E コマースにおけるユーザー行動ログと大規模言語モデルを活用したクエリ拡張のための辞書作成

浅野孝平　稲田和明　張信鵬
株式会社 MonotaRO
{kohei.asano, kazuaki.inada, xinpeng_zhang}@monotaro.com

概要

情報検索において，ユーザーの意図をより正しく捉え，適切な検索結果に改善する方法の 1 つとしてクエリ拡張が知られている．クエリ拡張の実現には人手で整備された辞書をはじめとする言語資源が必要だが，特定の E コマースなどの専門性の高いドメインで利用可能な言語資源は一般に公開されておらず，また人手での作成コストが高い．本研究では，E コマースの商品検索サービスのログデータにおけるユーザー行動の類似度に基づいて，言語資源のための候補データを抽出し，大規模言語モデルを用いて言語的な類似度の評価を行うことで，クエリ拡張のための高精度な言語資源を作成する．

1　はじめに

近年，E コマースが盛んに利用されることによって膨大なユーザー行動が蓄積しており，そのユーザー行動ログを用いた商品検索や商品推薦などの E コマースサービスが開発されている [1]．E コマースの商品検索では，ユーザーが入力した検索クエリの意図にマッチする商品を結果として提示することが重要であるが，特に間接資材などの専門性の高い商品を扱う E コマースでは，専門用語や商品の業界固有の呼称，誤字や表記ゆれによって適切な商品マッチングが難しい．しかしユーザーの意図にそぐわない商品を提示すると，E コマースの信頼の低下に繋がり， Customer Lifetime Value (CLTV) の低下を招く恐れがある．そのため，専門性の高い商品を扱う E コマースにおいて商品検索サービスを提供し高い CLTV を達成するためには，クエリに含まれるユーザーの意図を考慮した高精度の商品検索サービスが求められる．

本研究では，表面的なテキストの類似度を超えた，クエリに含まれるユーザーの意図を考慮する方法の 1 つである**クエリ拡張 (Query Expansion)** [2] に焦点を当てる．E コマースに蓄積されたユーザー行動のログデータと **大規模言語モデル (LLM: Large Language Model)** [3, 4, 5] を組み合わることで，クエリ拡張のための高品質な言語資源の獲得し，商品検索へ応用することで有用性を示す．まず，著者らの先行研究 [6] を用いて，ユーザー行動ログから検索クエリと商品の 2 部グラフを作成し，ユーザー行動が類似するクエリ間に含まれる単語のペアをクエリ拡張の辞書の候補として獲得する．そして，LLM を用いて候補の単語ペアの言語的な関連度の高い単語ペアに絞り込むことで，ユーザー行動と言語的な類似度の両方を考慮したクエリ拡張のための辞書を作成する．評価実験では，ユーザー行動ログと LLM を活用して獲得した辞書がクエリ拡張のデータとして適切であるかを目視で確認する．さらに，実際の E コマースの検索ログデータを用いた商品検索のシミュレーション実験によって，言語的な類似度を考慮した辞書を用いたクエリ拡張によって検索指標が改善することを示す．また，クエリ拡張とベクトル検索との比較を実例を用いて比較する．

2　関連研究

E コマースの商品検索としてよく用いられる，BM25 [7] などのテキストマッチスコアを用いる全文検索システムでは，多義語，同義語によって再現性が低下することが知られている [2]．この問題を緩和する方法として，クエリ拡張があり検索クエリを再定式化することで検索パフォーマンスを向上させる．クエリ拡張の手法の 1 つとして，シソーラス辞書を用いる方法が挙げられる [8] が，専門性が高いドメインにおいて，高品質な辞書の作成はコストが高い作業である．一方，ユーザー行動ログから同義語の候補を抽出し，辞書を獲得する研究も行われている [6, 9]．本研究は著者らの先行研究 [6] のユー

ザー行動の類似度と LLM を組み合わせることでより精度の高いクエリ拡張を目指す．また近年では，LLM を用いてクエリ拡張を行う手法も提案されている [10, 11]．LLM を用いたクエリ拡張手法 [10, 11] は，拡張する単語を LLM のみを用いて生成する．

