卒業研究配属選択支援システムの開発

二村 彰[1,a)]　峯 恒憲[2,b)]

概要：教員の研究内容についての情報や，研究に関する専門知識を十分に持たない学部学生にとって，卒業研究の指導教員選択は困難な作業であり，多くの学生が少ない情報を頼りに選択を行っている．そこで本研究は，九州大学工学部電気情報工学科での卒業研究の指導教員選択を例として，指導教員選択支援システムを提案する．提案手法は，卒業研究選択の材料として提供される研究テーマ集や Web などの情報をトピックモデリングに基づき分析を行い，研究分野の関連の強さで教員をクラスタリングした結果を視覚化する．これにより，興味あるトピックに関連する指導教員の絞り込みを可能とするだけでなく，教員間の研究分野での関連性についても把握可能とする．実験の結果，提案手法の有効性を確認した．

Development of a Laboratory Assignment Support System for Graduation Research

1. はじめに

卒業研究配属は，学生にとって，極めて重要な選択である．研究内容は研究室によって異なり，それが学生に大きく影響を与える．特に，大学院を卒業した後の就職先も各研究室で異なるので，教員選択は卒業後の就職先にも大きく関わっている．

しかし，指導教員選択は重要な選択であるとともに，学生にとって，非常に難しい選択である．多くの学部 3 年生は，専攻教育科目の授業を受けているものの，専門的な研究に関しては未経験であるため，自分がどのような研究を行いたいのか，または，自分がどの分野に興味があるのか，明確に決めることができない．また，学生に提供される各教員の研究内容に関する情報は，九州大学工学部電気情報工学科の場合，主に卒業研究テーマ集のみであり，十分とは言えない．さらに，研究内容に含まれているキーワードは専門的であり，学生にとって具体的に理解しづらい．そのため，教員のもとでどのような研究ができるのか，明確に把握できないことが多く，教員選択は難解なものとなっている．

そこで，本研究では「多くの学生は，卒業研究配属の選択が難しいと感じている」ことを問題提起とする．ここで，学生が卒業研究配属の選択の際に，必要な情報としては「教員を専門分野，研究テーマにおいて分類した情報，特に自分が興味のある分野の研究を行っている教員情報」「自分が興味ある教員と関連性が高い研究を行っている教員情報」が重要であると考えている．以上の点を踏まえ，本研究では学生が卒業研究の配属選択の際に感じる困難さを緩和することを目的とし，卒業研究配属選択支援システムを開発する．具体的には，学生が確認できる情報源から，教員の研究内容や研究室情報を収集し，Latent Dirichlet Allocation（LDA）を用いて分析を行い，得られたトピック分布からクラスタリングを実施する．その結果を Web アプリケーション上で可視化を行う．これにより，学生が興味あるトピックに関連した教員や，興味ある教員と関連性が高い教員を直感的に理解することを支援し，「卒業研究配属の選択が難しい」という課題の緩和に貢献できると期待される．本研究の主な貢献は以下の通りである．

- 九州大学工学部電気情報工学科の学生向けの卒業研究配属選択を支援する Web アプリケーションの開発
- LDA を活用した教員の研究分野の分析とそのクラスタリング
- クラスタリング結果の可視化による教員選択支援
- 提案手法の定量的評価

2. 関連研究

本研究の関連研究として iTOChat [1] [2] が挙げられる．

[1] 九州大学工学部電気情報工学科
[2] 九州大学大学院システム情報科学研究院
[a)] futamura.shou.710@s.kyushu-u.ac.jp
[b)] mine@ait.kyushu-u.ac.jp

iTOChat は，ChatGPT を利用した高校生向けの九州大学工学部の学科推薦を行う LINE Bot である．高校生は専門性が培われておらず，全ての学部学科について理解することは難しいという課題があるため，対話形式での学科推薦を行う手法を採用している．一方，本研究では九州大学工学部電気情報工学科の卒業研究配属を支援するシステムの開発である．工学部の各学科では研究内容が大きく異なり，学科の数は少ないが，各教員の研究内容は類似性が見られることが多く，人数も多い．このため，ChatGPT による会話形式での推薦手法を用いると，精度の高い推薦を行うことが難しくなる可能性がある．そのため，本研究では，Latent Dirichlet Allocation（LDA）[3] を用いて分析を行い，教員をクラスタリングし，その可視化を行うことで，学生が直感的に教員間の研究分野での関連性を把握できる手法を採用している．

可視化手法については，様々なものが提案されている内容的な類似性に基づく科学技術マップの開発による，研究分野の構造を直感的に把握できる仕組みの構築 [4] や，科学研究費助成事業データベースを用いた研究機関のネットワークの可視化 [5]，論文データをトピックマップとした可視化する手法 [6] 等が挙げられる．ここで挙げられた関連研究の手法は，主に論文データを分析対象として扱っており，目的は研究分野の構造の理解，研究機関間の連携の促進，新たな研究テーマの発見となっている．それに対して，本研究の目的は，学生の卒業研究配属の選択の困難さを緩和することである．論文は，ある特定の分野に特化し，その分野の専門家を対象としている．そのため，論文データを利用して教員をグループ分けした場合，一つのグループ内の人数が少なく，グループ数が多くなり過ぎたり，抽出されたキーワードも専門的になり過ぎて，分析結果を可視化した時，学生が十分に理解できない可能性がある．

