

緩和度付きキーワードからなるクエリの構造に基づく質問修正

金子 恭史[†] 中村 聡史[†] 田中 克己[†]

[†] 京都大学情報学研究科社会情報学専攻 〒 606-8501 京都市左京区吉田本町

E-mail: †{kaneko,nakamura,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし 我々はこれまで、クエリ内の各キーワードに対してユーザがどの程度他の語に置き換えてほしいと感じているかの度合いを緩和度と定義し、ユーザから伝えられた緩和度とキーワード間の関係性を基にユーザの意図を推定して適切な語を補い、それらの語を基に新たにクエリを生成してより多様な話題を含む検索結果を提示する手法を提案してきた。しかし、これまでは得られた検索結果中のページをユーザの意図を考慮して並び替えることができておらず、下位に適合ページが埋もれてしまうことがあった。そこで本稿では、ユーザの意図を考慮する上で重要と考えられるクエリの構造について述べ、このクエリの構造を推定しこれに基づいてページの並び替えを行うことでユーザの意図に適合するページを上位に並び換えることを可能にする手法を提案する。

キーワード Web, 意図推定

1. はじめに

近年、情報検索をする際にキーワード入力型の Web 検索エンジンを利用することが一般化している。ユーザは対象とする情報が記述されたページが含まないようなキーワードの組を基にクエリを作り、入力ボックスにそのクエリを入力することでその情報にたどり着こうとする。しかし、適切なクエリを生成することはユーザにとって容易なことではなく、クエリによっては検索結果数が膨大になり過ぎたり逆に絞り込み過ぎたりするため対象とする情報を含むページを得られないことが多い。この大きな要因として、既存の検索エンジンにはクエリ中のキーワードそれぞれに対する意図を伝えることができないという点が挙げられる。

例えば、「京都で豆腐などの食材を使った和食料理を食べに行きたい」と考え、「京都 豆腐 和食」というクエリで検索を行ったとする。このとき、ユーザは「豆腐などの食材」という検索意図を伝えるためにとりあえず「豆腐」という語をキーワードとして入力しているが、「湯葉」や「豆乳」などでもかまわないと考えている。しかし、検索エンジンにはそうしたユーザの意図がわからないため、ただ単に「京都」「豆腐」「和食」の3語を含むページを検索結果として返すだけである。ユーザが「湯葉」「豆乳」などを含むページが欲しいと思っている場合は、新たにクエリを作り検索する必要がある。

これに対して我々はこれまで、ユーザがこだわっていないキーワードを判別するための尺度として緩和度を定義し、ユーザから伝えられた緩和度とキーワード間の関係性を基に意図推定を行い、こだわっていないキーワードに対してその語を補う適切な語を取得し、それらの語を用いて元のクエリを置換することで新たにクエリ群を生成し、そのクエリ群によって得られる複数の検索結果を一つにまとめてユーザに提示する手法を提案してきた[1]。

しかし、複数の検索結果をまとめて提示することで、ページ

集合が大きくなってしまい、下位に埋もれる適合ページが増加しやすくなるという問題も生じる。これを解決するためには、得られたページの中でユーザの意図に沿うものをうまく上位に並び換える必要があるが、これまでの研究ではクエリ内のキーワードを多く含むページを上位に並び換えるという単純な手法を用いており、このページの並び替えにおいて十分にユーザの意図を考慮することができていなかった。

そこで本稿では、ユーザの意図を考える上で重要な「クエリの構造」に注目し、このクエリの構造に基づき検索結果中の各ページをユーザの意図に沿うように並び替える手法を提案する。

本稿の構成は以下の通りである。まずは2章で関連研究と比較することで本研究の位置付けについて述べる。3章で本研究における「クエリの構造」とはどのようなものかについて述べ、その推定手法についても説明する。4章でこれまでの検索結果取得手法と、本研究の「クエリの構造」を考慮したページの並び替え手法について述べ、5章でシステムの実装について述べる。最後に6章でまとめと今後の課題について述べる。

