ウェブ検索における類似例を用いた検索意図明確化手法

Query Disambiguation by Using Similar Examples in Web Search

田島　敬史（Keishi TAJIMA, Dr. Sci.）
京都大学　情報学研究科　社会情報学専攻　教授
(Professor, Department of Social Informatics, Graduate School of Informatics, Kyoto University)
Association for Computing Machinery　情報処理学会　日本ソフトウェア科学会
研究専門分野：Web 情報処理　ソーシャルネットワーク分析　情報検索

あらまし

現在のウェブ検索エンジンが不得意な状況の一つに，検索式が複数の意味に解釈でき，ユーザが意図する解釈がそのうちの主要なものではない場合がある．例えば，とうもろこしで肥料を作る方法を求めて「とうもろこし 肥料」という検索を行った場合，検索解上位は「とうもろこしのための肥料」に関するもので占められ，望む情報が得られない．本稿では，このような場合に，意図する解釈と文脈が似ている検索（例：「生ごみ 肥料」）および意図しない解釈と文脈が似ている検索（例：「あさがお 肥料」）をユーザに指定させ，この情報を用いて検索意図に合う解を上位にランクする手法を提案する．提案手法では，まず，似ている検索の解中では検索語周辺に頻出するが，似ていない検索の解中では頻出しないようなフレーズ（例：「○○から肥料を」「○○を肥料に」）を抽出し，これらを用いたフレーズ検索で解候補を取得する．次に，似ている検索の解集合を正例，似ていない検索の解集合を負例とみなし，これらの情報を用いて候補解をランキングする．主要でない検索意図による検索では本手法が有効であることを実験により確認した．

1．研究の目的

ウェブ検索エンジンなどの，現在，広く用いられている文書検索システムでは，一つまたは複数の検索語をクエリ（検索式）として入力し，それらに適合すると思われる文書がランキングの形で表示される方式が一般的である．ユーザが求める情報を上位に出すランキング技術は長年研究されており，現在のウェブ検索エンジンは，多くの検索に対して適切な文書を上位にランクすることができる．

しかし，現在のウェブ検索エンジンが適切な文書を上位に表示できないような種類の検索も依然として存在する．そのような種類の検索の一つに，ユーザのクエリがどのような検索意図を表しているのかについて複数の解釈が可能で，かつ，ユーザが意図している解釈が，それら複数の解釈の中であまり主要ではない場合というものがある．例えば，余ったとうもろこしで肥料を作る方法を知りたいユーザが「とうもろこし 肥料」というクエリで検索を行う場合を考える．この時，この二語を含む文書のトピックとしては，「とうもろこしで肥料を作る」というトピックよりも「とうもろこしに与える肥料」というトピックの方が一般的である．実際，2020 年 4 月 20 日時点で，Google で「とうもろこし 肥料」というクエリを用いて検索を行った結果，上位 100 件の検索結果のほとんどが，とうもろこしに肥料を与える話題をなんらかの形で含むものであり，とうもろこしから肥料を作る話題は（土地改良材として用いる話の一件を除くと）一件もない（図 1）．

図 1：「とうもろこし 肥料」による検索結果．とうもろこしに与える肥料に関する文書で占められており，トウモロコシから肥料を作る話題はない．

そのため，このユーザは，このクエリでは求める情報を得られない．

ウェブ検索における類似例を用いた検索意図明確化手法

Query Disambiguation by Using Similar Examples in Web Search

このような，クエリが多義的であり，検索結果上位が意図するものとは異なる解釈に対応するもので占められてしまう場合に，われわれがよく行うのは，検索語を追加することで，クエリの曖昧性を低減することである．この手法は，「キーボード」のように全く分野が異なる複数の意味（PC 周辺機器と楽器）を持つ多義語の曖昧性解消には有効で，例えば，「キーボード 楽器」などのクエリに変更することで曖昧性を低減できる．しかし，「とうもろこしから肥料を作る」と「とうもろこしのための肥料」は非常に近い分野に属する話題であるため，適切な追加検索語を発見することは難しい．