そこで，本研究では，卒業論文テーマ集や公式ホームページ等，学生に提供されている情報を分析対象とする．これにより，学生にとってわかりやすい結果の提供を行うことができるだけでなく，卒業論文テーマ集や公式ホームページ等，学生に提供されている情報と提案手法が学生に提供する情報で比較し，学生の卒業研究配属選択における有用性を評価することができる．

3. 提案手法

本研究では，以下の手法から卒業研究配属選択支援システムの開発を行う．

3.1 研究情報の収集

教員を研究内容でクラスタリングするために，各教員の研究内容を収集する必要がある．そこで，今回の研究では，卒業研究配属を選択する際に学生に提供されている以下の 3 つの主要な情報源からデータを収集する．

(1) **九州大学 大学院システム情報科学府の研究室紹介ページ** [7]
各研究室の概要や研究内容に関するキーワードが記載されている．各教員の所属研究室を特定し，所属している教員の研究内容を把握する．

(2) **総合 Web システム 卒業研究テーマ集** [8]
大学院システム情報科学府に所属している教員の研究テーマをまとめた PDF と研究内容に関するキーワードが取得できる．

(3) **九州大学 研究者情報データベース** [9]
研究室紹介ページやテーマ集に掲載されていない教員の研究情報を補完するために利用する．

3.2 クラスタリング手法と評価指標

本研究では，各教員の研究内容に関するキーワードリストからトピックモデリングを行うことで，教員のトピック分布を作成し，その分布を基にクラスタリングを行う手法を採用する．

トピックモデリングには複数の手法が存在するが，結果が他の手法に比べて解釈しやすいという点で LDA を選択する．LDA ではトピック数を手動で設定する必要があるため，評価指標を利用して適切なトピック数を決定する．今回は R の ldatuning パッケージを用いる．本パッケージで計算可能な評価指標である，CaoJuan2009 [10]，Arun2010 [11]，Griffiths2004 [12]，Deveaud2014 [13] の 4 つを用いて分析を行い，最適なトピック数を選定する．

クラスタリング手法については，K-means 法，DBSCAN を採用し，実験で比較検証を行う．

3.3 データ整形

トピックモデリング（LDA）を実行する前に，教員の研究内容に関する，テキストデータの前処理を行う必要がある．LDA の入力としては，トピック数，テキストデータからキーワードを収集し，各キーワードに id を割り当てた辞書，Bag-of-Words で変換した各文書，の三つである．しかし，「同じ意味を表す単語が違う表現で存在していた場合，それらの単語は同じであると認識されずに，誤った学習を LDA が行う」という課題がある．例えば，「人工知能」と「AI」は同じ意味だが，辞書には違う単語として登録されてしまうため，前処理の段階で類義語を統一する必要がある．さらに，「嗅覚」や「食品」等の研究室情報全体からみて一回しか出現していない単語は LDA 実行のノイズになってしまうため，その単語も削除しなければならない．

ここで，卒業論文テーマ集や，研究室紹介ページに記載されているキーワードに加え，研究内容に関するテキストデータから ChatGPT で抽出したキーワードを各教員に関して収集する．収集したキーワードリストに対して，データ整形を行う．この処理は，ChatGPT に全キーワードを

与え，グループ化とノイズになりうる単語の発見を実行させることで，グループ化による類義語の統一とノイズ発見による，単語削除を効率的に行う．

これらの前処理を通じて，トピック分布の精度の評価指標のスコアを向上させ，トピックごとの一貫性を改善し，LDA の学習をより適切に行えるようにする．

3.4 クラスタリング結果の可視化

本研究では，学生が直感的に各教員の研究内容の関係性について理解できるように，クラスタリング結果を Web アプリケーションで可視化を行う．可視化は，ネットワークグラフの描画するライブラリである Cytoscape.js を利用して，クラスタリング結果を表現する．

ネットワークグラフは，ノードとエッジによって構成される．今回はそれぞれ以下のように設定する．

- **ノード**
  - 各教員をノードとして配置し，ノード内に名前を記載する．
  - クラスターを容易に確認できるようにするため，クラスタごとに異なる色を付与する．
- **エッジ**
  - エッジはクラスタ内の教員間とクラスタの重心間のみ表示する．
  - **クラスタ内の教員間**
    - エッジの長さでは，長さが短いほど研究内容が近い，ということが表現できると考えたため，ここで以下の式で各教員間の cos 類似度を計算する．

$$S_{i,j} = \frac{\mathbf{p}_i \cdot \mathbf{p}_j}{\|\mathbf{p}_i\|\|\mathbf{p}_j\|} \tag{1}$$

