もある

当研究室での取り組み

- 従来のキーワード検索への付加的機能として、**関連研究者の発見技術**を開発
  - 国内研究者の研究業績を**言語モデル**に入力、計算機上で専門性を特徴表現
  - 専門特徴の近さを計測し、類似している研究者を提示

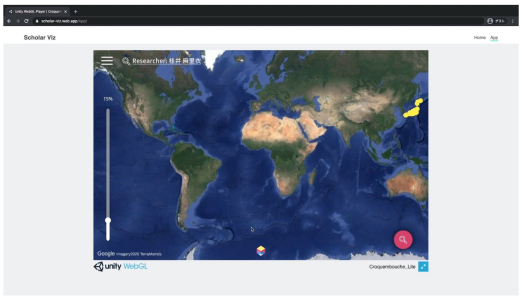


- 最近は業績情報のみならず、所属、発表場所（コミュニティ）など、**多様な情報も一緒に埋め込む方法**を構築中
  - 下図のようなグラフ情報を整備する方法も並行して開発



- これらの成果は国際会議SDP2024と国内学会NLP2025で発表
  - Researcher Representations Based on Aggregating Embeddings of Publication Titles: A Case Study in a Japanese Academic Database（Nagao & Katsurai）
  - 「学術情報推薦におけるグラフ構造の有効性検証」（長尾、桂井）

研究者検索インタフェースの開発



A Novel Researcher Search System Based on Research Content Similarity and Geographic Information
(Takahashi, Tango, Chikazawa, & Katsurai, ICADL2020)

まとめ

- 本講演では学術データの大規模分析に関する最新の成果を3つ紹介
  - 情報抽出、トレンド分析、パートナー探索技術
- これらの成果で培った知見はビジネス面への応用が見込める
  - 社内PDFのデータベース化、社員のスキルモデリング、関連性提示
  - 任意の研究分野の共同研究関係を可視化→研究企画戦略に生かす
- 様々な種類のウェブデータでの研究経験を生かした技術開発
  - 具体的な研究トピックは研究室ウェブサイトをご覧ください

研究室ウェブサイト

メールアドレス

katsurai@mm.doshisha.ac.jp

今回ご紹介した学術データ分析に限らず、**創造・嗜好データの利活用のあり方**を幅広く研究しています。新たなコラボレーション機会を楽しみにしております！

サイエンスコミュニケーション研究センター

- 学術情報分析、メディア研究、各種AIの開発に係る教員で構成しています。
- 2025年春から本格始動予定です。センター宛のご依頼もお待ちしています！

2025年3月時点の構成員		
理工学部	准教授	桂井 麻里衣（センター長）
社会学部	准教授	阿部 康人（副センター長）
理工学部	准教授	木村 共孝
文化情報学部	准教授	飯尾 尊優
文化情報学部	助教	阿部 真人
文化情報学部	准教授	中西 義典
免許資格課程センター	教授	佐藤 翔
免許資格課程センター	助教	山口 洋介
ハリス理化学研究所	助教	桝 太一
生命医科学部	教授	野口 範子
研究開発推進機構	嘱託研究員	下原 勝憲

主に30~40代の研究者で構成
ウェブサイトは5月頃公開予定