大規模検索ログを用いない特定ドメイン向けサジェスト手法 ― 単語分散表現とトピック別出現頻度の統合 ―

鈴木 琴音[1] 岩本 和真[1] 安藤 一秋[2]
[1]香川大学大学院 創発科学研究科 [2]香川大学 創造工学部
{s25g357, s24g351, ando.kazuaki}@kagawa-u.ac.jp

概要

本研究では，大規模な検索ログの取得が困難な小規模かつ特定ドメインの文書検索において，単語分散表現に基づく意味的類似度と，ドメイン内におけるトピック別語彙頻度という 2 つの要素を結合するサジェスト手法を提案する．実験では，ユーザが検索時に考案した追加キーワードを用いて提案手法による再現性を検証した．実験の結果，単語の類似度と頻度のスコアを適切に調整することでユーザが考案したサジェスト語の再現性が向上することを確認した．特定ドメインにおける検索支援には，意味的類似性とトピック特有の語彙頻度の 2 要素を用いることが有効であることを確認した．

1 はじめに

検索支援におけるサジェスト機能は，ユーザの検索意図を先読みして入力候補を提示することで，入力コストの削減やクエリの具体化を助け，検索効率を向上させる有効な手段である．一般に，Google に代表される大規模な Web 検索エンジンにおけるサジェスト機能は，膨大な検索ログに基づく確率モデルを活用して実現されている．しかし，企業や大学といった組織における部署単位のローカル環境を対象とした検索システムや，その導入段階においては，大規模な検索ログの収集が困難である．そのため，既存手法をそのまま適用することは難しく，このような環境に適したサジェスト機能が求められている．

そこで，本研究では，検索対象文書のドメイン特性を活用したサジェスト機能の実現を目指す．ローカル環境における検索では，部署単位の業務文書や特定の学術分野に属する文書など，専門的かつ限定的なドメインを対象とする場合が多い．このような場合，ユーザが求める情報や適切なサジェスト語は，ドメイン特有の用語体系や出現頻度などに依存する．

我々の先行研究[1]では，ドメイン文書集合における第一クエリ（ユーザが入力したクエリ）とサジェスト候補語間との意味的距離に着目し，単語分散表現に基づく類似度がサジェスト提示における有効な指標になり得ることを確認した．また，ドメイン文書内における単語の出現頻度に着目した研究[2]において，頻出語がサジェスト候補として一定の妥当性を持つことを確認した．

しかし，ユーザが的確な情報を得るためには，検索タスクへの習熟度やドメイン知識の深さに応じて，柔軟に制御されたサジェストの提示が必要となる．クエリの表現方法は多様であり，単語間類似度や出現頻度といった単一指標に基づく固定的なフィルタリングでは，広範な概念を指す一般語と，特定の技術や概念を指す専門用語を，状況に応じて適切に提示することが困難である．

そこで本稿では，単語間の意味的類似度と文書内における出現頻度という 2 つスコアを統合したサジェスト手法を提案する．また，提案手法がサジェスト候補提示手法として妥当であるかを評価する．

2 関連研究

ドメイン特化型単語分散表現の有効性　検索支援において，クエリとサジェスト候補語との意味的関係性を計算する手法として，単語分散表現が広く用いられている．Diaz ら[3]は，Wikipedia といった大規模汎用コーパスで学習した単語分散表現より，検索対象となる特定ドメインの文書集合で学習した分散表現のほうが，ドメイン固有の専門用語や意味関係をより正確に捉えることができ，検索精度も向上すると述べている．本研究においても，Wikipedia で事前学習した Word2vec に対して，検索対象ドメイン文書で追加学習したモデルを用いる．

語の出現頻度と意味的類似度の統合　サジェスト候補のランキングにおいて，単語の意味的類似度だけでなく，対象ドメインにおける単語重要度（出現頻度）を統合するアプローチ[4,5,6]が提案されてい