多義的検索の意図明確化に用いられるもう一つの手法として，フレーズ検索がある．フレーズ検索とは「とうもろこしから肥料を」のように，複数語からなるフレーズを用いる検索である．Google などの一部の文書検索システムでは，検索語や検索フレーズを二重引用符で囲むことにより完全一致を要求する機能も利用可能であり，「"とうもろこしから肥料を"」のようにフレーズによる完全一致検索を用いると，このフレーズがこの形で現れる文書を検索できる．しかし，2020 年 4 月 20 日時点で，Google で「"とうもろこしから肥料を"」で検索すると，このフレーズが現れる文書はないという結果が示される（図 2）．

図 2：「"とうもろこしから肥料を"」というフレーズによる完全一致検索の検索結果．完全一致する文書は 1 件もなく，完全一致でない検索結果がその下に示されているが，やはり，とうもろこしに与える肥料に関する文書で占められている．

なお，Google ではこのような場合，自動的に完全一致を要求しない検索として「とうもろこしから肥料を」が実行されるが，その検索結果の上位は，やはりとうもろこしのための肥料に関するもので占められる．そこで，次に「"とうもろこしを肥料に"」というフレーズで検索してみると，3 件の解が表示され，うち 1 件は，とうもろこしを肥料にする話が含まれる文書である．（残り 2 件は，肥料ではなく飼料にする話で肥料と誤記しているものである．）

このように，完全一致と組み合わせたフレーズ検索はクエリの曖昧性の低減に有効だが，一方，検索意図が絞り込まれるようなフレーズは解として得られる文書数が大幅に少ないことが多く，解が 0 件という場合も頻発する．すなわち，適合率（＝解として表示した文書のうちの正解の比率）は高いが，再現率（＝存在する正解である文書のうち解として表示できたものの比率）が低い．そのため，求める情報を得るには，上述の例のように，有効と思われるフレーズを多数案出し，それらを一つずつ試していく作業が必要となることが多い，しかし，存在する正解文書の中で検索語がどのようなフレーズで出現するかは多様であり，これを予測して解が 0 件とならない有効なフレーズを発見する作業はユーザにとって大きな負担となる．

本研究では，この例のように，多義性を排除したクエリを考えることが容易でなく，かつ，意図する解釈が主要な解釈ではないために，求める情報を得るのが難しいという場合の検索を支援する手法を提案する．

２．提案手法の概要

提案手法では，まず，意図する解釈に対応するものに検索解を絞りこめるようなフレーズを，多数，自動生成し，これらを一つずつ使って検索を行うことで解候補の集合を取得する．続いて，得られた全ての解候補を適切にランキングし，一つの解リストとして表示する．よって，以下の二つの要素技術が必要となる．

1. 意図する解釈に解を絞るために有効なフレーズを，多数，自動生成する方法
2. これらのフレーズによる検索で得られた解候補を適切にランキングする方法

この二つを実現するために，提案手法では，ユーザ

ウェブ検索における類似例を用いた検索意図明確化手法

Query Disambiguation by Using Similar Examples in Web Search

にクエリとあわせて，以下の二つを指定させる．

- そのクエリの意図する解釈と文脈が似ているクエリ
- そのクエリの意図しない解釈と文脈が似ているクエリ

例えば，上述の「とうもろこし 肥料」の例の場合，それぞれの例として以下のようなものが考えられる．

- 意図する解釈と似ているクエリ：
  「生ごみ 肥料」「麦わら 肥料」
- 意図しない解釈と似ているクエリ：
  「あさがお 肥料」「植木 肥料」

「生ごみ」「肥料」という二語が近接して出現する場合は，「生ごみから肥料を作る」という文脈で出現することが多いと予想され，よって，意図するクエリ解釈と文脈が似ていると期待される．一方，「あさがお」「肥料」という二語が近接して出現する場合は，「あさがおに与える肥料」という文脈で出現することが多いと予想され，よって，意図しないクエリ解釈と文脈が似ていると予想される．以降は，簡単のため，両者をそれぞれ「類似クエリ」「非類似クエリ」と呼ぶ．