また専門用語や誤字や表記ゆれの問題の解決策として，Semantic Vector Search [12, 13] や Retrieval Augmented Generation (RAG) [14] は，LLM などから得られるベクトル表現を用いた情報検索によって成果を上げている．本研究では，クエリ拡張をベースとした提案手法とベクトル検索の検索結果を比較し，それぞれの手法の特性を評価する．

LLM を活用した言語資源の獲得として，人間の持つ常識を知識グラフ化する研究が存在する [15, 16]．本研究においてもプロンプトを用いて LLM を利用するが，より効率的に LLM を活用するためのプロンプトチューニングに関する様々な研究が行われている [17]．プロンプトチューニングには，LLM への指示に加えていくつかの正解事例を入力として与える few-shot と，指示のみを与える zero-shot があり，特に zero-shot は擬似相関を回避しロバストな推論が可能である [3]．Open AI の GPT シリーズや Google の PaLM シリーズなどの LLM は公開されている Web ページから学習されているため，本研究で着目する専門用語と一般用語の関係性をすでに獲得していると期待できる．そこで，本研究では zero-shot によって単語の関連性を評価する．

3 ユーザー行動ログと LLM を用いた辞書作成

ユーザー行動ログと LLM を用いた辞書作成では，はじめにログデータからユーザー行動が類似する検索クエリのグルーピングを行い，ユーザー行動が類似する単語のペアを候補として獲得する．その後，候補の単語ペアに対して LLM を用いて言語の意味的な関連度を算出することで，クエリ拡張のための辞書を作成する．

3.1 ユーザー行動ログから関連語候補の抽出

本研究では，E コマースのログデータとして蓄積された膨大な種類の検索クエリに含まれる単語から辞書を作成する．しかし，すべての単語の組み合わせの類似度を LLM を用いて総当りで評価することは非現実的であるため，検索意図が類似している検索クエリに含まれる単語のペアをユーザー行動ログから抽出する．先行研究 [6, 18] で用いられる検索クエリと商品の 2 部グラフを応用して，ユーザー行動が類似している検索クエリをクラスタリングする．

検索クエリ q と関連する商品の集合を $P(q)$ とし，検索クエリ q の検索においてクリックされた商品集合と定義する．2 つの検索クエリ q_i, q_j 間のユーザー行動としての類似度 $S_Q(q_i, q_j)$ を $S_Q(q_i, q_j) = \mathrm{Jaccard}(P(q_i), P(q_j))$ と定義する．ここで Jaccard 係数は集合 A, B に対して $\mathrm{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$ で定義される．

各クエリ q_i に対して，q_i とユーザー行動としての類似度の高いクエリクラスタを $C_{1\mathrm{hop},i}$ として以下で定義する．

$$C_{1\mathrm{hop},i} = \{q_j : S_Q(q_i, q_j) > \tau\} \tag{1}$$

ここで，τ は類似度を調整するためのパラメータである．

q_i のクリックデータが不十分である場合，1hop クラスタ $C_{1\mathrm{hop},i}$ では q_i と比較できる検索クエリ数が少なくなることが考えられる．そこで，1hop クラスタ $C_{1\mathrm{hop},i}$ を 2 つまで連結することを許容し，以下の式で定義される検索クエリクラスタ C_i に拡張して，以降の処理に用いる．

$$C_i = C_{1\mathrm{hop},i} \cup \bigcup_{q_j \in C_{1\mathrm{hop},i}} C_{1\mathrm{hop},j} \tag{2}$$

次に単語間のユーザー行動の類似度を計算する．C_i に含まれる検索クエリを単語分割することで得られる単語集合を T_i とする．ただし，T_i の単語が類似しているとは限らない．例えば，クエリクラスタが「石鹸 工業用」と「ピンク石鹸」である場合，単語である「工業用」と「ピンク石鹸」は類似していない．そのため，T_i を辞書作成のための単語ペアとしては不十分であると考えられる．そこで，類似した単語は類似した単語集合に分布すると考え，単語 t_i, t_j のユーザー行動としての類似度 $S_T(t_i, t_j)$ を $\mathrm{Jaccard}(D(t_i), D(t_j))$ で評価する．ここで，$D(t)$ は単語 t を含む単語集合のインデックス集合で，$D(t) = \{i : t \in T_i\}$ である．

