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

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

Eコマースの商品検索において，クエリに含まれるユーザの意図を正しく解釈することは重要あり，検索結果にその意図を適切に反映することで，さらなるユーザ体験の向上につながると考えられる．本研究では，Eコマースの商品検索におけるログデータを用いて検索されたクエリをクラスタリングし，その結果に対して詳細な分析を行うことで，クエリに含まれるユーザ意図についての理解を深める．さらに，類似した意図を持つクエリのひとつのバリエーションである表記ゆれの問題に着目し，ユーザーの意図を正しく反映できるクエリ集合が作成できたことを示す．

1　はじめに

近年Eコマースが活発に利用されることに伴い，日々膨大なログデータが蓄積されている．蓄積されたログデータは，各Eコマースの商品検索や推薦といったサービスにおけるユーザ体験の向上に活用されている．商品検索では，ユーザが入力したクエリにマッチする検索結果を提示することが重要であるが，入力されるクエリにはユーザの意図が反映されることで多くの曖昧性が含まれているため，クエリからユーザの意図を正しく読み取って検索結果に反映することは難しい．さらに，クエリに含まれる意図の推定精度が十分に高くない商品検索サービスを提供すると，ユーザへ悪い印象を与える可能性があり，コマースサービスの全体的な信頼度の低下やCuster Lifetime Valueの低下を招くと恐れがある．そのため，実際の商品検索のサービスとして提供するためには，非常に高精度なクエリの意図推定が必要となると考えられる．

商品検索におけるクエリの分析として，様々な研究が実施されている．Guoらは，クエリに適切な分割を与えたり，逆に結合したりすることで，検索精度を向上させるクエリ整形の手法を提案している[1]．中山らは，日本語のEコマースにおける商品検索のクエリに対して，商品の詳細を示す属性値の抽出・活用したクエリ整形を実施しているが，その調査規模は限定的である[2]．このように，特に日本語を対象としたEコマースの商品検索におけるクエリの理解に関する研究は少なく，クエリに生じる曖昧性に関する分析も十分であるとは言えない．またクエリの分析には，ユーザ意図の類似したクエリをクラスタリングするアプローチが有用であることが知られており[3]，Caoら[4]のクラスタリングをベース手法として採用した．

そこで本論文では，Eコマースの商品検索におけるクエリと商品クリックなどのログデータの関係に着目してクエリをクラスタリングすることで，意図が類似したクエリの集合を抽出し，その特徴を分析してクエリの理解を深める．目視の観察によって，同じ商品を求めているクエリには，表記ゆれや言い換えによる曖昧性によって異なる表現や，絞り込みを目的とした属性値，意味的な含意が生じているクエリペアが存在することを確認した．さらに，分析によって発見した表記ゆれの問題に対して，高い精度の同義関係の判定ルールを提案し，検証を行った．

2　クエリのクラスタリング

2.1　クエリクラスタリングのベース手法

本研究では，Caoら[4]のクエリクラスタリング手法をベースとする．Algorithm 1にCaoらのクエリクラスタリング手法を示す．Caoらは，クエリと結びつきのある商品ページのURLを重み付き2部グラフとみなし，2部グラフが類似しているクエリ間をクラスタ化している．

Algorithm 1 Query clustering

Input: the query set Q, the diameter threshold $D_{\max}$
Output: the set of clusters Θ

$dimarr[d] \leftarrow \varnothing$ (for $d = 1, \ldots, m$)
for query $q_i \in Q$ **do**
  $C_{\text{cand}} \leftarrow \varnothing$
  for $d \in \{j : \vec{q}_i[j] \neq 0\}$ **do**
    $C_{\text{cand}} \leftarrow C_{\text{cand}} \cup dimarr[d]$
  end for
  $C \leftarrow \arg\min_{C' \in C_{\text{cand}}} \text{dist}(q_i, C')$
  if diameter$(C \cup \{q_i\}) \leq D_{\max}$ **then**
    $C \leftarrow C \cup \{q_i\}$
    update the centroid and diameter of C
  else
    C = new cluster$(\{q_i\})$
    $\Theta \leftarrow \Theta \cup C$
  end if
  for $d \in \{j : \vec{q}_i[j] \neq 0\}$ **do**
    if $C \notin dimarr[d]$ **then**
      link C to $dimarr[d]$
    end if
  end for
end for
return P

以下に，Algorithm 1 で用いる表記について記す．クエリの集合を $Q = \{q_1, \ldots, q_n\}$，商品の集合を $P = \{p_1, \ldots, p_m\}$ とする．n, m はそれぞれクエリ数と商品数である．クエリ q_i による商品検索によってクリックされた商品 p_j の間にはパス $e_{i,j}$ が発生し，そのパスのクリック数を $w_{i,j}$ とする．また，クエリの集合であるクラスタを $C_k \in \Theta$ を $C_k = \{q_1, \ldots, q_l\}$ とする．k, l はそれぞれクラスタ ID とクラスタに含まれるクエリ数である．