提案手法では，これらの類似クエリと非類似クエリの情報を用いて有効なフレーズを生成する．まず，これらの検索それぞれのランキング上位解となる文書の集合を取得する．そして，類似クエリの解中では検索語周辺に頻出するが，非類似クエリの解中では頻出しないようなフレーズを抽出する．例えば，「○○から肥料を」というフレーズは，「生ごみ 肥料」という検索の解集合中にも，「麦わら 肥料」という検索の解集合中にも，共通して，「生ごみから肥料を」「麦わらから肥料を」という形で頻出するが，一方，「あさがお 肥料」や「植木 肥料」という検索の解集合中では，「あさがおから肥料を」「植木から肥料を」というフレーズはほとんど現れない．よって，この「○○から肥料を」というフレーズは，「とうもろこし 肥料」という検索式の複数の可能な解釈のうち，ユーザが意図する解釈のみに特有な文脈を表すフレーズであると期待される．よって，このような条件を満たすフレーズを多数抽出し，意図する解釈に対応するものへ検索解を絞り込むのに有効なフレーズとして用いる．

解候補のランキングも，類似クエリ・非類似クエリの情報を用いる．類似クエリの解集合を正例の集合，非類似クエリの解集合を負例の集合とみなし，この正例集合と類似度が高く，かつ，負例集合とは類似度が低い解候補を上位にランキングする．

３．関連研究

検索意図を明確化するクエリの推薦は古くから研究されている．代表的なアプローチには，全ユーザの検索ログから関連するクエリを発見する手法[1]や，クリックスルーデータ（どの検索でどの解がクリックされたかのデータ）を用いる手法[2]があり，商用ウェブ検索エンジンでも実用化されている．これらの手法は，頻繁に用いられたクエリ中の語や，頻繁にクリックされた文書中の語を推薦する．これは，多くのユーザに頻繁に選択されるものが，現在のユーザにとっても有用である確率が高いという仮定に基づいており，本研究で対象とするような多義的クエリの主要でない解釈に対応する検索要求には有効ではない．

類義語辞書を用いて，クエリ中の語の同意語，上位語，下位語などを推薦する手法も存在する[3,4]．しかし，類義語辞書に載っている同意語，上位語，下位語などの数は少なく，多様な検索意図に対応できるような多様な推薦は生成できない．例えば，下位語を用いて多義的クエリの曖昧性低減ができるのは，意図する検索要求にちょうど対応するような下位語が存在する時に限られる．

ここまでの二つのアプローチは，検索意図を明確化する検索語を推薦する手法だが，与えられた検索の解のランキングを工夫する手法も提案されている．

まず，検索結果上位に多様な話題のウェブページを含むように，互いに異なる話題の文書を上位にランクする手法が研究されている[5,6]．多様化により，主要でない検索意図に適合する解も上位に含まれる可能性が高まると期待される．しかし，多様化の手法では，ランキング最上位にはやはりまず主要な解釈に対応するものが現れることになり，提案手法のようにユーザが意図する解釈を明示的に指定して，それが最上位に出るようにすることはできない．また，意図する解釈に対応する解が他の解釈に対応する解と混ざってランキング中に現れるため，意図する解釈に対応する解を