2. 関連研究

2.1 クエリ内のキーワード間の関係性

野田らは、クエリ内の2つのキーワード間において、それらが「の」で結ばれる関係にあるかを調べることで、主題を表すものと話題を表すものを判別する手法を提案している[2]。我々はこの手法を基にクエリ内の任意の2キーワード間に成り立つ関係性を取得し、得られた関係性を利用して意図推定を行う。

2.2 クエリ修正

我々の研究では、クエリ修正をする際に元のキーワードを補うための語を取得している。これは大島らが提案した手法[3]を基に、キーワードの関係性まで考慮することでより検索意図に沿った同位語を取得するというものであり、4.3節で改めて述べる。ある語に対してそれに関連する語を見つけ出す研究としては他にも、クエリログ[4]やHTML構造[5]や大規模辞書[6]

を利用したものなど多く存在するが、辞書が不要で柔軟に語を取得できることと、比較的高速に関連する語を取得できることから大島らの手法を利用している。

クエリ修正に関する研究は多く存在し、我々の研究と同様に元のキーワードを他の語に置き換えようとする研究も多い[7][8]。しかし、クエリの構造やキーワード間の関係性の考慮、また辞書に頼らない語の取得手法などの点からこれらの研究とは異なるものだと言える。

2.3 検索意図推定

ユーザの検索意図を推定しようとする研究は昔から盛んに行われてきた。Kang ら [9] や Lee ら [10] は検索目的を分類することで意図推定を行う手法を提案している。また、ユーザの行動履歴に基づき検索結果を変える適合フィードバックに関する研究 [11][12] も有名であるが、これは事前に反復的な操作を必要とする点で我々と異なる。他にも、田らは2キーワードから成るクエリにおいてキーワード間に成り立つ関係を考慮して検索結果内のページをランキングする手法を提案している [13]。この手法では元の検索結果のランキングを行うのに対して、我々は元の検索結果には含まれない検索結果も含めてユーザに提示する点で異なる。

3. クエリの構造

本稿で述べるクエリの構造とは、クエリ中の各キーワードが持つ役割やキーワード間において成り立つ関係性のことである。クエリの構造のイメージを図1に示す。例えば「京都 豆腐 和食」というクエリが投げられた時、ユーザの意図としては「京都で豆腐などを扱った和食を楽しみたい」といった場合や、「京都の豆腐を食べてみたい、和食で」といった場合がある。前者の場合、ユーザは「豆腐」と「和食」のつながりを重要と考えており、「京都」には場所を表す役割を期待している。一方で後者の場合、ユーザは「京都」と「豆腐」のつながりを重要と考えており、「京都」には「豆腐」を形容する役割を期待している。このように、ユーザが想定しているキーワード間のつながりや各キーワードに対して期待する役割は、同じキーワードからなるクエリであっても検索意図によって異なる。ユーザは検索意図を検索クエリに落とし込むとき、このようなクエリの構造を想定しているはずだが、既存のシステムではキーワードのみしか入力できないため、システムにこのクエリの構造をうまく伝えることができない。このクエリの構造の欠如により、既存の検索エンジンによって得られるのは検索クエリのキーワードに適合した検索結果になってしまっている。一方でユーザが求めているのは、キーワードに適合する検索結果ではなく検索意図そのものに適合する検索結果のはずである。そこで、本研究ではこのクエリの構造を推定し、検索結果に反映させることで、ユーザの検索意図に沿った検索結果を提示する手法を提案する。

我々はこれまでの研究で、キーワード間の関係性の強さを考慮してキーワードの適切な置換語を取得する手法を提案してきた。本研究では各キーワードに対してユーザがどのような役割を期待しているのかという部分に注目し、次節でキーワードに期待される役割の分類と推定手法について述べる。クエリの構

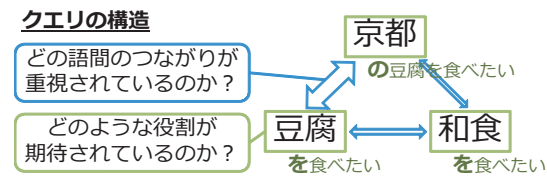


図1 クエリの構造

表1 格助詞一覧

種類	役割 (一部)
が	主体
の	修飾
を	動作の対象
に	動作の対象・目的地
へ	目的地
と	相手
から	起点・原因
より	起点・比較
で	場所・手段・媒介物

