あいまい文献検索と文章クラスタリングによる学術研究データベース検索機能の提案

松井我颯[1]　中島陽子[1]　本間宏利[1]
[1]釧路工業高等専門学校 創造工学科
{yoko，honma}@kushiro-ct.ac.jp

概要

データベース検索時にユーザーが適切な研究分野や検索に効果的なキーワードを知らない場合，必要な情報を見つけることは困難である．本研究では形態素解析と tf-idf 値を用いて各研究分野内で特徴的な単語を収集し，word2vec を用いて単語の意味や類似度を数値的に表す．これにより，学術研究データベースに掲載されている研究記事の検索において，検索ワードが具体的な名称でなくとも利用者の要求する研究記事をより柔軟に提示可能な検索機構を提案する．

1　はじめに

近年，研究課題や研究事例のデータベースが数多く存在する．それらは先行研究や最新研究の動向を知るために利用できるが，掲載されているデータ量が多くなるほど, 検索ワードそのものを含んでいる意図しない情報も検索してしまい，利用者が望む情報を検索するためにはさらに絞り込むことが必要になってくる [5]. 研究事例や論文検索のためのデータベースを利用する際には，研究分野や研究内容に関するキーワードなどを用いることが一般的である．必要な情報にたどり着くためには検索キーワドを探す労力や時間的コストがかかってしまう [6]．利用者が検索キーワードを入力する際の単語や文を分析し, 利用者が所望していると推測される情報を提示するための情報検索支援サービスが求められている．

本研究は検索の際に最適な探索ワードを入力しなくとも，利用者が入力した文章や単語を分析と専門分野において詳細な分類を選択することでより専門的な情報を取得可能にすることを目的としている．提案する機能は，利用者の未知の智を補完することで既存のデータベースから適切なデータを獲得することが可能になる．また，文章クラスタリングの技法を用いて，研究データベース内の類似する内容の研究を自動的にグループ化することで，これまでの研究手法やそれらの研究成果を時系列にピンポイントに獲得することが容易になり，新たな課題の発見や問題解決への効果的なアプローチを効率的に獲得が可能になる．

本研究では，日本国内の研究者や研究課題のデータベースサイトである「日本の研究.com」[1]に掲載されている研究関連記事の検索において，検索キーワド補完と研究分野においての小分類のクラスタリングを用いて，利用者が所望する研究記事をより柔軟に提示可能な検索機能の提案と実装を行う．

2　「日本の研究.com」について

「日本の研究.com」は，日本国内で研究されている研究課題や研究者についての国内最大級のデータベースサイトで，科学研究費助成事業や独立行政法人科学技術振興機構などから配信されているニュースやプレスリリースを公開している．研究課題のタイトルや概要文などの文章から単語に分解し処理を施し，研究分野を推定しており，人文科学，社会科学，医歯薬学，生命科学，理学，工学の 6 つの大分野に分類される (図 2)．1 つの大分野はさらに 6 つの小分野に区分される．従って，各データの推定分野は，大分野 6× 小分野 6 の合計 36 分野で推定されるタグが付与されている.「日本の研究.com」では，検索キーワドによる検索機能，推定分野を用いた検索機能，統計データやランキングなどの機能を提供している．

本研究では，利用者が所望する情報を得る場合に利用する検索機能に注目し，データベースをより有効に利用するための機能を提案する.「日本の研究.com」において検索機能を利用する場合，次に述べる 2 点の事象に直面する．

[1]日本の研究.com：https://research-er.jp/

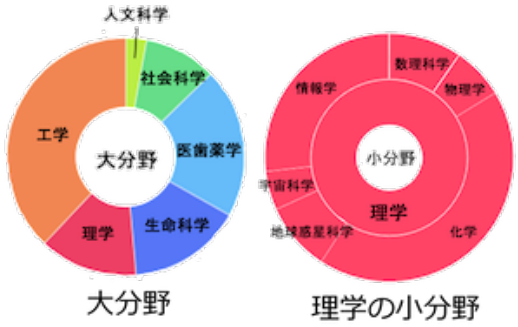

図 1: 推定分野の構成

1 点目は，利用者が検索キーワドに入力した単語とは関連のない分野内から記事を取得してしまう場合がある．現在の検索機構では，利用者が検索した単語が含まれている記事を，研究内容に関わらず提示している．そのため，研究内容が利用者が望んでいない分野の記事でも検索結果に表示されてしまう．例として，利用者が「歯」という単語で検索した場合に，小分野が『歯学』や『生物学』に属する記事を提示することが望ましいが,「歯止めがきかない」のような単語が本文中に含まれている場合や,「医歯薬学分野の研究グループに所属している」のような研究者についての紹介文が含まれている場合には，検索ワードである「歯」とは関係のない分野の記事でも検索結果として提示される．