多数集めたい場合には，そうでない解を多数スキップしながら，ランキングの下の方まで見なければならない．

多義的クエリの検索結果をクラスタリングして各解釈に対応するクラスタに分類し，ユーザーにクラスタを選ばせる手法も研究されている[7]．しかし，主要でない意図に適合する文書数は極端に少ないことが多く，前述の「とうもろこし 肥料」の例のように，与えられたクエリの上位解に全く含まれないか，含まれても少数であることが多い．そのため，与えられたクエリの上位解をクラスタリングする手法では，当初の解の上位にそもそも適合解が含まれていない場合には対応できず，また，少数のみ含まれていた場合は，うまく一つのクラスタになりにくい．クラスタ数を多くすると分離できる可能性は高まるが，クラスタ数を多くし過ぎると，生成された多数のクラスタの中から自分の意図に適合するものを選ぶユーザの手間が大きくなる．一方，本研究では，特定の検索意図に特有なフレーズをまず発見し，これを用いて検索を行うことによって，主要でない検索意図に適合する少数の解を発見できる確率を高めている．

検索結果ランキングの改善手法としては，適合フィードバックと呼ばれる手法も広く知られている[8]．この手法では，初期検索結果上位の各文書がクエリ意図と適合しているかどうかをユーザが判定し，その結果に基づいて検索結果を再ランキングする．しかし，ユーザが判定できる上位解の数は限られ，意図する検索意図が主要でない場合は上位解の中に適合解がまったく含まれない場合も多く，その場合は適合フィードバックは機能しない．一方，本研究の提案手法では，適合解と非適合解ではなく，類似クエリと非類似クエリをユーザに指定させることにより，初期検索結果内に適合解と非適合解が含まれるかということに依存せずに，適合フィードバックと同様の効果を得ることができる．また，クエリの形で指定することにより，少数のクエリを指定するだけで，多数の文書に対して判定を行ったのと同様の効果を得られる．

適合フィードバックにおけるユーザの判定作業の負荷をなくすために，初期検索の上位解を無条件に適合とみなして再ランキングを行う疑似適合フィードバックという手法もある．しかし，疑似適合フィードバック適用後の検索結果は，初期検索結果上位で多数派となるクエリ意図に結果がさらに偏るため，主要でないクエリ意図による検索では逆効果となる．そこで，後述の実験で，提案手法と疑似適合フィードバックの比較を行う際には，通常の疑似適合フィードバックとは反対に，初期検索解の上位を無条件に非適合と判定する．

４．提案手法の詳細

本章では提案手法の詳細について述べる．ここでは，二つの検索語からなるクエリを想定して説明するが，提案手法は一つ，あるいは，三つ以上の検索語からなるクエリにも適用可能である．

４．１．フレーズクエリの生成

今，二つの検索語からなる目的クエリ$Q = (k_1, k_2)$と，やはり二つの検索語からなる類似クエリ$S_1 = (s_1^1, s_2^1), \ldots, S_n = (s_1^n, s_2^n)$および非類似クエリ$D_1 = (d_1^1, d_2^1), \ldots, D_m = (d_1^m, d_2^m)$が与えられたとする．本手法では，まず初めに各類似クエリS_i，および，その検索語の順番を入れ替えたものでウェブ検索を行い，それぞれ上位最大 10,000 件（それ以下しか得られない場合もある）の文書を取得する．順番を入れ替えたものも使用するのは，今回使用したウェブ検索エンジンの結果が検索語の語順で変わるためである．そして，各文書のスニペットから，

- s_1^i, s_2^iのうち先に出現する方の検索語（s_x^iとする）の直前の 2 語（p_p）
- 二つの検索語の出現の間のフレーズ（p_m）
- s_1^i, s_2^iのうち後に出現する方の検索語（s_y^iとする）の直後 2 語（p_f）

を抽出し，以下の 8 通りのフレーズパターンを生成する．

- $X\ Y$
- $p_p\ X\ Y$
- $X\ p_m\ Y$
- $X\ Y\ p_f$
- $p_p\ X\ p_m\ Y$
- $p_p\ X\ Y\ p_f$
- $X\ p_m\ Y\ p_f$