各クエリ q_i のベクトル表現 $\vec{q}_i \in \mathbb{R}^m$ の j 番目の要素は，式 (1) で定義される．

$$\vec{q}_i[j] = \begin{cases} \dfrac{w_{i,j}}{\sqrt{\sum_{\forall j} w_{i,j}^2}} & (e_{i,j}\text{が存在する場合}) \\ 0 & (\text{上記以外}) \end{cases} \tag{1}$$

クラスタ C_k のセントロイドベクトル $\vec{c}_k \in \mathbb{R}^m$ は，クラスタ C_k に属するクエリベクトル $\{\vec{q}_1, \ldots, \vec{q}_l\}$ の L_2 正規化された平均ベクトルで求められる．dist(q, C) と diameter(C) はそれぞれ式 (2), (3) で定義される．

$$\text{dist}(q, C) = \sqrt{\sum_{p_j \in P} (\vec{q}[j] - \vec{c}[j])^2} \tag{2}$$

$$\text{diameter}(C) = \sqrt{\frac{\sum_{q \in C} \sum_{q' \in C} \|\vec{q} - \vec{q}'\|^2}{|C|(|C| - 1)}} \tag{3}$$

また，$D_{\max}$ はクラスタサイズを調整するためのハイパーパラメータである．

表 1　クラスタリング結果

	Cao らの手法	拡張後
クラスタ数	80,283	79,472
クラスタの平均クエリ数	2.19	4.63

2.2　クラスタの拡張

Algorithm 1 によるクラスタリングでは，ひとつクラスタに含まれるクエリがひとつのみあるシングルトンクラスタが多数生成される [4] が，シングルトンクラスタに属するクエリは，比較対象のクエリが存在しないため意図分析が困難である．そこで本研究では，類似したクラスタをさらにグループ化してシングルトンクラスタの数を減らすことで，解析対象のクラスタ数の拡大を行う．

各クラスタのセントロイドベクトル $\vec{c}_k$ において $\vec{c}_k[j] \neq 0$ を満たす次元を 1，それ以外の次元を 0 とするバイナリベクトル b_k を作成する．ある二つのクラスタ C_x, C_y について，$\vec{b}_x = \vec{b}_x \odot \vec{b}_y \neq \vec{0}$ を満たすとき，C_x と C_y が関連しているとみなす．さらに $\vec{b}_k$ に対して $\vec{b}_k \odot \vec{b}_{k'} = \vec{b}_k$ を満たす $C_{k'}$ が Θ に存在しない場合，C_k を極大クラスタと定義する．ある極大クラスタ C_k を基準として，式 (4) で得られるクラスタの和集合を，類似したクラスタをグループ化した **拡張クラスタ** $\mathcal{C}_k$ を獲得する．

$$\mathcal{C}_k = \bigcup_{i \in \left\{i:\, \vec{b}_k \odot \vec{b}_i = \vec{b}_i\right\}} C_i \tag{4}$$

同様に極小なクラスタを基準とした拡張クラスタも獲得する．

2.3　クラスタリング結果

クラスタリング対象のクエリとして，monotaro.com[1] の 2022 年 1 月〜12 月における検索頻度の高いクエリ約 17 万 5 千件を採用した．なお前処理として，公序良俗に反する表現などを含むクエリ，信頼性の低いと考えられるクリック数の少ないクエリ-商品間のパス，各クエリにおいて他の商品よりも相対的に重みの小さいクエリ-商品間のパスを削除している．

表 1 に Algorithm 1 のクラスタリング結果と，2.2 節のクラスタ拡張を適用した結果を示す．なお，$D_{\max} = 0.75$ とした．

導出された拡張クエリクラスタの例を表 2 に示

1）https://www.monotaro.com/

表 2 拡張クエリクラスタの例

例 1	はんだコテ, 半田ごて, はんだごて はんだこて, はんだ小手, ハンダこて ハンダゴテ, 半だごて, 半田ゴテ ハンダゴテ□精密
例 2	コンパウンド□バフ, ポリシャースポンジパッド ウレタン□バフ, スポンジバフ□ 180, パフ スポンジパッド, bafu, モノタロウ□バフ バフ□仕上げ
例 3	ステンレス□千枚通し, 千枚通し, 千枚どうし, きり, 錐
例 4	職人大学, 軍手□薄手, 薄手軍手
例 5	六角ボルト□全ねじ□ステンレス , sus □ m8 六角ボルト□ m6 □ステンレス m8 □ボルト□ sus m10 □ボルト□ sus