      ここで，$\mathbf{p}_i$ と $\mathbf{p}_j$ は各教員のトピック分布を表すベクトルである．
    - この cos 類似度の逆数をエッジの長さとすることで，エッジの長さが短いほど，教員間の研究内容が近いことを表現している．
  - **クラスタの重心間**
    - 各クラスタの重心は，以下の式で求め，その重心と一番 cos 類似度が大きいトピック分布である教員を重心とする．

$$\mathbf{c}_k = \frac{1}{|C_k|} \sum_{i \in C_k} \mathbf{p}_i \tag{2}$$

      ここで $\mathbf{c}_k$ はクラスタ k の重心，$|C_k|$ はクラスタに属する教員の数，$\mathbf{p}_i$ は教員 i のトピック分布を表す．
    - エッジの長さは，クラスタ内の教員間と同様に，cos 類似度の逆数に設定する．

図 1 は作成したネットワークグラフの一部である．また，検索機能を設けており，ユーザの入力内容に対応した教員のノードの枠を太くすることで強調表示できるようにしている．

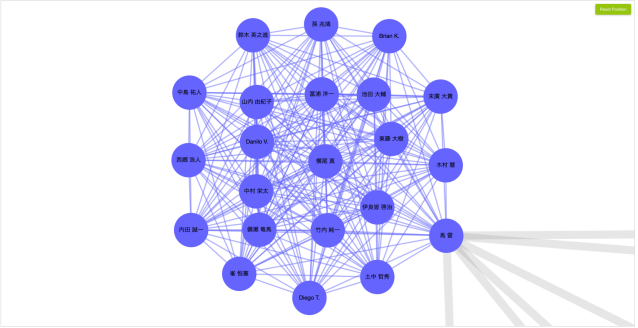

図 1 ネットワークグラフによるクラスタリング (K-means) の可視化の一部

4. 実験方法

4.1 利用するデータ

本研究において，利用するデータは，3.1 から収集した教員情報と研究内容に関するキーワードである．本研究では，88 名の教員のキーワードを収集した．そして，キーワードはデータ整形後，全てで 106 種類となった．

4.2 比較手法

本研究で．比較する手法は以下の通りである．

- クラスタリング手法として，K-means 法と DBSCAN の比較
  Silihouette Score で定量的に評価行った．また，ネットワークグラフでの可視化結果の比較も行った．
- LDA に基づいたクラスタリングと embedding ベースのクラスタリングの比較
  どちらも K-means 法を用い，Silihouette Score で定量的に評価行った．
- データ整形を行う場合と行わなかった場合の比較
  LDA が生成した各トピックに関連づけられたキーワードの確認や，クラスタリング結果を「クラスタの要素数」「専攻の分布」「クラスタ内の評価」という視点で評価を行った．

これらの比較を行い，分析を行った．

4.3 Research Question

以上を踏まえて，本研究では以下の 4 つの Research Question を設定し，実験を通じて回答を行った．

- **RQ1: トピックモデリングの適切なトピック数は幾つなのか？**
  トピックモデリングはトピック数を決めて実行する必要があるため，適切なトピック数を設定する必要がある．そのため，評価指標を利用して最適なトピック数を決定する．