3.2 LLM を用いた関連度の評価

ユーザー行動のログデータを用いて，ユーザー行動としての関連が強い単語ペア集合を獲得したが，言語の意味的な評価を行っていないため，クエリ拡張の辞書に用いるデータとして不十分であると言える．そこで，LLM を用いて単語ペアの言語的な意味を評価することで，クエリ拡張に適した辞書を作成

する．本研究では，GPT-3.5-turbo[1] (以下，GPT-3.5) と PaLM2 chat-bison[2] (以下，PaLM2) を LLM として用いた．これらの LLM は幅広い Web ページを用いて学習されているため，一般的な言語的意味に基づく関連性の評価が期待できる．本研究では，図 1 に示すプロンプトを用いて zero-shot で LLM に問い合わせを行い，単語間の関連度を評価する．実験では，一組の単語ペアに対して最大で 3 回の評価を行い，その平均値を関連度として用いる．

> 一般的な日本語の知識に基づいて、EC サイトの検索キーワードの分析をしてください。入力される 2 つの検索キーワードの意味が同じかどうかを推定してください。2 つの単語の関連性を 1 ～ 5 段階で評価してください。1 は関連性が低く、5 になるにつれて関連性が高いことを表しています。「id term1 term2」の形式で入力されるので、「id:score」の形式で出力してください。

図 1: 単語の関連性評価用のプロンプト

4 評価実験

4.1 実験設定

クエリ拡張の辞書を作成するために用いる検索クエリ - 商品クリックのユーザー行動データは，モノタロウ[3] の 2022 年 1 月 1 日から 2022 年 12 月 31 日のログから作成した．式 (1) のパラメータは $\tau = 0.5$ とし，GPT-3.5 及び PaLM2 の温度パラメータ t と TOP-P パラメータ p_{P} はそれぞれ $t = 0.8$, $p_{\mathrm{P}} = 0.8$ とした．ユーザー行動から獲得できた単語ペア候補は約 2,400,000 件で，そのうちユーザー行動の類似度の上位 55,000 件を LLM で関連度を評価した．各 LLM の関連度が 3 以上の単語ペアをクエリ拡張として利用可能な単語ペアとして定め，GPT-3.5 及び PaLM2 を適用してフィルタリングすると，それぞれ 34,031 件，50,657 件の単語ペアが得られた．

4.2 辞書の評価

表 1 に獲得した辞書の単語ペアの実例を示す．

表 1 の結果から，LLM が表記ゆれ（例 1）や専門的な言い換え（例 2）の関係を適切に獲得できていることを確認できる．さらに，例 3 では口語的なクエリの意味関係を評価していて，例 4 ではルンバという固有商品名と一般商品名の関係も獲得できている．また例 5〜8 の不適切と考えられる事例から，過度に一般化している単語ペアについても高い関連度を与えていることが確認できる．

GPT-3.5 と PaLM2 の利用可能な単語ペアをそれぞれランダムに 50 件サンプルし，著者らが目視で評価したところ，GPT-3.5 は 92% (46/50)，PaLM2 は 90% (45/50) の Precision であることが確認できた．

表 1: GPT-3.5 と PaLM2 両方で関連度の高い事例

判定	例	単語 1	単語 2
適切	1	ふたつきバケツ	フタ付きバケツ
	2	ホロセットボルト	六角穴付止めネジ
	3	o 型	丸型
	4	お掃除ロボット	掃除機ルンバ
不適切	5	足湯	足湯用バケツ
	6	チラシ	チラシポスト
	7	オイスターナイフ	貝
	8	安全刃折器ポキ L 型	折

4.3 検索への応用

次に実際の E コマースの商品検索ログを用いて，クエリ拡張の効果を定量的に評価する．モノタロウにおける 2023 年 1 月 1 日から 2023 年 11 月 30 日の期間の検索ログのうち，検索回数が一定以上かつ検索結果件数が少ない約 15,000 件の検索クエリを対象クエリとした．対象クエリが含まれる対象期間中の検索セッションでバスケットに追加された商品を集計し，上位 200 件の商品を関連商品集合とした．ドキュメントとして，モノタロウの商品データを約 4,400,000 件を用いた．