ウェブ検索における類似例を用いた検索意図明確化手法

Query Disambiguation by Using Similar Examples in Web Search

- $p_p\ X\ p_m\ Y\ p_f$

XとYはs_x^iとs_y^iの出現位置に置かれる変数で，ここに後で特定の語を代入する．次に，得られた各フレーズパターンphの重要度$R(ph)$を以下の式で求める．

$$R(ph) = \frac{(\prod_{i=1}^{n} \#(ph[s_1^i, s_2^i])/\#(S_i))^{\frac{1}{n}}}{(\prod_{i=1}^{m} \#(ph[d_1^i, d_2^i])/\#(D_i))^{\frac{1}{m}}}$$

ここで，$\#(q)$はクエリqによる完全一致検索の解数，$ph[s_1^i, s_2^i]$はph中のX, Yそれぞれをs_1^i, s_2^iのうちの対応する方で置き換えたフレーズ，$ph[d_1^i, d_2^i]$も同様である．

上式右辺の分子はフレーズパターンphが各類似クエリの解内でその類似クエリの語とともに出現する相対頻度の幾何平均である．幾何平均を用いているため，どの類似クエリの語と組み合わせても安定した出現率で出現するフレーズパターンに対しては値が大きくなり，反対に，類似クエリの中にどれか一つでも，フレーズパターンがその類似クエリの語とともには出現しないものがあると値が 0 となる．一方，上式右辺の分母は同じものを非類似クエリについて計算している．これにより，どの類似クエリの語とも高い率で出現し，非類似クエリの語とは低い率でしか出現しないフレーズの重要度が大きくなる．この重要度Rの上位 20 件の候補フレーズを選び，その $X,\ Y$をk_1, k_2で置き換えたものを一つずつ用いて検索を行い，それぞれの上位 100 件の解を目的クエリの解候補として収集する．

しかし，類似クエリの解から抽出されるフレーズパターンの数はかなり多く，上のRを全てのフレーズパターンについて計算すると，フレーズパターンの数に比例した回数のウェブ検索が必要となりコストが高すぎるため，今回の実験では以下の方法で候補を絞り込んでからRを計算した．

まず，各phについて$ph[k_1, k_2]$による検索を行い，その結果が 0 件となるものを候補から除外する．これは，そのようなフレーズパターンは，後で$ph[k_1, k_2]$を用いて解候補を取得する際に解が 0 件となり有用ではないからである．そして，残った全てのフレーズパターンについて，下記のスコアSを計算する．

$$S(ph) = (\prod_{i=1}^{n} N(S_i, ph[s_1^i, s_2^i]))^{\frac{1}{n}}$$

$N(q,p)$はクエリqの検索解（最大 10,000 件，それ以下しか得られない場合もある）中でのフレーズpの出現回数である．よって，$S(ph)$はフレーズパターンphの各類似クエリの解中での出現数の幾何平均である．これは，フレーズパターンの個数とは関係なく，各類似クエリの検索結果の最大 10,000 件を取得すれば求めることができる．このスコアSで上位 300 件のものを選び，それらについて，前述の重要度Rを計算し上位 20 件を選んだ．

４．２．候補解のランキング

続いて，解候補をランキングするために，類似クエリと非類似クエリの情報を用いて，ユーザの検索意図を表現するベクトル（クエリベクトル）を生成する．そのために，まず，ユーザが指定した目的クエリ，類似クエリ，非類似クエリを以下の手順でベクトル表現へ変換する．

まず，別の文書集合で学習済の Sentence-BERT を用意する．近年，与えられた文書集合内で各語の前後にどんな語が出現するかの情報を用いて，各語（あるいは文書全体）をその語（文書）がどんなトピック（および文構造）にどの程度関係するかを表すベクトルに変換する手法が多く開発されており，Sentence-BERT もその一つである．