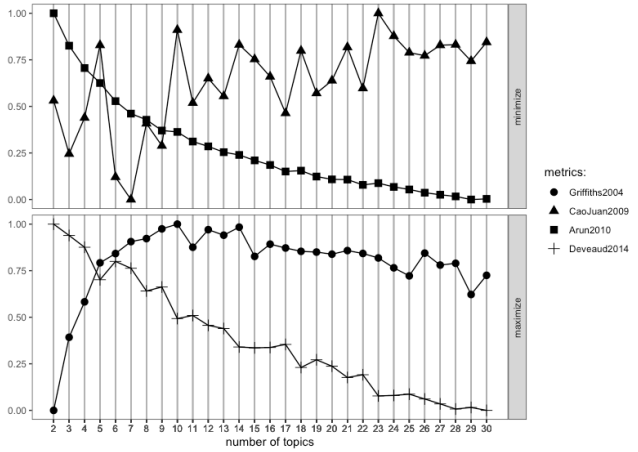

図 2 各トピック数における評価指標

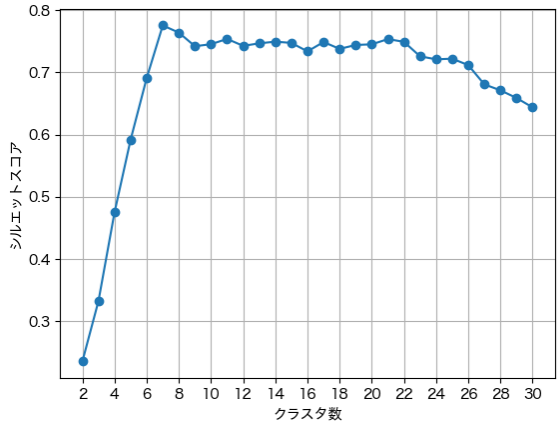

図 3 各クラスタ数における Silhouette Score

- **RQ2: 適したクラスタリング手法は何か？**
  トピックモデリングによって，各教員のトピック分布を計算し，それを基にクラスタリングを行う．ここで，適切なクラスタリング手法は何か，を検証する．また，本研究ではトピックモデリングを介したクラスタリングと，キーワードのエンベディングを直接クラスタリングする手法との比較も行う．
- **RQ3: データ整形はトピックモデリング，クラスタリングの精度の向上に貢献しているか？**
  提案手法の中で，類義語の統一やノイズとなる単語の削除などのデータ整形を行った．この前処理がトピックモデリング結果，クラスタリング結果に与える影響を，データ整形を行わない場合と比較して評価する．
- **RQ4: クラスタリング結果はどう評価できるのか？**
  クラスタリングを行い，その結果を可視化した．ここでは，作成されたクラスタが各教員の研究分野と整合性があるかを検証し，その妥当性を評価する．

5. 実験結果

5.1 R1 の回答

トピックモデリング (LDA) の適切なトピック数を決定する実験を行う．整形した研究内容データを入力として，2 から 30 のトピック数で評価指標を計算した．CaoJuan2009，Arun2010，Griffiths2004，Deveaud2014 の計算結果は図 2 のようになった．CaoJuan2009，Arun2010 は値が小さいほど，Griffiths2004，Deveaud2014 は値が大きいほど，適しているとする評価指標である．Arun2010 はトピック数が大きいほど小さくなっている．CaoJuan2009 はトピック数 7 で最小値を取っている．Griffiths2004 はトピック数 7 から 14 まで大きい値を維持している．Deveaud2014 はトピック数が大きくなるほど値は小さくなっている．以上の結果から，LDA の適切なトピック数は 7 であると考えられる．

5.2 RQ2 の回答

トピックモデリングによって得られた各教員のトピック分布を基にクラスタリングを実施し，結果を可視化した．その際のクラスタリング手法として，K-means 法と DBSCAN の二つを使用し，比較する．さらに，トピックモデリングを行わない embedding ベースのクラスタリング手法と提案手法の比較も行う．

5.2.1 クラスタリング手法の比較

5.2.1.1 K-means 法

K-means 法はクラスタ数を手動で決定する必要があるため，それぞれクラスタ数 2 から 30 まで Silhouette Score を計算した．結果を図 3 に示す．これより，クラスタ数 7 で Silhouette Score が 0.775397 で最大値を取っているので，クラスタ数 7 が最適であると考えることができる．

5.2.1.2 DBSCAN

DBSCAN は eps と min samples という 2 つのパラメータを手動に決定する必要がある．まず，eps の決定には k=3 の k 近傍法を用いた．その結果，データ点間の距離分布において，0.9 から 1.1 の範囲で急激な変化が見られたため，この範囲が適切な eps の候補を考えることができる．

次に min samples を決定するために，min samples を 2 から 7，eps を 0.9 から 1.1 まで変化させた時の Silhouette Score を計算した．結果を図 4（ヒートマップ）に示す．

計算結果より，Silhouette Score が 0.730199 で最大値を取っている，eps=1.09，min sample=2 が最適なパラメータであると考えることができる．そのパラメータで DBSCAN を実行すると，クラスタ数はノイズデータを除くと，8 だった．ノイズ点の割合は 7.95%で，一般的にノイズ点の割合は 5 から 15%は許容されるため，パラメータ設定は妥当であると示された．

結果をまとめると，定量的評価の観点からは大きな差異はないが，K-means 法の方がわずかに良いスコアだった．一方，可視化の観点からは，DBSCAN はノイズ点により，

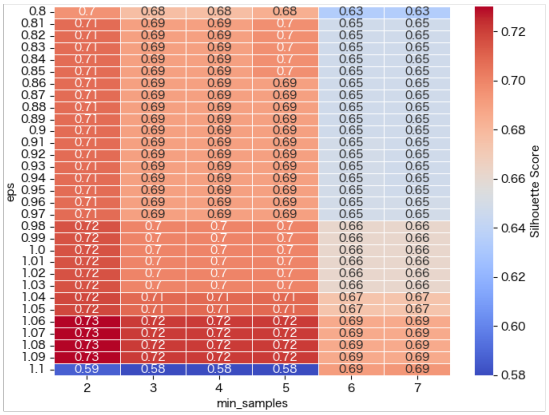

図 4 DBSCAN の各パラメータと Silhouette Score のヒートマップ

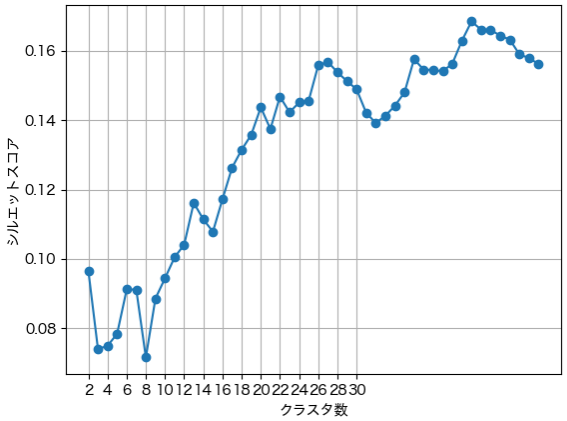

図 5 各クラスタ数における Silhouette Score(embedding ベース)