2 点目は，推定分野を利用して記事検索する際に，特定の分野への絞り込みが難しい場合がある．現在サイトで定義されている小分野においてそれぞれに含まれる研究領域が広いため，検索ワードのみで小分野から記事の取得を試みる場合，ピンポイントで所望の記事を取得するのが難しい．

この 2 点の事象に対応するために，あいまい文献検索機能と小分野をさらに細分化した小分類の生成を提案する．

3　提案機能の概要

あいまい文献検索機能と小分類生成についての概要図を図 3 に示す．あいまい文献検索機能は検索ワードと研究記事のテキストデータを入力とし，検索ワードから類義語と連想語を求め，研究記事から生成した小分野ごとの重要語リストの単語と比較し重要度の高い類義語と連想語が含まれる研究記事を提示する．小分類生成は，小分野ごとのテキストデータを入力とし，6 クラスにクラスタリングを行い小分類の定義を行う．

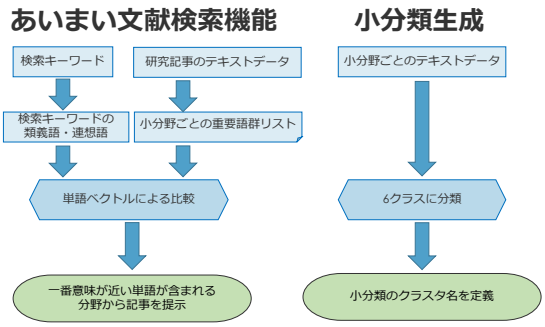

図 2: あいまい検索機能と小分野の細分化の概略

4　あいまい文献検索機能

あいまい文献検索機能は検索ワードが最適なキーワードではなくとも利用者の望む検索結果を提示する．手順を以下に示す．

1. 小分野ごとに，各分野内で特徴的な単語をまとめた重要語群リストを作成する．
2. 利用者が入力した検索ワードとその類義語・連想語を，小分野ごとの重要語群リストと，単語ベクトルを用いて比較する．
3. 検索ワードとその類義語・連想語が重要となっている分野の研究記事を検索結果として提示する．

類義語・連想語および重要語群リストについては以下で説明する．

重要語群リストの作成

重要語群リストは，小分野ごとの特徴的な単語を自動的に抽出しリスト化する．重要後は分野内の特徴的な語を表している．自動抽出にはある文書を特徴づける重要な単語の重要度を示す手法の一つである，Tf-Idf 値を用いる．

tf (Term Frequency) は, ある記事内の単語頻度であり，頻出する単語に高い数値を示す，式 1 で表される． $f(t,d)$ は記事 d 内における単語 t の出現回数を表し， $\sum_{t\in d} f(t,d)$ 記事 d 内における全単語の出現回数の和を表す．

$$tf(t_i,d_j) = \frac{f(t_i,d_j)}{\sum_{t_k\in d_j} f(t_k,d_j)} \tag{1}$$

idf (Inverse Document Frequency) は，逆文書頻度であり，複数の文章で頻出する単語に低い数値を示す，式 2 で表される．N は全記事数を表し， $df(t)$ は単語 t が出現する文書数を表す．

$$idf\left(t_i\right) = \log\left(\frac{N}{df\left(t_i\right)}\right) + 1 \tag{2}$$

Tf-Idf 値は tf と idf の積で算出する式 3 で表される．

$$tfidf = tf \times idf \tag{3}$$

記事のテキストデータの形態素解析[2]を行い名詞と固有名詞を抽出し, 各小分野内における単語ごとの Tf-Idf 値を求める [3]．小分野ごとに Tf-Idf 値上位 50 単語を採用し，それらを重要語として重要語群リストを生成する．

小分野ごとに重要語を抽出した重要語群リストの各 50 単語のうちの一部を表 4 に示す．

類義語と連想語の取得

検索ワードの類義語と連想語は，類義語シソーラスおよび連想語シソーラス ([2, 日本語語彙体系] , weblio[3]) を用い取得する．類義語と連想語は検索ワードの拡張検索ワードとして利用する．検索ワードで取得した類義語と連想語の例を表 4 に示す．