続いて，目的クエリによるウェブ検索の上位解 100 件を取得し，これらをそれぞれ Sentence-BERT を用いてベクトルに変換し，これらの平均ベクトルを目的クエリのベクトル表現V_Qとする．同様の方法で，各類似クエリと各非類似クエリについてもベクトル表現$V_{S_1}, \ldots, V_{S_n}$と$V_{D_1}, \ldots, V_{D_m}$を得る．

そして，ユーザの検索意図を表現するクエリベクトルV_qを以下の式によって求める．

$$V_q = V_Q + \frac{1}{n}\sum_{i=i}^{n} V_{S_i} - \frac{1}{m}\sum_{i=i}^{m} V_{D_i}$$

つまり，目的クエリのベクトル表現に類似クエリのベクトル表現の平均を加えるとともに，非類似クエリのベクトル表現の平均を減じる．

そして，各解候補を同様に Sentence-BERT を用いてベクトル化し，このクエリベクトルV_qとのコサイン類似度を求め，これを各解候補のランキングに用いる．

本研究では，このクエリベクトルとのコサイン類似度に加えて以下の二つも用いてランキングを行う．

- その解候補を取得するのに用いられたフレーズの重要度R.
- そのフレーズをクエリとしてその解候補を取得した際の，その解候補の順位.

これは，重要度Rの高いフレーズほど，ユーザの検索意図をよりよく反映しており，また，そのようなクエリの検索結果内で上位の文書ほど，その検索意図により適合していると考えられるためである．これら，合計三つの要素の全てまたは一部を考慮して最終的なランキングを生成するが，複数の要素を用いてランキングを生成する方法の詳細については本稿では省略する．

５．実験

次に，提案手法の有効性を示すために行った実験について説明する．まず，提案手法においては，提案手法で生成した 20 個のフレーズをそれぞれ用いて検索を行い，各々の上位 100 件，合計 2000 件を候補解とした．この候補解のランキング方法としては，以下の四つを比較した．

- 前述の三要素を全てランキングに用いるもの
- フレーズの重要度Rを使用しないもの
- フレーズ検索結果内の順位を使用しないもの
- フレーズ重要度Rもフレーズ検索結果内順位も使用しないもの

比較手法としては，以下の 3 手法を用いた．

- 目的クエリによる通常の検索（Google）
- 疑似適合フィードバック
- 多様化（MMR）

疑似適合フィードバック，多様化手法のいずれにおいても，クエリと文書のベクトル化には，古典的な文書ベクトル化手法である tf-idf 法を用い，目的クエリの Google による検索結果の上位 100 件を再ランキングした．提案手法では 2000 件を再ランキングしているので，比較手法で上位 100 件のみを用いるのは不公平に見えるが，比較手法で再ランキングする範囲をさらに下位まで広げると，ノイズとなる解が増えてむしろ性能が低下するため，ここでは，上位 100 件のみを用いた．

疑似適合フィードバックでは，検索結果の上位 3 件を不適合と判定して再ランキングを行う．上位 3 件を不適合と判定するのは，本実験で用いる検索が全て，主要でないクエリ意図で検索する場合を想定しているためである．多様化手法としては MMR[5]を用いた．

実験に用いた 30 個のクエリ（全て英語）の一部を以下に示す．各クエリの右に，今回想定する意図と，括弧内に主要な意図を示し，各クエリの下に，使用した類似クエリと非類似クエリを示す．単数形と複数形のある検索語については，どちらも検索語として用い，出現数を計算する際などには両者を同一の語とみなした．

- 「corn fertilizer」：とうもろこしを肥料にする方法（≠とうもろこしに与える肥料）

  類似クエリ：「waste fertilizer」「straw fertilizer」「feces fertilizer」

  非類似クエリ：「tomatoes fertilizer」「vegetables fertilizer」
- 「Microsoft office」：Microsoft 社のオフィス（≠Microsoft 社の製品 Office）