ネットワークに接続されない点が存在してしまうため，比較すると K-means 法が視覚的にわかりやすいと言える．しかし，ノイズ点は，「この教員は他の教員が行わないような研究をやっている」という情報と捉えることができ，ノイズ点が欠点とは必ずしも言うことはできない．

以上より，開発したシステムでは，定量的評価から，K-means 法を採用した．しかし，どちらの手法がより適しているのかを明確に結論づけることはできなかった．今後の課題として，それぞれのクラスタリング結果をユーザに提供し，どちらの方がより適したクラスターを作成できていたのか，ノイズ点の有無は教員検索においてどう思われるのか，という点についてのユーザアンケートを実施することで，適した手法の評価を行う必要があると考えられる．

5.2.2 embedding ベースと提案手法の比較

embedding ベースのクラスタリングについて提案手法と比較検証する．まず, embedding 生成に関しては, OPENAI 社が提供しているモデル（text-embedding-3-small）を利用した．トピックモデリングで使用したものと同様のキーワードリストを一つの文字列として結合し，この文字列を入力として，embedding を生成した．

この embedding を用いて，K-means 法によるクラスタリングを行った．クラスタ数 2 から 50 までの Silhouette Score を計算した．その結果を図 5 に示す．図からわかるように，どのクラスタ数においても Silhouette Score が 0.2 を下回っている．教員数は 88 なので，クラスタ数 50 以降は教員数に近づくにつれ値は小さくなると考えられる．図 3 と比較すると，明らかに embedding ベースのクラスタリングの性能が低いことがわかる．以上のことから，embedding ベースのクラスタリングと比較すると，提案手法の方が適切なクラスタリングを実行できていることを示せた．

5.3 RQ3 の回答

提案手法では．適切な学習を LDA が行えるよう，類義語の削除やノイズの削除などのデータ整形を行った．ここでは，LDA の結果とクラスタリング結果に基づいて，データ整形を行わない場合と行う場合の比較検証を行う．

5.3.1 LDA の結果

ここでは，LDA が生成した各トピックに関連づけられた上位 6 個のキーワードを確認することで評価を行う．データ整形を行なった場合は表 1，行わなかった場合は表 2 のようになった．

データ整形なしの場合，トピック内にあまり関連性が見られない単語が含まれている部分がある．例えば，トピック 9 には「半導体」と「オークション」，トピック 4 には「SDGs（持続可能な開発目標）」と「無線通信」などがある．また,「コミュニケーション理解（ミーティング分析）」や「状況に埋め込まれた知能」など，研究内容を表すキーワードとして，専門的すぎるキーワードが含まれている．一方，データ整形ありの場合，関連性がないキーワードが同じトピック内に含まれておらず，一貫性が見られる．さらに，各キーワードが研究内容を表すキーワードとしてわかりやすいものである．

5.3.2 クラスタリング結果

ここでは，データ整形有無それぞれで K-means 法でクラスタリングした結果について「クラスタの要素数」「専攻の分布」「クラスタの評価」の 3 つの観点から比較を行う．ここで，評価指標をもとにクラスタ数は決定するが，データ整形有無どちらもトピック数と同じ数であった．さらに，各クラスタに含まれる教員のトピック分布を確認すると，クラスタごとに特定のトピックが優勢であることが明確に確認できた．そのため，以下では，トピック番号とクラスタ番号は対応させている．具体的には，クラスタ 1 に含まれる教員のグループは，主にトピック 1 に関連する研究を行っていると解釈する．

表 1 各トピックに関連づけられた上位 6 個のキーワード (データ整形あり)

トピック	キーワード
トピック 1	センサ，センシング，医療，ナノテクノロジー，バイオテクノロジー，バイオケミカルと分子解析
トピック 2	ナノテクノロジー，材料科学，半導体，センサ，バイオテクノロジー，環境
トピック 3	教育技術，データサイエンス，可視化技術，機械学習，HCI，ラーニングアナリティクス
トピック 4	人工知能，機械学習，深層学習，データサイエンス，センシング，量子デバイス
トピック 5	データサイエンス，材料科学，環境，機械学習，通信技術，集積回路
トピック 6	超伝導，新計算原理，次世代計算機システムアーキテクチャ，量子コンピュータ，コンピュータアーキテクチャ，量子デバイス
トピック 7	機械学習，データサイエンス，最適化，深層学習，アルゴリズム，人工知能

表 2 各トピックに関連づけられた上位 6 個のキーワード (データ整形なし)