る．ALMarwi ら[4]は，Word2Vec を用いた意味的類似度と，WordNet，単語の出現頻度を統合することで，クエリ拡張語の選択と重み付けする手法を提案している．しかし，先行研究の多くは，意味的類似度と出現頻度の重みを固定して利用しており，多様なトピックが混在する文書集合を対象とした環境において，トピックごとに最適な重みや閾値を動的に調整する手法は十分に検討されていない．

3 サジェスト語のスコアリング手法

提案手法は，第一クエリとサジェスト候補語の意味的類似度に基づくスコアと，文書集合内における各トピックの語彙出現頻度に基づくスコアの 2 つを統合したサジェスト手法である．以降，各手法について説明する．

3.1 意味的類似度スコア

先行研究[1]に基づき，本稿では，検索対象ドメインに特化した用語間の意味関係を捉えるため，ドメイン特化型の単語分散モデルを構築する．具体的には，事前学習された単語分散モデルに対して，検索対象となるドメイン文書集合を追加学習することにより，一般的な語彙体系に加え，専門ドメイン特有の文脈に基づく単語間の意味類似度を算出可能とする．本稿では，東北大学自然言語処理グループが公開する日本語事前学習モデル[7]を追加学習させた分散表現モデルを用いる．

第一クエリqとサジェスト候補語wの分散表現ベクトル間のコサイン類似度$Sim(q,w)$を算出する．ただし，極めて近い意味を持つ派生語の提示を避けるため，類似度上限閾値 τ を設定し，$cos(q,w) > \tau$ を満たす語は候補から排除する．

$$Sim(q,w) = \begin{cases} \dfrac{cos(q,w)}{\tau} & if\ \cos(q,w) < \tau \\ 0 & if\ \cos(q,w) \geq \tau \end{cases} \quad (0.0 \leq \tau \leq 1.0) \tag{1}$$

3.2 トピック別出現頻度スコア

検索時に選択されたトピック内の文書集合 T におけるサジェスト候補語wのトピック別出現頻度スコア $Freq(w,w',T)$ を用いる．まず，ドメイン特有の語が上位に出現しやすくなるために，東ら[8]の手法を用いて，ドメインに依存しない普遍的な語をストップワードとして排除する．トピック別出現頻度スコア $Freq(w,w',T)$ は，$Count(w,T)$を，トピック内の単語集合 T 内で最も頻出する語w'の出現回数 $Max\ Count(w',T)$で正規化することにより，0 から 1 の範囲の値として算出する．

$$Freq(w,w',T) = \frac{Count(w,T)}{Max\ Count(w',T)} \tag{2}$$

3.3 スコア統合

意味的類似度スコアとトピック別出現頻出スコアを統合し，最終的なサジェスト語のスコアとする．

$$Score(w) = \alpha \times Sim(q,w) + (1-\alpha) \times Freq(w,w',T) \quad (0 \leq \alpha \leq 1.0) \tag{3}$$

ここで，α は，意味的類似度とトピック別出現頻度スコアの寄与度を調整するパラメータである．

最終的に，$Score(w)$ の上位 N 件をサジェスト候補として提示する．

4 評価データの構築

提案手法を評価するための評価データ収集法とサジェストが検索性能に与える影響について分析する．

4.1 検索対象文書

本研究では，社内文書や特定ドメイン文書のローカル文書検索におけるサジェスト機能を想定している．本実験では，ドメイン要素を含む文書集合として，情報科学フォーラム（FIT）において，2020～2024 年までの一般発表論文，計 2,122 件を対象とする．

4.2 評価データ収集

提案手法の有効性を検証するため，Elasticsearch[i]による検索システムを用いて，57 名の協力者から以下の流れで検索ログを収集した．

1. 単体クエリ検索（1st-query）：ユーザが最初に想起したキーワード（第一クエリ）のみで検索
2. サジェスト併用検索（1st-query + user-suggest）：第一クエリに加え，ユーザ自身が追加で入力したキーワード（第二クエリ）を併用して検索