キーワードリストから適切な小分野の選択

次に，キーワードリストリストと小分野の重要語群リストを用いて，検索ワードに最も意味が近い小分野を選択する手法について説明する．単語の意味は周囲の単語によって形成されるという分布仮定に基づいて単語の意味ベクトルを分散表現として表現する Word2Vec[4]を用いて，単語を分散表現に変換する．
単語を分散表現として扱うことで，単語同士の類似度や意味の関係を数値的に表現することが可能になり，単語間の関係性をより正確に捉えることができる．

図 4 に示すように，単語の分散表現を用いて，キーワドリストの単語である検索ワードとその類似語および連想語と小分野ごとの重要語リストの各単語の類似度を求める．求めた類似度は小分野ごとに和を求め，その値が最大になる小分野を検索ワードに関連する小分野として選択する．ただし，各単語の類似度がしきい値以下の場合は関連が弱いとして除外する．キーワードリストとの類似度が最も高かった小分野に属する記事を優先的に検索結果として出力することで，検索ワードと関連の弱い分野内からの記事が出力されにくくなる．

[2]MeCab: https://taku910.github.io/mecab/
[3]https://ejje.weblio.jp/
[4]https://code.google.com/archive/p/word2vec/

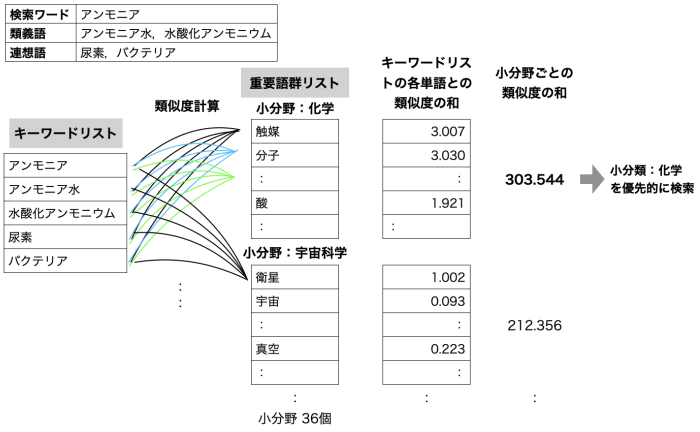

図 3: 検索ワードから関連小分野を推定するための重要語群リストの各単語とキーワードリストの各単語の類似度計算

5　小分類の生成

小分野を細分化するために，各小分野の記事に対しクラスタリング手法を用いる．以下の処理を小分野ごとに行い，各小分野に対し小分類を生成する．

1. 最適なクラスタ数をエルボー法により求める
2. 研究記事を求めたクラスタ数でクラスタリングする
3. 各クラスタ名を人手により付与し，それらを小分類の項目と定義する

クラスタリング

クラスタリング手法には，データを k 個のクラスタに分類する非階層型クラスタリング手法のひとつである k-means 法を採用する．k-means 法は，各クラスタを代表する中心点を求め，それぞれのデータ点を最も近い中心点に属するクラスタに分配する．その後，中心点を再計算し，全てのデータのクラスタが定まるまでこの操作を繰り返すことで，最終的にデータを k 個のグループに分類することが可能である．クラスタ数 k は，エルボー法を用いて決定する．

クラスタリング結果の各クラスには各クラスタに属する文の形態素解析を行い，名詞，動詞，形容詞の Tf-Idf 値を求め特徴語を抽出し，科学研究費助成事業データベース[5]で使用されている分野名を参考にクラスタ名を付与する [4]．クラスタ名は小分類名とする．小分野 36 個に対し同様の処理を行い小分類を生成する．小分野「情報学」の研究記事１００件をクラスタ