トピック	キーワード
トピック 1	超伝導，深層学習，磁性ナノ粒子，フォトニックデバイス，機械学習，レーザー
トピック 2	機械学習，ラーニングアナリティクス，行動情報処理，CSCL (Computer Supported Collaborative Learning)，状況に埋め込まれた知能，行動・状況分析
トピック 3	教育工学，ラーニングアナリティクス，データマイニング，IoT，CMOS 回路，AI
トピック 4	SDGs（持続可能な開発目標)，磁性ナノ粒子，無線通信，コミュニケーション理解（ミーティング分析)，光通信，IoT (Internet of Things)
トピック 5	人工知能，機械学習，深層学習，電力システム，半導体，集積回路
トピック 6	機械学習，データマイニング，レーザー，無線センシング，有機材料，ナノテクノロジー
トピック 7	機械学習，強化学習，根圏ケミカル，ニューラルネットワーク，味覚センサ，嗅覚
トピック 8	最適化，マッチング理論，ミクロ経済学，マルチエージェントシステム，ゲーム理論，量子ドット
トピック 9	半導体，ミクロ経済学，スピントロニクス，オークション，水素，語学学習支援システム

5.3.2.1 クラスタの要素数

各クラスタに属する教員数が，データ整形の有無で変化があるのかを確認する．結果は表 3 のようになった．データ整形なしの場合は，クラスタの要素数が 7 から 14 にあり，比較的均一に分布しているのに対して，データ整形ありの場合は，クラスタの要素数が 6 から 22 にあり，ばらつきが増加している．さらに，標準偏差より，データ整形ありの方がクラスタリングの結果が偏りのある形になっていることがわかる．また，クラスタ数に関しては，データ整形を行うと減少している．

これらの結果から，データ整形を行うことでクラスタごとの要素数の偏りが増加し，特定のクラスタに多くの教員が集まる傾向が見られた．

5.3.2.2 専攻の分布

各教員は，「情報理工学専攻」または「電気電子工学専攻」いずれかに属しており，同じ専攻に属する教員は，研究内容に一定の関連性があると考えられる．ここで，各教員をクラスタリングした結果，各専攻の教員がどのようにクラスタに分布するかを確認する．これにより，クラスタリング結果の妥当性を比較評価できる．例えば，同じ専攻の教員がまとまってクラスタに分布していれば，研究内容に基づいた適切なクラスタリングが行われたと考えられる．データ整形を行った場合のクラスタごとの専攻の分布は図 6 のようになった．データ整形を行わなかった場合は図 7 のようになった．ここで，クラスタ数に違いがあるが，これは，評価指標によって，それぞれ最適なトピック数が違っているからである．

結果から次のようなことがわかる．まず，データ整形を行わなかった場合，クラスタごとに，「情報理工学専攻」と「電気電子工学専攻」がバラバラに分布しており，同じ専攻の教員がまとまってクラスタリングされていない．それに対して，データ整形を行った場合は，クラスタごとに専攻の偏りを確認することができ，クラスタ 3，7 のように同じ専攻の教員がまとまってクラスタリングされている．

5.3.2.3 クラスタの評価

各クラスタの適切性を評価するために，データ整形の有無ごとに，クラスタに属する教員の研究内容を確認し，そのクラスタに属する教員の研究内容が関連性があり，妥当であるかを分析する．

- データ整形あり
  情報理工学専攻の分布が偏っているクラスタ 7 を確認する．トピック番号と対応しているので，表 1 より，クラスタ 7 は「機械学習，データサイエンス，最適化，深層学習，アルゴリズム，人工知能」に関する研究を行なっている教員が属していると考えることができる．クラスタ 7 に属する教員の研究内容を確認すると，4 人程度，トピック 7 との対応が見られない教員が含まれていたが，大半はトピック 7 のキーワードに対応していた．対応が見られない教員を具体的に言及すると，「3D ビジョン，HCI，VR，機械学習」が主要なキーワードの教員が属しており，トピック 7 とは少し違う研究内容であり，トピック 3 に属するべき教員が含まれていた．
  同様にクラスタ 3 を確認すると，トピック 3 との対応が見られない教員が 3 割程度確認された．
- データ整形なし
  データ整形ありと同様に，クラスタ 2 を確認する．ク

表 3　データ整形の有無によるクラスタごとの教員数と統計値の比較

分類	1	2	3	4	5	6	7	8	9	標準偏差
整形有	10	12	17	6	12	9	22	-	-	4.95
整形無	10	14	7	8	12	11	9	10	7	2.20

注: 1，2，3... はクラスタ番号を表す．

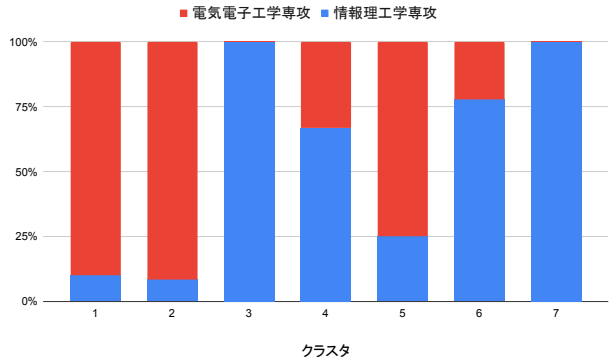

図 6　クラスタごとの専攻の分布 (データ整形あり)