各検索結果の上位 10 件の論文に対して，ユーザに「求める情報であったか」を 4 段階（1 : 低～4 : 高）の適合度として評価してもらう．本実験では，検索システムによる自動サジェストは用いず，ユーザに

[i] https://www.elastic.co/jp/elasticsearch

「本来欲しかったサジェスト語」を直接入力してもらうことで,サジェスト語の正解セットを構築する．収集の結果，重複ペアを除いた第一クエリと第 2 クエリのペア数は 49 件となった．

4.3 検索性能の比較分析

ユーザがサジェスト語を併用した場合の検索性能への影響を定量的に分析する．具体的には，単体クエリ検索と，サジェスト併用検索を比較し，検索性能が向上したかを検証する．評価指標として，Precision@k (k=3, 5, 10)，MRR を用いる．

単体クエリ検索（1st-query，53 件）と，サジェスト併用検索（1st-query + user-suggest，49 件）の検索性能の比較結果を表 1 に示す．

分析の結果，Precision@3 においては，サジェスト語を併用した場合のほうが単体時より向上した．また，MRR は，単体時の 0.553 に対し，併用時は 0.603 と向上した．この結果は，ユーザが第一クエリに対して情報を絞り込む第二クエリを追加することで，検索結果の最上位層における適合率が向上し，ランキング上位で適合文書を発見できたことを示している．一方，Precision@10 においては，単体クエリ検索と比較して微減した．この要因として，ユーザが求める文書が検索対象に十分に含まれていなかった可能性が考えられる．

表 1 クエリ別検索性能

	1st-query	1st-query + user suggest
Precision@3	0.370	**0.384**
Precision@5	**0.322**	0.321
Precision@10	**0.276**	0.255
MRR	0.553	**0.603**

5 サジェスト評価実験

提案手法が，ユーザ自らが思考して入力した第二クエリをどの程度再現できるかを検証する．

5.1 実験設定

サジェスト併用検索時において，ユーザが高い適合度を付与した論文が検索結果に含まれていた検索クエリに対応する第二クエリを正解サジェスト語と定義する．提案手法のパラメータ（τ, α）を変動させた際に，正解サジェスト語がランキング上位に出現するかを，Recall@10 および MRR で評価する．本実験では，$0.3 \leq \tau \leq 0.8$ (0.05 刻み)，$0.0 \leq \alpha \leq 1.0$ (0.1 刻み）の範囲で設定する．なお，データセットの規模に起因する未知語の影響を排除するため，単語分散表現モデルが語彙として認識可能であった有効クエリ候補語ペア（51 組）のみで再集計した．

5.2 評価結果

図 1 に，調整パラメータ α と類似度上限閾値 τ を変動させた際の Recall@10 の推移を，また，図 2 には，同様の条件下における MRR の推移をヒートマップとして示す．図 1 より，$0.4 \leq \alpha \leq 0.6$,$0.55 \leq \tau \leq 0.7$周辺の設定において，Recall@10 が最大値 0.157 を記録したことがわかる．一方，意味的類似度のみを用いた設定 ($\alpha = 1.0, \tau = 0.6$) の Recall@10 の最大値は 0.059 にとどまり，提案手法はこれに対して約 2.66 倍の性能を示した．また，トピック別出現頻度のみを用いた設定 ($\alpha = 0$) における最大値 0.098 と比較して約 1.6 倍の性能を示した．