表 3 類似した意図を持つクエリの観察結果

分類名（クラスタ数）	例
表記ゆれ (39)	台車 / daisha / だいしゃ 単三電池 / 炭酸電池 バッテリー / バッテリ
言い換え (81)	錐 / 千枚通し 尿素水 / アドブルー 三相スイッチ / bs230b3
絞り込み (72)	六角ボルト□ m12 × 45 t シャツ□モノタロウ 軍手□綿 100%
含意 (39)	八角皿 / 中華食器 トラテープ / トラテープ□屋外
その他 (31)	vg32 □オイル / オイル□ vg32 フィルム□ヒーター / usb ヒーター

す．例 1 と例 2 から，ひとつの商品を指す言葉に様々なクエリが存在していることがわかる．また例 3 と例 4 では，俗称や固有商品名を用いて，ひとつの商品を異なる表現で表現で検索されていることが確認できる．さらに例 5 では，クエリに商品名の「ボルト/六角ボルト」に加え材質である「sus」や寸法を指す「m8」などの属性値が使用されていることがわかる．ここで，「m6」，「m8」，「m10」はことなる寸法を指す属性値であるが，monotaro.com では属性値のみが異なる同一商品を，まとめてひとつの商品ページで取り扱っていることが原因である．以上の観察結果から，ユーザ意図が類似したクエリをグループ化できていることが確認できた．

なお以降の分析では，$|\mathcal{C}| > 1$ を満たす拡張クエリクラスタに含まれるクエリのみを用いた．

3 クエリクラスタの分析

類似した意図を示すクエリにどのような事例があるかを調査するために，拡張クラスタからランダムサンプルした 200 個のクラスタに対して目視チェックを実施した．本研究では，類似した意図を示すクエリに表れる事例を，表記ゆれ，言い換え，絞り込み，含意，その他の 5 種類に分類し，各事例の観察結果を表 3 に示す．ただし，ひとつのクラスタに複数の事例が混在しているため，表 3 のクラスタ数の合計は 200 を超える点に注意されたい．

表記ゆれ 表記ゆれとは，文字の変換の有無，誤字，長音や促音の有無，濁音・半濁音の差異によって，同じ単語を指す表現が異なる文字列となっていることを指す．テキストベースの TFIDF や BM25 を用いた一般的な全文検索では，表記ゆれの生じたクエリ間の検索結果がそれぞれ異なる [5, 6]．また，表記ゆれの差異による意図の差はほぼなく，基本的にあるひとつの意図を示していると考えられる．そのため表記ゆれにより，十分な検索件数が得られない場合や，ユーザが意図しない検索結果を提示してしまうことは，ユーザ体験の大きな悪化に繋がると考えられるため, 4 章で表記ゆれに焦点を当てた同一意図のクエリ抽出について詳しく分析する．

言い換え 言い換えとは，あるひとつの商品を指す表現が複数存在することを指す．たとえば，商品の一般的な名称と固有商品名による検索，業界固有の商品呼称による検索，型番による検索などによって発生する．特に業界固有の商品呼称に関しては，商品説明文などに明示的に記述されていたり，言語資源が存在したりしないため，言い換えの中でも特に困難な問題であるといえる．

絞り込み 絞り込みとは，商品の絞り込みを目的としたブランドや大きさ・色などの属性値の有無によって，クエリとしての表現が異なる物を指す．2.3 節で説明したとおり，monotaro.com では属性値のみが異なる商品をまとめてひとつの商品ページで取り扱っていることで，類似した意図を持つクエリとして複数の事例が観測された．

含意 含意とは，クエリ内のある表現がすでに他の表現を含意していることで，異なるクエリでも同じ意図を示す事例を指す．たとえば「雨傘 / ジャンプ傘」では，ジャンプ傘は雨傘の一種であるため，雨傘がジャンプ傘を含意しているといえる．意味的含意の判断には，ドメイン知識を反映させた WordNet [7] のような意味関係を持つ言語資源の作成が必要であり，対応が困難な問題であると言える．

その他　また上記以外にも，スペースの有無や単語の並び順が異なるだけの軽微な差異である事例が 18 件，分類自体が困難な事例が 13 件存在した．分類が困難な事例は，2 部グラフでユーザの意図を適切に捉えられていないようなクリックログや商品検索結果を持つクエリであり，本手法で抽出した類似していない意図を持つクエリ集合，すなわち負例といえる．

4　同じ意図の表記ゆれクエリの獲得

4.1　表記ゆれの判定

まず，表記ゆれによって表層が異なるが同じ意図を指すクエリ郡を獲得する基本的な対応方法として，クエリのカナ読みに着目する．しかし，「さらさ」という読みを持つ商品には以下の三つ異なる商品が存在するため，同じ読みの意図が異なるクエリを同じ意図として判定してしまう危険性がある．

