D-014

# ローカル検索用サジェスト生成に向けたユーザ属性とドメイン文書の語彙分析
# Vocabulary Analysis of User Attributes and Domain Documents for Generating Local Search Suggestions

鈴木 琴音† 岩本 和真† 安藤 一秋‡
Kotone Suzuki Kazuma Iwamoto Kazuaki Ando

## 1. はじめに

検索支援において，ユーザの検索意図を先読みして入力候補を提示するサジェスト機能は，検索効率の向上に有効な手段の一つである．既存のサジェスト機能は，膨大な検索ログに基づくが，ローカル環境の文書を対象とする検索システムでは，大規模な検索ログの収集が困難であるため，既存のサジェスト手法の適用が難しい[1]．また，ローカル環境による検索では，社内の部署単位の業務文書といった特定ドメインに属する文書を対象とすることがある．その場合，ユーザが求める情報やサジェスト語も，そのドメインに強く関連していると考えられる．そこで本研究では，ローカル環境における検索対象文書とユーザの属性情報を活用することで，よりユーザの検索意図を反映したサジェスト機能の実現を目指す．

我々の先行研究[2]では，ドメインが限定された文書（ドメイン文書）におけるクエリとサジェスト候補間の意味的距離を分析し，両者の類似度がサジェスト提示に有効であることを確認した．しかし，意味的距離のみでは，より適切なサジェスト語を厳選することは困難である．そこで本稿では，トピック情報に基づいてサジェスト候補を絞り込むアプローチの有効性を検証するための初期検討として，ドメイン文書内の単語出現頻度に着目し，頻出語がサジェスト候補としてどの程度適切であるかを分析する．

## 2. 分析データの構築

頻出語によるサジェストの有効性を確認するために利用するデータについて述べる．まず，検索対象とするドメイン文書群から，前処理としてストップワードを除去した後，残りの単語の出現頻度を算出して「頻出語リスト」を作成する．次に，特定のドメインにおけるユーザが選んだキーワードを「正解サジェスト候補」とする．

### 2.1 分析対象とするドメイン文書

社内文書の入手が困難であることから，本稿では，人工知能学会全国大会（JSAI）論文集から収集した過去 3 年間分（2021 年～2023 年）の概要をドメイン文書として利用する．これらの文書は，特定の専門分野における語彙が豊富に含まれており，ドメイン固有のサジェスト機能の分析に適すると考えられる．収集した総文書数は 1,152 件である．

### 2.2 正解サジェスト候補の選定

正解サジェスト候補として，JSAI 論文概要に付与されているキーワードを用いる．このキーワードは，論文で扱う主題や重要な技術を示す語として著者自身が選定したものであり，サジェスト候補として利用できるといえる．

本稿では，JSAI で定義されている一般セッション 11 トピック（①AI と社会 / ②AI 応用 /③ Web インテリジェンス / ④エージェント / ⑤ヒューマンインタフェース / ⑥ロボットと実世界 / ⑦基礎・理論 / ⑧機械学習 / ⑨画像音声メディア処理 / ⑩知識の利用 / ⑪言語メディア処理）を分析対象として選定し，各トピックからランダムに 20 件の論文概要とそれに紐づくキーワードを抽出して利用する．そして，合計 220 件の概要に付与された全キーワード（重複排除済み）を，各トピックにおける正解サジェスト候補とする．

## 3. ストップワードの作成

ストップワードの選定に関しては，様々なアプローチ[3-5]が存在する．東ら[3]は，文書頻度（Document Frequency, DF）に着目し，多くの文書に出現する語をストップワードとして抽出し，さらに意味的に類似する単語を合わせて排除する手法を提案している．本研究においても，東らの研究[3]を参考に，DF を用いてストップワードのベースリストを作成する．

まず，220 件の論文概要（ドメイン文書）に対して，Mecab（+NEologd）で単語分割し，名詞を抽出する．なお，連続する名詞は複合名詞として扱う．重複を排除した結果，得られた名詞の総数は 9,666 語となった．これらの名詞に対して DF を算出し，DF が単語集合全体のおよそ 10%を超える語をストップワード候補とする．本稿では，DF の閾値を 95 に設定する．次に，DF の閾値に基づいて選定されたストップワードに対し，Word2Vec を用いて類義語を追加する．より厳密に意味的類似性の高い語に限定するため，本稿では，類似度の閾値を 0.7 に設定する．そして，閾値を超える類似語をストップワードに追加する．