検索クエリと商品データのマッチングは BM25 [7] を用いて行った．BM25 のインデックスには商品名の形態素解析した表層系 (surface) を利用し，数詞と記号は除外した．ハイパーパラメータ k_1, b は，$k_1 = 2.0$, $b = 0.75$ とした [7]．なお，形態素解析には Sudachi[4] [19] を用いた．

検索クエリ q のトークンとして，スペース区切りの単語集合と形態素の和集合を用いた．すなわち検索クエリは以下で表せる．

$$Q = \{q \text{ の形態素}\} \cup \{q \text{ のスペース分割}\} \quad (3)$$

クエリ拡張は，検索クエリに含まれる単語 $t \in Q$ と関連のある単語集合 $\mathrm{Dict}(t)$ が存在するとき，下記のように処理する．

$$Q_e = Q \cup \bigcup_{t \in Q} \mathrm{Dict}(t) \quad (4)$$

1) https://platform.openai.com/docs/models/gpt-3-5
2) https://cloud.google.com/vertex-ai/docs/generative-ai/model-reference/text-chat
3) https://www.monotaro.com/
4) https://github.com/WorksApplications/SudachiPy

表 2: 定量評価

手法	Precision@100	Recall@100
base	6.03%	25.93%
log	6.29%	26.51%
GPT-3.5	6.31%	**26.73%**
PaLM2	**6.33%**	26.48%

BM25 の商品ランキングと関連商品集合の関連度は Precision@k と Recall@k ($k = 100$) を用いて評価した．ベースライン手法としてクエリ拡張を行わない base，ユーザー行動ログに基づいた単語の類似度 $S_T(t_i, t_j)$ から作成した辞書を用いた log を評価した．log で用いる単語辞書は $S_T(t_i, t_j) > 0.8$ を満たす 20,000 件の単語ペアをランダムサンプルして作成した．提案手法として LLM で作成した辞書を用いた GPT-3.5 / PaLM2 を評価した．また，GPT-3.5 / PaLM2 の辞書は利用可能な単語ペアからそれぞれ 20,000 件の単語ペアをランダムにサンプルして作成した．

評価結果を表 2 に示す．base と log，GPT-3.5, PaLM2 の比較より，クエリ拡張を行うことで，Precision と Recall が向上することが確認できる．また，log に比べ，GPT-3.5 では Precision と Recall，PaLM2 では Precision の改善が確認できる．これは LLM で言語的な意味を評価することで，より高精度な辞書を作成できたためであると考えられる．

4.4 ベクトル検索との比較

最後にクエリ拡張とベクトル検索の実際の検索結果を比較して定性的に評価する．ベクトル検索エンジンとして，モノタロウの商品ページで Fine Tuning した Google の Vertex AI Search[5] を用いた．表 3，4 にそれぞれ「エビ金具」と「表彰筒」の検索結果の上位 3 件の商品名を示す．

表 3 のエビ金具とは，トラックのあおりを固定するための道具で，「バネカン」や「エビ金ハンドル」とも呼ばれる．作成した辞書では，「エビ金具」と「バネカン」や「エビカン」の対応を獲得できていたので，クエリ拡張では適切な検索結果となっている．ベクトル検索の検索結果はすべて「エビ金具」とは無関係な商品となっていて，ベクトル検索が必ずしも単語の同義関係を検索結果に反映できないことが示唆される．

表 4 の例では，クエリ拡張は 2 番目の検索結果は正しいが，1，3 番目の検索結果は表彰状であるため不適切である．これはテキストマッチで，表彰という単語が表彰状と多くマッチしていることが原因であると考えられる．ベクトルでマッチした商品はすべて表彰状をしまうための筒を提示しているため精度が高く，検索クエリ全体の意味を考慮する例では有用であると示唆される．