- さらさ（P&G 社の洗剤）
- サラサ（ゼブラ社のボールペン）
- SARASA（ファロス社の鍼）

そこで表記ゆれによる同じ意図のクエリ獲得では，2.2 節で獲得した拡張クラスタ内 $\mathcal{C}$ で行う．拡張クラスタ内のクエリはユーザーの意図が類似してるため，上述の三つのクエリをひとつにまとめた不適切なクエリ集合の抽出を防ぐことができると考えられる．

本研究では，クエリをスペース区切りで分割 $q = [t_1, \ldots, t_L]$ した上で，拡張クラスタ内の全てのクエリの構成要素のペア (t_i, t_j) に対して，カナ読みの完全一致と編集操作 [8]（挿入・削除・置換）を用いた比較を適用する．編集操作として，「エアチューブ」と「エアーチュウブ」間のような表記ゆれを吸収できるように，ふたつの文字列間における長音の挿入・削除・置換，かな読みの大文字小文字の置換，濁音・半濁音への置換といった特定の文字の操作を許容した．なお，カナ読みの付与には SudachiPy[2] を用い，t が英数字のみの文字列である場合，t の訓令式ローマ字とみなしてカナ読みを求める．

表 4　表記ゆれの解消

例 1	オイルフィルター, oirufiruta, オイルフィルタ, オイルフイルター, oil フィルター
例 2	タフレッド, 田フレッド, タフレット

4.2　分析結果

2.3 節のクエリに対して，提案した「表記ゆれ」による類似した意図を示すクエリ集合の抽出手法を適用したところ，5,468 ペア，12,892 クエリが抽出された．表 4 に抽出された表記ゆれ表現のペアの例を示す．

表 4 の例 1 では，「オイルフィルター」の意図を持つ表記ゆれのクエリ郡をひとつのグループとして集約できていることが確認できた．一方例 2 では，「タフレッド」はアトム社が販売するゴム手袋の商品名，「タフレット」は NACHI（不二越）社が販売するロールタップの商品名であるため，異なる意図のクエリを集約してしまった．クエリ「タフレット」における商品検索結果で，ユーザーが「タフレッド」の商品をクリックすることが多数あったことが，クラスタの誤判定の原因であるとわかった．

5　おわりに

本論文では，ユーザ行動に基づいたクラスタリング手法を用いて，E コマースの商品検索におけるクエリの分析を行った．クエリクラスタリングによってユーザ意図が類似したクエリをグループ化することができ，類似したクエリとして表記ゆれや商品名の言い換えなどが含まれていることと，それぞれの課題について確認した．さらに，表記ゆれが生じている単語の同義関係の判定方法を提案した．

今後の課題として，より多くのクエリの意味関係を捉えるためには，本研究の分析で発見した言い換えや含意についても考慮する必要がある．そのためにも，言い換えや含意，業界固有の商品呼称に関する言語資源の作成が長期的な課題であると言える．また，クエリに含まれる絞り込みを目的とした属性値について正しく同定し，検索結果に反映することで，より良い検索体験の提供につながることが期待できる．

2）　https://github.com/WorksApplications/SudachiPy

参考文献

[1] Jiafeng Guo, Gu Xu, Hang Li, and Xueqi Cheng. A unified and discriminative model for query refinement. In **Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval**, pp. 379–386, 2008.

[2] 中山祐輝, Chen Zhao, Erick Mendieta, 村上浩司, 新里圭司. E コマースにおける検索クエリの整形と属性値抽出への適用. 言語処理学会 第 28 回年次大会 発表論文集, pp. 1578–1582, 2022.

[3] Claudio Carpineto and Giovanni Romano. A survey of automatic query expansion in information retrieval. **Acm Computing Surveys (CSUR)**, Vol. 44, No. 1, pp. 1–50, 2012.

[4] 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.

[5] Yuki Amemiya, Tomohiro Manabe, Sumio Fujita, and Tetsuya Sakai. How do users revise zero-hit product search queries? In **Advances in Information Retrieval**, pp. 185–192. Springer International Publishing, 2021.

[6] Gyanit Singh, Nish Parikh, and Neel Sundaresan. Rewriting null e-commerce queries to recommend products. In **Proceedings of the 21st International Conference on World Wide Web**, pp. 73–82, 2012.

[7] George A Miller. Wordnet: a lexical database for english. **Communications of the ACM**, Vol. 38, No. 11, pp. 39–41, 1995.

[8] Dan Gusfield. **Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology**. Cambridge University Press, 1997.