  類似クエリ：「Sony office」「Facebook office」

  非類似クエリ：「Microsoft Azure」「Microsoft Windows」
- 「China visa」：中国での VISA カードの利用（≠中国へ，または，中国からの渡航ビザ）

  類似クエリ：「China Master」「China Amex」「China JCB」

  非類似クエリ：「Russia visa」「America visa」「EU visa」
- 「apt install」：パッケージ管理ソフト APT のインストール（≠パッケージ管理ソフト APT を用いたパッケージのインストール）

  類似クエリ：「Chrome install」「Firefox install」「Python install」

  非類似クエリ：「yum install」「pip install」「npm install」
- 「apple」：りんご（≠Apple 社）

  類似クエリ：「orange」「grape」「lemon」

  非類似クエリ：「Google」「Microsoft」「Amazon」
- 「amazon」：熱帯雨林のアマゾン（≠EC サー

ビスの Amazon）
類似クエリ：「rainforests」「animals」「Brazil」
非類似クエリ：「Google」「Microsoft」「Facebook」

これらを含む 30 個のクエリについて，各手法によって生成されたランキングを Presicion@k ($k = 5, 10, 20, 30$)，すなわち，解上位k件のうちの検索意図に適合するものの比率によって評価し，この値の全 30 クエリに対する平均値で各手法を比較する．

表 1 は前述の四つの提案手法の評価結果である．最も性能が良いのは，クエリベクトルとのコサイン類似度とフレーズ重要度Rの二つのみを用いたものである．

一方，表 2 は前述の三つの比較手法の評価結果である．提案手法の方がこれらの手法よりも良い性能を示していることがわかる．

また，表 3 は提案手法で生成したフレーズクエリの検索結果を各比較手法を使って再ランキングした場合の評価結果である．これらよりも表 1 に示した数値の方が高くなっており，提案手法の性能に，フレーズクエリによる候補解取得だけでなく，提案するランキング手法も貢献していることがわかる．

- 表 1：提案手法の Presicion@kによる評価結果

ランキング手法	$k = 5$	$k = 10$	$k = 20$	$k = 30$
V_q	0.324	0.331	0.366	0.374
V_q，R	0.428	0.410	0.402	0.399
V_q，ページ順位	0.407	0.372	0.357	0.354
V_q，R，ページ順位	0.372	0.372	0.374	0.360

- 表 2：比較手法の Presicion@kによる評価結果

ランキング手法	$k = 5$	$k = 10$	$k = 20$	$k = 30$
Google	0.145	0.131	0.129	0.128
疑似適合フィードバック	0.145	0.138	0.134	0.122
多様化（MMR）	0.145	0.124	0.124	0.129

- 表 3：提案手法によるフレーズ検索＋比較手法によるランキングの Presicion@kによる評価結果

ランキング手法	$k = 5$	$k = 10$	$k = 20$	$k = 30$
疑似適合フィードバック	0.338	0.338	0.334	0.346
多様化（MMR）	0.338	0.348	0.362	0.353

６．まとめと今後の課題

ウェブ検索に代表されるような，一つまたは複数の検索語からなる検索式を用いる文書検索においては，検索式が複数の意味に解釈できる場合があり，ユーザが意図する解釈がそのうちの主要なものでない場合は，検索解上位が主要な解釈に対応するもので占められ，望む情報が得られない．本稿では，このような場合に，意図する解釈と文脈が似ている検索，および，意図しない解釈と文脈が似ている検索をユーザに指定させ，この情報を用いて検索意図に合う解を上位にランクする手法を提案した．

提案手法では，まず，似ている検索の解中では検索語周辺に頻出するが，似ていない検索の解中では頻出しないようなフレーズを抽出し，これらを用いたフレーズ検索で解候補を取得する．次に，似ている検索の解集合を正例，似ていない検索の解集合を負例とみなし，これらの情報を用いて解候補をランキングする．複数の意味に解釈できる多義的な検索 30 件について，ユーザの意図が主要な解釈ではない場合を想定して，検索結果のランキングの評価を行った結果，そのような検索においては，提案手法が従来手法よりも優れたランキングを生成することを確認した．