ラスタ 2 は一番専攻の偏りがあるにも関わらず，「半導体」「機械学習」「VR」など，教員の研究内容に一貫性が見られなかった．これは，データ整形なしの場合，トピックに関連づけられるキーワードに一貫性があまり見られないことが原因だと考えられる．他のクラスタも同様の結果であった．

以上の結果から，データ整形を行うことで，クラスタごとの要素数の偏りが増加する傾向が見られたが，専攻ごとの分布を確認すると，クラスタリングの妥当性が向上しているため，要素数の偏りは，クラスタリングの質の低下を意味するわけではないことがわかった．さらに，データ整形がない場合は，クラスタ内の教員の研究内容にあまり一貫性が見られなかったが，データ整形がある場合は，一部誤ったクラスタに属している教員も存在するが，一定の一貫性が見られた．

これらの結果から，トピックモデリングにおいては，データ整形により得られる各トピックの一貫性を強めており，クラスタリングにおいては，専攻の分布を向上させ，クラスタごとの一貫性を向上させる効果があることが示された．したがって，データ整形はトピックモデリングとクラスタリングの精度向上に大きく貢献していると結論づけることができる．

5.4　RQ4 の回答

ここでは，提案手法でクラスタリングによって作成された各クラスタが各教員の研究分野を分類できているかを検証する．具体的には，Jaccard 係数を用いて，「クラスタ間の分離性」「クラスタ内の一貫性」について評価を行った．

5.4.1　クラスタ間の分離度

以下の手順で分析を行う．

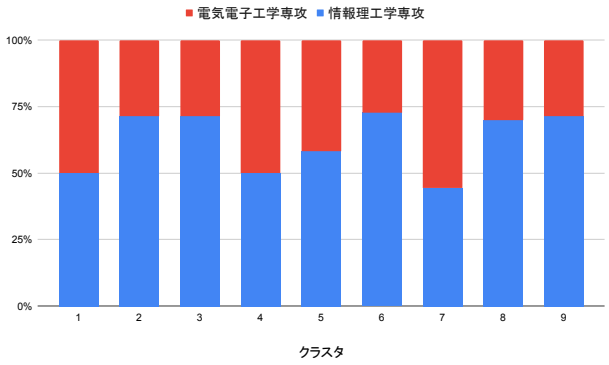

図 7　クラスタごとの専攻の分布 (データ整形なし)

(1) 各クラスタに含まれる教員の研究キーワードをまとめてリスト化．

(2) キーワードの頻度数で順位づけを行い，各クラスタの上位 10 個のキーワードを抽出．抽出したキーワードをそのクラスタのキーワードリストとする．

(3) クラスタ間の Jaccard 係数を以下の式で計算し，異なるクラスタがどれだけ異なるキーワード集合を持つかを評価する．

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

ここで，A および B はそれぞれ異なるクラスタのキーワードリストを表し，$|A \cap B|$ は両者に共通するキーワードの数，$|A \cup B|$ は両者の総キーワード数を示す．Jaccard 係数が低いほど．クラスタごとのキーワードが異なっていることを意味し．クラスタが適切に分離されていることを示す．

各クラスタのキーワードを抽出し，Jaccard 係数を計算した結果を図 8 に示す．この結果より，ほとんど 0.2 を下回った結果になっていることがわかり，適切にクラスタが分離されている部分が多いことが示された．一方，クラスタ 1 と 2 の間と，クラスタ 2 と 5 の間の Jaccard 係数が 0.3 を超えた結果となっている．ここで，図 6 より，クラスタ 1，2，5 は電気電子工学専攻に偏ったクラスタで構成されているため，電気電子工学専攻のクラスタリングの境界が曖昧であることが原因と考えられる．

5.4.2　クラスタ内の一貫性

以下の手順で分析を行う．ここで，各クラスタに対応するトピックに関連づけられた上位 10 個のキーワードを，そのクラスタの研究分野を表す代表キーワードとする．

(1) 以下の式で各教員について Jaccard 係数を計算する．

$$J(C,D) = \frac{|C \cap D|}{|C \cup D|} \times \frac{|D|}{\max(|D|)} \tag{4}$$

ここで，C はクラスタの代表キーワード，D は教員のキーワードリスト，$|D|$ は教員のキーワード数，$\max(|D|)$ は全教員の中で最大のキーワード数を示す．

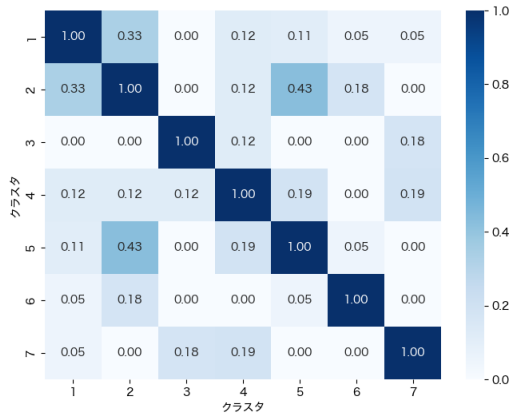

図 8 クラスタ間の Jaccard 係数

表 4 各クラスタの所属する教員の Jaccard 係数の平均値と外れ値の割合

	1	2	3	4	5	6	7
平均値	0.15	0.09	0.16	0.10	0.09	0.24	0.13
外れ値の割合（%）	10.0	33.3	47.1	16.7	16.7	22.2	27.3