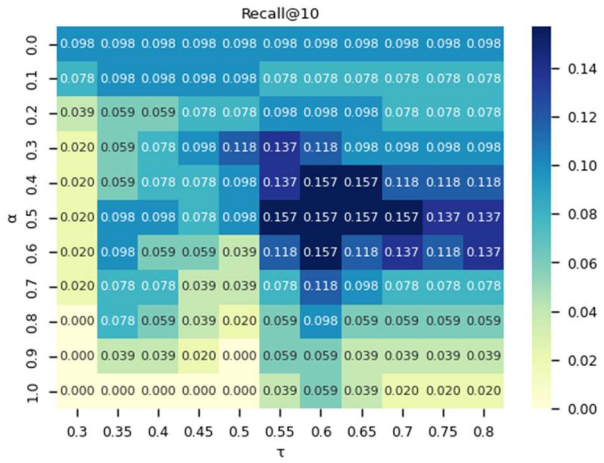

図 1 Recall@10 のヒートマップ

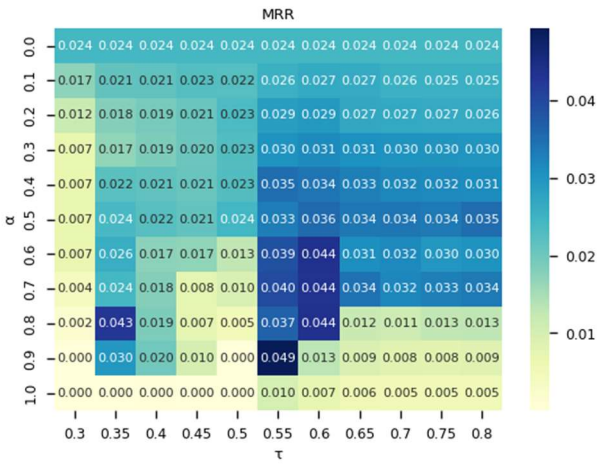

図 2 MRR のヒートマップ

MRR は図 2 の結果から，$\alpha = 0.9, \tau = 0.55$において最大値 0.049 を示した. また, $0.6 \leq \alpha \leq 0.8, 0.55 \leq \tau \leq 0.6$周辺において，MRR が 0.040 以上となった. 意味的類似度のみを用いた設定（$\alpha = 1.0, \tau = 0.55$）では，MRR が 0.014 と約 4.4 倍，出現頻度のみを用いた設定（$\alpha = 0$）における最大値 0.024 と比較しても約 1.83 倍向上した. MRR が最大である$\alpha = 0.9, \tau = 0.55$ではサジェスト順位において,「対話-コーパス」「雑談-コーパス」が 1 位，「傾聴-共感」が 2 位であり，特定の正解サジェストが高順位に提示されたため，MRR が大幅に向上したと考えられる.

次に，Recall@10 および MRR の両方で高い値を示した$\alpha = 0.6, \tau = 0.6$設定における上位 10 件のクエリ･サジェストペアを表 2 に示す. 提案手法により，単一スコアのみでは上位に現れにくい正解サジェスト語が上位に位置する傾向が確認された．例えば，「共感 - 感情」ペアは，トピック別出現頻度のみを用いた場合（$\alpha = 0.0$）では 5 位，意味的類似度のみを用いた場合（$\alpha = 1.0$）では 233 位であるのに対し，両スコアを統合した $\alpha = 0.6$では 1 位であった．他ペアでも, 単一スコアのみでは順位が低い場合でも，2 つのスコアを統合することで，ユーザの意図に沿った候補を適切に上位に提示できる傾向がみられた.

表 2 $\tau = 0.6$の α変動時のサジェスト順位

クエリ	サジェスト	α=0.6	α=0.0	α=1.0
共感	感情	**1**	5	233
対話	コーパス	**4**	19	8
自己開示	対話	**5**	45	101
小説	感情	**5**	**5**	1,129
傾聴	共感	**7**	197	9
雑談	コーパス	**8**	19	6,699
応答	感情	8	**5**	16,210
触覚	視覚	**9**	381	**9**
AI	会話	21	**6**	26,472
小説	対話	206	**45**	6,767

5.3 考察

個別ログの分析から，閾値変動による再現性向上の要因分析と，再現性が低かった正解サジェストに対してエラー分析する.