上記の手順により生成されたストップワードを統合し，重複を排除した結果，2,483 語のストップワードリストが得られた．このリストには，「研究」や「論文」「情報」「システム」など一般的な学術用語や，「こと」や「もの」「ため」など形式名詞などが含まれている．

## 4. 頻出語によるサジェスト分析

作成したストップワードリストを用いて，ドメイン文書からストップワードを排除し，残りの単語群の出現頻度を集計する．そして，この頻出語リストが，どの程度正解サジェスト候補に含まれているか分析する．

### 4.1 トピック別頻出語

トピック別の論文概要集合を対象に，ストップワード除去後の名詞の出現頻度を算出し，頻出語の特徴を分析する．

各トピックにおける頻出語リストを表 1 に示す．表 1 より，多くのトピックにおいて，そのトピックに直接関連する専門用語が頻出語の上位に出現する傾向が見られた．例

† 香川大学大学院創発科学研究科 Graduate School of Science for Creative Emergence, Kagawa University
‡ 香川大学創造工学部 Faculty of Engineering and Design, Kagawa University



Copyright © 2025 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

表 1 各トピックにおける頻出語リスト

| Topic | Top1 | Top2 | Top3 | Top4 | Top5 |
|---|---|---|---|---|---|
| ① | 企業 | 信頼 | 支援 | 提示 | 人々 |
| ② | 実施 | 特徴量 | 発生 | 学習データ | 企業 |
| ③ | 興味 | 特性 | 発見 | プール | 類似度 |
| ④ | 戦略 | シミュレーション | 協力 | 発生 | メカニズム |
| ⑤ | 共感 | 収集 | 促進 | 理解 | 感情 |
| ⑥ | ロボット | 物体 | 動作 | 獲得 | 教示 |
| ⑦ | 意味 | 発見 | 臨床の知 | 制御 | 嗜好 |
| ⑧ | 強化学習 | 探索 | 獲得 | 複雑 | 報酬 |
| ⑨ | 物体 | 動画 | 撮影 | 検出 | 理解 |
| ⑩ | 知識グラフ | 知識 | 協働ロボット | 機能 | 属性 |
| ⑪ | 単語 | 日本語 | 対話 | BERT | 推論 |

えば，⑧の機械学習トピックでは，「強化学習」「探索」，⑪の言語メディア処理トピックでは，「対話」「BERT」が上位に現れた．一方で，「興味」「利用」「提示」といった一般的な語も高い頻度で出現していた．これらの語は，ストップワード処理で除去できなかった汎用性の高い語であるか，ドメインに関する論文で多用される語である可能性が考えられる．

### 4.2 正解サジェスト候補のカバー率と出現順位

2.2 で定義した正解サジェスト候補が頻出語リストに含まれているか（カバー率），また，どの順位に出現するか（出現順位）を調査・分析する．

調査の結果，正解サジェスト候補の約 60%（394 件）は，頻出語リストに存在していなかったことから，頻出語リストのカバー率は低いといえる．しかし，正解サジェスト候補は，必ずしも概要に含まれているとは限らないことから，本文を用いて分析する必要がある．

次に，頻出語リストに含まれる正解サジェスト候補の出現順位を分析する．図 1 に各ランキング帯における正解サジェスト候補の割合の分布を示す．図 1 に示すグラフは，1〜500 位を 50 位ごとに刻んでいる．図 1 より，正解サジェスト候補は特定の順位に集中せず，1 位から 250 位の間に広く分散していることがわかる．低頻度帯では，トピックごとで顕著な偏りが見られた．「AI 応用」トピックでは 72.4%が，「言語メディア処理」トピックでは 46.4%が 501 位以上の階級に分類されていることから，これらのトピックのキーワードは，頻出語リストの上位には現れにくい傾向を確認した．この偏りの要因としては，ストップワード処理で除去される一般的な単語が正解サジェスト候補に多く含まれていなかったため，当該トピックにおける正解サジェスト候補の母数自体が他トピックよりも多くなったことが挙げられる．