このような多義的クエリの主要でない解釈を意図した検索の頻度は必ずしも高くはない．しかし，そのような検索が 1000 回に 1 回であったとしても，そのような，求める情報を見つけるのが困難な検索に要する時間が，通常の容易な検索に比べて数十倍かかるとすると，われわれが検索に費やす時間の数パーセントはそのような検索に費やされていることになる．世界中の人々が，日々，情報の検索に費やしている時間は膨大であり，その数パーセントを大幅に短縮することができれば，生産性の向上に大きく貢献することができる．

類似クエリと非類似クエリを指定することは，ある程度，検索に慣れているユーザであれば，決して難しいものではない．また，特許検索など，主に専門家が用いる検索システムにおいては，提案手法のような検索方法を理解して使用することは容易であろう．

また，提案手法を使用する際のインタフェースとしては，従来通りの検索ボックスに

ウェブ検索における類似例を用いた検索意図明確化手法

Query Disambiguation by Using Similar Examples in Web Search

とうもろこし 肥料 like:(生ごみ 肥料)　unlike:(あさがお 肥料)

のような形で入力させる形であれば，従来のインタフェースに変更は必要なく，この機能を使用しないユーザの邪魔にもならない．

さらに，提案手法は，検索結果の多様化を用いる手法などと違い，ユーザが類似クエリ・非類似クエリを特に指定しない場合は，従来の検索結果と同じものを返す．そのため，主要でない検索意図に対応するために，より主要な検索意図に対する性能を犠牲にするという問題は起こらない．

なお，適合解と非適合解の例を指定するのではなく，適合解を多く生成すると思われる検索式と非適合解を多く生成すると思われる検索式を指定することにより，より容易に多くの情報を得て検索結果ランキングを改善しようという手法は，文書検索以外にも応用可能であり，そのような応用は有望な今後の研究課題である．

謝辞

本稿で紹介した研究成果は，京都大学工学部情報学科，および，京都大学情報学研究科社会情報学専攻に在籍していた，橋本泰平氏，田中雄也氏との共同研究によるものである．

参考文献

[1] Baeza-Yates, R., Hurtado, C. and Mendoza, M.: Query recommendation using query logs in search engines, International Conference on Extending Database Technology, Springer, pp. 588-596 (2004).

[2] Cui, H., Wen, J.-R., Nie, J.-Y. and Ma, W.-Y.: Query expansion by mining user logs, IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 4, pp. 829-839 (2003).

[3] Fang, H.: A re-examination of query expansion using lexical resources, Proceedings of ACL-08: HLT, pp. 139-147 (2008).

[4] Pal, D., Mitra, M. and Datta, K.: Improving query expansion using WordNet, Journal of the Association for Information Science and Technology, Vol. 65, No. 12, pp. 2469-2478 (2014).

[5] Carbonell, J. G. and Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries, Proceedings of the 21st annual international ACM SIGIR Conference, pp. 335-336 (1998).

[6] Santos, R. L., Macdonald, C. and Ounis, I.: Exploiting query reformulations for web search result diversification, Proceedings of the 19th International Conference on World Wide Web, ACM, pp. 881-890 (2010).

[7] Leuski, Anton, Evaluating document clustering for interactive information retrieval, Proceedings of the tenth International Conference on Information and Knowledge Management, pp. 33-40 (2001).

[8] Rocchio Jr, J.: Relevance feedback in information retrieval: The Smart System-Experiments in Automatic Document Processing, ed. Salton, G (1971).

この研究は，平成２７年度ＳＣＡＴ研究助成の対象として採用され、平成２８～３０年度に実施されたものである．