閾値変動によるランキング結果の変動　表 2 の$\alpha = 0.6, \tau = 0.6$の設定では，単一スコアのみを用いた場合には上位に現れない正解サジェスト語が上位に位置していた．例えば，「共感 - 感情」や「自己開示 - 対話」といったペアは，単語分散表現上の意味的類似度のみを用いた設定（$\alpha = 1.0$）では 100 位以下に位置しており，出現頻度のみを用いた場合でも，よりトピックに特化した専門語に埋もれる傾向が見られた.

しかし，提案手法では，トピック内で適度に頻出するという統計的性質と，類似度上限閾値 τ によってクエリと過度に類似度した語を除外した意味的類似性を統合することで，これらの語が上位に出現した．これは，特定ドメインにおけるユーザの検索意図が，単なる類義語や頻出語ではなく，当該分野の文脈において重要度の高い語に集中していることを示唆している.

エラー分析　ランキング下位に位置したサジェスト語の分析から，特定トピックにおける語彙不足が一因であることがわかった．例えば，「nfs – 検索」というペアにおいて，各語の出現頻度自体は一定数確認されたが，意味的類似度を重視する設定では順位が低下した．これは，両語が同一文脈内で共起する回数が少なく，単語分散表現上での類似度が低く算出されたことが考えられる．このような傾向は，1 学会のみを対象とした限定的なコーパスの利用に起因しており，ユーザが想定する広範な専門用語や意味関係を十分に網羅できていない可能性を示している．よって，今後の課題としては，語彙の網羅性の拡張がサジェストの性能に与える影響を評価する.

6 おわりに

本稿では，大規模な検索ログに依存せず，特定ドメインの文書集合からサジェスト語を生成する手法の実現に向け, 単語分散表現による意味的類似度と，ドメイン別の出現頻度を統合したサジェスト手法を提案し，ユーザの検索ログを用いて評価した．実験の結果，$0.4 \leq \alpha \leq 0.6, 0.55 \leq \tau \leq 0.7$ の範囲で，Recall@10 が最大となり，$0.6 \leq \alpha \leq 0.8, 0.55 \leq \tau \leq 0.6$の範囲で MRR が向上し，提案手法のサジェスト提示における有効性が示された.

今後は，単語分散表現における単語間の類似性向上のための外部コーパス活用や，トピックに応じた動的パラメータ調整法を検討し，文書量の少ない分野においても文脈的な関連性を捉えられる性能の高い検索支援を目指す.

参考文献

[1] 鈴木琴音, 岩本和真, 安藤一秋, “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).

[2] 鈴木琴音, 岩本和真, 安藤一秋, “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”, 第 24 回情報科学技術フォーラム.第二分冊.pp.173-174. (2025).

[3] F. Diaz, B. Mitra, N. Craswell, “Query Expansion with Locally-Trained Word Embeddings”, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.367-377, 2016.

[4] H. AMarwi, M. Ghurab, I. Al-Baltah, “A hybrid semantic query expansion approach for Arabic information retrieval”, *J Big Data*, **7**, 39 (2020).

[5] J. Gabín, M.E. Ares, J. Parapar, “Keyword Embeddings for Query Suggestion”, ECIR 2023, Lecture Notes in Computer Science, vol.13980. Springer, Cham. (2023).

[6] K. Manojkumar, A. Grigor, “Personalized Prefix Embedding for POI Auto-Completion in the Search Engine of Baidu Maps” , KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp.2677-2685. (2020)

[7] 東北大学 乾・鈴木研究室. 日本語 Wikipedia エンティティベクトル, 2016. http://www.cl.tohoku.ac.jp/~m-suzuki/jawiki_vector/ (参照 2026-01-08).

[8] 東和幸, 高橋仁, 中川博之, 土屋達弘, “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).