キーワード数が異なる教員の比較を行うため，重み付きで計算している．

(2) 各クラスタに属する教員の Jaccard 係数の平均値を計算し，その値に対して大幅に下回る教員の割合を調べることで，クラスタ内にどれだけ誤った分類をされた教員がいるのか，を分析する．

分析の結果，各クラスタに属する教員の Jaccard 係数の平均値とその平均値の 50%以下の値をとった教員（以下，「外れ値」と省略）の割合は表 4 のようになった．また，クラスタ全体での外れ値の割合は約 27.3%だった．結果より，全体的に 3 割程度の教員を誤分類をしていることが示された．また，クラスタ 3 の外れ値の割合は 47.1%と高く，一貫性が一番低いと考えられる．

以上の結果から，本研究のクラスタリング手法の精度に関しては一定の成果が見られる一方で，いくつかのクラスタについてはまだ改善の余地があることが示された．

6. おわりに

本稿では，九州大学工学部電気情報工学科に所属する学部 3 年生に向けた卒業研究配属選択支援システムについて紹介し，定量的に評価を行った．評価の結果，embedding ベースの手法や，データ整形を行わない場合と比較して，良いクラスタリング結果を示した一方で，クラスタリング手法については，改善する部分が残っていることもわかった．さらに，本研究ではクラスタリング結果を評価指標など，定量的な評価しか行っておらず，実際のユーザのフィードバックからの評価は行っていない．

そのため，今後はクラスタリング手法や可視化手法の改良を加えたのちに，学部 3 年生に利用してもらい，フィードバックを通じて更なる評価と改善を行う予定である．

謝辞 本研究の一部は，科研費（JP19KK0257，JP 23K20734（旧 JP21H00907））の支援を受けて行われました．

参考文献

[1] S. Futamura, I. Nakao, S. Kita, R. Fujimoto, and T. Mine, “itochat2024: Development and evaluation of a department recommendation bot for open campus,” ICEA 2024, 2024.

[2] N. Eguchi, I. Nakao, K. Saito, W. Juntao, H. Motomatsu, and T. Mine, “itochat: a mobile bot recommending engineering departments and open campus events,” 2023 Fourteenth International Conference on Mobile Computing and Ubiquitous Network (ICMU), pp.1–6, IEEE, Nov. 2023.

[3] D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent dirichlet allocation,” J. Mach. Learn. Res., vol.3, no.null, p.993–1022, March 2003.

[4] 川村隆浩，江上周作，渡邊勝太郎，“研究内容の類似性に基づく科学技術マップの開発,” Japio YEAR BOOK 2018 寄稿集，pp.216–227，日本特許情報機構，Nov. 2018.

[5] 吉田光男，“科研費データの研究者所属情報に基づく研究機関マップの試作,” 人工知能学会全国大会論文集，vol.JSAI2024，pp.3Xin242–3Xin242，2024.

[6] 高沢健太，矢吹太朗，佐久田博司，“論文マッピングによる研究知識の可視化手法の提案,” 全国大会講演論文集，第 72 回，pp.393–394，mar 2010.

[7] 九州大学大学院システム情報科学府，“研究室紹介ページ,” 2025. 参照日: 2025 年 1 月 29 日. https://www.isee.kyushu-u.ac.jp/laboratory_ist.html

[8] 九州大学，“総合 web システム 卒業研究テーマ集,” 2025. 参照日: 2025 年 1 月 29 日. https://idp.kyushu-u.ac.jp/idp/profile/SAML2/Redirect/SSO?execution=e1s2

[9] 九州大学，“九州大学 研究者情報,” 2025. 参照日: 2025 年 1 月 29 日. https://hyoka.ofc.kyushu-u.ac.jp/html/home_ja.html

[10] J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang, “A density-based method for adaptive lda model selection,” Neurocomputing, vol.72, no.7, pp.1775–1781, 2009. Advances in Machine Learning and Computational Intelligence. https://www.sciencedirect.com/science/article/pii/S092523120800372X

[11] R. Arun, V. Suresh, C.E. Veni Madhavan, and M.N. Narasimha Murthy, “On finding the natural number of topics with latent dirichlet allocation: Some observations,” Advances in Knowledge Discovery and Data Mining, eds. by M.J. Zaki, J.X. Yu, B. Ravindran, and V. Pudi, pp.391–402, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[12] T.L. Griffiths and M. Steyvers, “Finding scientific topics,” Proceedings of the National Academy of Sciences, vol.101, no.suppl_1, pp.5228–5235, 2004. https://www.pnas.org/doi/abs/10.1073/pnas.0307752101

[13] R. Deveaud, E. Sanjuan, and P. Bellot, “Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval,” Document numérique - Revue des sciences et technologies de l’information. Série Document numérique, pp.61–84, June 2014. https://hal.science/hal-01002716