表 3: 「エビ金具」の検索結果

検索方法	検索結果 (上位 3 件の商品名)
クエリ拡張	1. エビカン 2. バネカン受金具 (鍛造品) ステンレス 3. バネカン 小
Vertex AI	1. 固定金具樹脂プラグ 2. カールプラグ 3. セキュリティハンガーセット

表 4: 「表彰筒」の検索結果

検索方法	検索結果 (上位 3 件の商品名)
クエリ拡張	1. 表彰状 2. 表彰状用 紙筒 3. 表彰状用紙
Vertex AI	1. 丸筒ワニ皮 2. 角筒ワニ皮 3. ワニガワ丸筒

5 おわりに

本研究では，E コマースにおけるクエリ拡張のための辞書作成方法を提案し，作成した辞書の評価を行った．商品検索におけるユーザー行動ログから単語の候補を獲得し，LLM を用いて単語の関連度を評価することで，E コマースのクエリ拡張に適した辞書を作成した．作成した辞書の定性評価で辞書の妥当性を，実際の検索データを用いた定量実験で検索指標の改善を確認した．また，作成した辞書を用いたクエリ拡張とベクトル検索の検索結果を比較して，明示的な言語関係を用いる有用性を示唆した．

今後の課題として，本研究の言語資源は単語間の関連度の評価にとどまっているため，単語間の同義性，上下関係も評価することでより多機能な辞書の作成が挙げられる．また実際の商品検索サービスに活用する際には，ユーザー行動のフィードバックから不適切な単語関係を削除し，継続的に辞書の品質を改善する仕組みづくりも重要である．そのために実際の E コマースにおける商品検索サービスにおいて，作成した辞書でクエリ拡張を適用し，ユーザー体験の向上に寄与できているかを評価していく予定である．

5) https://cloud.google.com/generative-ai-app-builder/docs/try-enterprise-search

参考文献

[1] Prajyoti Lopes and Bidisha Roy. Dynamic recommendation system using web usage mining for e-commerce users. **Procedia computer science**, Vol. 45, pp. 60–69, 2015.

[2] Claudio Carpineto and Giovanni Romano. A survey of automatic query expansion in information retrieval. **ACM Comput. Surv.**, Vol. 44, No. 1, jan 2012.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[4] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[5] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.

[6] 浅野孝平, 稲田和明, 張信鵬. ユーザ意図を考慮したEコマースにおける商品検索クエリの調査と分析. 言語処理学会 第 29 回年次大会, 2023.

[7] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. **Found. Trends Inf. Retr.**, Vol. 3, No. 4, p. 333–389, apr 2009.

[8] Carolyn J. Crouch and Bokyung Yang. Experiments in automatic statistical thesaurus construction. In **Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '92, p. 77–88, New York, NY, USA, 1992. Association for Computing Machinery.

[9] Aritra Mandal, Ishita K Khan, and Prathyusha Senthil Kumar. Query rewriting using automatic synonym extraction for e-commerce search. In **Proceedings of eCOM Workshop@the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval**, 2019.

[10] Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. Query expansion by prompting large language models, 2023.

[11] Liang Wang, Nan Yang, and Furu Wei. Query2doc: Query expansion with large language models, 2023.

[12] Shirui Wang, Wenan Zhou, and Chao Jiang. A survey of word embeddings based on deep learning. **Computing**, Vol. 102, pp. 717–740, 2020.

[13] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[14] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[15] Peter West, Chandrasekhar Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. Symbolic knowledge distillation: from general language models to commonsense models. In **North American Chapter of the Association for Computational Linguistics**, 2021.

[16] 井手竜也, 村田栄樹, 堀尾海斗, 河原大輔, 山崎天, 李聖哲, 新里顕大, 佐藤敏紀. 人間と言語モデルに対するプロンプトを用いたゼロからのイベント常識知識グラフ構築. 言語処理学会第 29 回年次大会, 2023.

[17] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. A survey on in-context learning, 2023.

[18] Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, and Hang Li. Context-aware query suggestion by mining click-through and session data. In **Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining**, pp. 875–883, 2008.

[19] Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. Sudachi: a japanese tokenizer for business. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Paris, France, may 2018. European Language Resources Association (ELRA).