以上より，低カバー率でかつリストに含まれる場合も出現ランクは低位にまで広く分散するという結果は，サジェスト候補と論文概要における単語の出現頻度との間に大きな乖離があることを示している．よって，専門性の高い単語では，必ずしも概要文書全体での単純な出現頻度と相関しないことを示唆している．

### 4.3 ランク外の単語に対する定性的分析と考察

本節では，ランク外にある単語の性質を定性的に分析し，今後のサジェスト性能の向上に向けた課題を考察する．

ランク外にある正解サジェスト候補の例としては，言語メディア処理トピックの「自然言語処理」「テキストマイニング」「トランスフォーマー」，機械学習トピックの

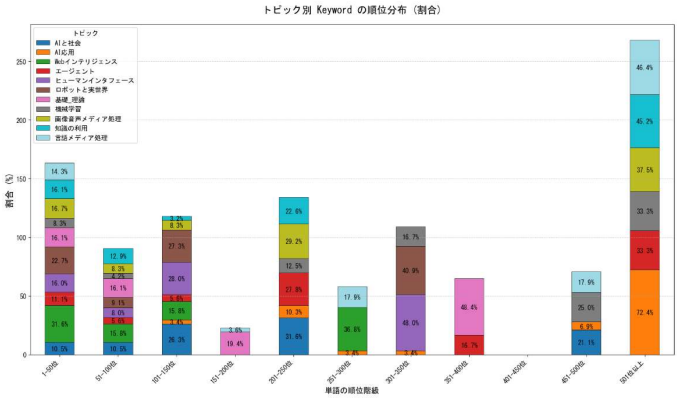


図 1 トピック別正解サジェストの出現順位割合の分布

「敵対的生成ネットワーク」「特異値分解」「深層学習」などが挙げられる．これらは，各トピックにおいて重要な専門用語といえる．つまり，概要文書集合内で高頻度に出現するドメイン関連語とは別に，出現頻度は低いが，論文の内容を示す重要語が存在することを示している．よって，サジェスト語の選択においては，出現頻度以外の指標も考慮する必要がある．

また，本分析は，論文概要のみを対象としたため，本文や参考文献を用いれば出現した可能性がある重要語を十分に抽出できていない点も考慮する必要がある．また，単純な頻出語に基づいたサジェスト候補絞り込みには，限界があるといえ，我々の先行研究である類似語による意味的距離手法を組み合わせた新しいサジェスト提示手法が必要となる．今後は，これら手法を組み合わせて，重要な専門用語を捉え，サジェスト性能の向上を目指す．

## 5. おわりに

本稿では，ローカル環境におけるサジェスト機能の実現に向けた基礎的分析として，ドメイン文書内の単語頻度とユーザが付与したキーワードとの関連性を分析した．作成したストップワードリストを用いてストップワードを除去した単語群から頻出語を抽出し，ドメイン文書の正解サジェストを比較した結果，多くの正解サジェスト候補は頻出語リストの上位には含まれず，カバー率も低いことを確認した．この結果から，出現頻度に基づくサジェスト手法には限界があり，意味的類似度やトピック情報など複数の特徴量を組み合わせることが必要であるといえる．

今後は，分析で明らかになった課題を踏まえ，文書集合全体を対象とした分析の他，類似語やトピック情報など複数の手法を組み合わせた高度なサジェスト手法を検討する．

### 参考文献

[1] 佐野, “階層的データ管理と複数データ領域の高速横断検索を実現する社内ナレッジシステム”, FIT2022 講演論文集, 第 4 分冊, pp.349-350, (2022).
[2] 鈴木他, “特定ドメイン向けローカル検索用のサジェスト提示に向けた分析”, NLP2025 論文集, pp.3248-3251, (2025).
[3] 東他, “単語の出現頻度と類似性に基づいたトピックモデル洗練化手法”, コンピュータソフトウェア, Vol.36, No.4, pp.25-31, (2019).
[4] 國府他, “内容推測に適したキーワード抽出のための日本語ストップワード”, 日本感性工学会論文誌, Vol.12, No.4, pp.511-518, (2013) .
[5] 桑原他, “BERT を用いた文書分類タスクにおけるストップワードの有効性の検証”, IPSJ 2022-DBS-175, 41, pp.1-6, (2022).



Copyright © 2025 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.