[5]科学研究費助成事業データベース：https://nrid.nii.ac.jp/

表 1: 各小分野の重要語群リストの例

小分野	重要語群リストの一部
宇宙化学	衛星，宇宙，放射線，軌道，材料，技術，月，地上，ロケット，地球，エンジン,...
化学	触媒，分子，構造，アンモニア，活性，糖，ナノ，化学，イオン，電子，材料,...
情報学	データ，技術，情報，モデル，手法，システム，画像，深層，精度，科学，知能,...
数理科学	流体，界面，乱，粘性，現象，流れ，細胞，図，方程式，領域，曲面，パターン,...
地球惑星科学	地質，軽石，地震，年代，断層，地層，地球，図，海洋，竜，オーロラ，海底,...
物理学	状態，陽子，理論，物質，量子，中間子，重力，エネルギー，電子，粒子，核，中性子,...
機械工学	滴，スプラッシュ，液，フィン，流速，風圧，乱，表面，濡れ，流体，ロボット,...
建築土木工学	コンクリート，硝酸，子ども，態，建物，セメント，免，汚泥，窒素，部屋，住宅,...
材料工学	液晶，電池，材料，太陽，半導体，構造，ミ，薄膜，皮膜，界面，温度，格子，セル,...
人間工学	細胞，脳，脊髄，声帯，患者，利き手，真皮，神経，筋，技術，図，アルギン酸，血管,...
総合工学	量子，磁性，ビット，レーザー，磁気，パルス，技術，電子，原子，光，エネルギー，磁場,...
電気電子工学	パケット，発振器，電力，コヒーレント，インターネット，アクチュエータ，信号，光，磁性,...

表 2: 検索ワードから取得した類義語，連想語の例

検索ワード	類義語	連想語
宇宙	宇宙空 万物 コスモ	地球，法則，人類 波動，生命，星 調和，次元，観測
地質	構造 地質学 ジオロジー	地層，地形，断層 花崗岩，地学，隆起 地盤，風化，堆積
地面	陸地 大地 土壌	着地，落下，両足 雑草，芝生，地上 足元，斜面，接地
失明	盲目	視神経，白内障，眼科 眼病，眼球，視力 透析，全盲，点字
チタン	–	軽量，磁気，カーボン 軽量化，アルミ，材質 スチール，ラバー

リングする実験を行った．エルボー法で求めた 4 クラスにクラスタリングを行った．小分類の各特徴語と付与した小分類名を表 3 に示す．

表 3: 各クラスタ内での特徴的単語

クラスタ名	特徴的単語
人工知能	人工，知能，深層
量子コンピュータ	光量子，コンピュータ，富岳
シミュレーション	パラメータ，モデル
ネットワーク	ネットワーク，無線

6 おわりに

本研究では，学術研究データベースの検索効率の向上を目標とした機能の提案を行った．利用者が入力した検索ワードの類義語と連想語を用い検索ワードに最も意味が近い小分野を自動的に選択するあいまい文献検索機能と小分野の細分化を行い新たに小分類のカテゴリの生成を行い，利用者が専門用語や検索技術に精通していなくともあいまい検索機能により検索ワードに関連する検索結果の提示と小分野よりさらにコアな小分類からより専門に近い記事を提示することが可能になった．小分野「情報学」のみで実験を行なったが，今後は，全ての小分野において実装し，提案機能の検証を行う予定である．また，あいまい文献検索機能において類義語･連想語が取得できない場合に対し，類義語･連想語シソーラスを更新し改善を行う予定である．

謝辞

本研究の遂行にあたり株式会社バイオインパクト，株式会社リバネスの両社から多大なご助言，ご協力を頂いたことに深く感謝します．本研究は JSPS 科研費 20K11093 基金基盤 (C) の助成を受けたものです．

参考文献

[1] 小野田崇, 坂井美帆, 山田誠二. k-means 法の様々な初期値設定によるクラスタリング結果の実験的比較. In: 第 25 回人工知能学会全国大会論文集, pp. 1J1OS91–1J1OS912011, 2011.

[2] 池原悟, 他. 日本語語彙大系 CD-ROM 版. 岩波書店, 1999.

[3] 原田大地, 荒木健治. 単語の分散表現及び tf-idf 法を用いた自動要約システム. 第 9 回 Web インテリジェンスとインタラクション研究会, pp. 49–50, 2016.

[4] 佃陽平, 森田和宏, 泓田正雄, 青江 順一. Web 検索エンジンを用いた分野連想語の自動抽出に関する研究. 言語処理学会第 12 回年次大会, pp. 648–651, 2006.

[5] 宮入暢子. 異種データの横断検索・分析ツールが切り拓く可能性. 第 16 回情報プロフェッショナルシンポジウム, 情報科学技術協会, pp. 43–48, 2019.

[6] 鹿島好央, 北山大輔. Word2Vec と Web 検索を用いた検索クエリ置換手法. 第 9 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2017), C6-1, 2017.