造に基づく検索結果の生成については本章の最後でそのアイディアを述べ、具体的な手法はこれまでの手法も交えて4章でまとめて述べる。

3.1 各キーワードの役割の分類

キーワードの役割としては、何かを形容するためのもの、場所を示すためのもの、目的を表すためのものといった分類も考えられるが、ここではその語がどのような各助詞を伴うかによって分類することにする。具体的には表1に示した9通りの格助詞によって分類する。このように分類する理由としては、ユーザの検索意図は文章のようなものであり、クエリとして用いた各キーワードは検索意図の段階では格助詞を伴っている場合が多いと考えたためである。例えば単に「京都」というクエリで検索をしたとしても、ユーザの検索意図としては「京都に行きたい」「京都の懐石料理を食べてみたい」「京都をもっと知りたい」といったように助詞を伴う文章として検索意図が頭の中にあることが多い。そのような場合、この検索における「京都」の役割を分類するとき、「京都の」「京都で」といった分類で十分に用途を分類できると考えたためである。

3.2 各キーワードの役割の推定

3.2.1 各キーワードと格助詞の共起頻度の集計

まずはクエリ中の各キーワードがそれぞれどのような格助詞を伴いやすいのかを取得する。より格助詞とのつながりを効果的に取得するため、ここでは対象のキーワードを、そのキーワードが格助詞を伴うすべてのパターンを検索オプションのORでつなげたものに置き換えたクエリの検索結果を基に行う。具体的には、元のクエリが「京都 豆腐 和食」で京都が伴いやすい格助詞を調べる場合、表1を基に「(京都が OR 京都の OR ... OR 京都で) 豆腐 和食」というクエリで検索結果を取得し、その上位100件中のタイトルやスニペット中で京都の後に現れる格助詞を集計する。ORで拡張したクエリから得られる検索結果の上位100件のみから格助詞を抽出する理由は、検索結果の上位にはクエリに適したページが現れやすく、そのようなページに現れる格助詞を抽出するだけでも十分と考えた

表 2 「京都 豆腐 和食」中の各キーワードに対する出現文書数

種類	京都	豆腐	和食
が	4	16	10
の	76	39	60
を	17	39	29
に	15	9	6
へ	4	0	0
と	2	31	15
から	3	0	5
より	1	0	0
で	26	6	5

ためである。クエリ「京都 豆腐 和食」中の各キーワードに対する各格助詞の出現文書数を図 2 に示す。

3.2.2 特徴的な格助詞の抽出

次に、出現数に基づきそのキーワードがどの格助詞を伴いやすいのかを測定する。ここで、格助詞そのものの出現頻度がそれぞれ異なるため、単純に出現数の多いものから順に伴いやすいと判断するべきではない。そこで、各格助詞の一般的な出現頻度をまず取得した。キーワードとして使われる名詞として、一般名詞や固有名詞を十分な数確保する必要がある。我々はこの名詞集合として Web 上で公開されている ipadic-2.6.3^(注1) という辞書に登録されている名詞を利用した。この ipadic-2.6.3 では 20 万語以上の名詞が人名、地域名などで分類されて登録されており、我々はこの中でクエリとして用いられやすいと思われる一般名詞、固有名詞、サ変型名詞、人物名、組織名、地域名に属する語を 100 語ずつ利用した。その各名詞に対して、それが格助詞を伴うすべてのパターンを OR でつなげて作られた「○○が OR ○○の OR ... OR ○○で」という形のクエリからえられる検索結果の上位 100 件から、その名詞の直後に格助詞の現れる文書数の平均を集計した。この集計結果を表 3 に示す。

この集計結果中の全体の平均値とクエリによって得られた値の各カイ二乗検定の値に符号をつけたものをそのキーワードのその格助詞の随伴度とする。随伴度が正の格助詞は出現文書数が表 3 の平均出現文書数よりも高いものであり、随伴度が負の格助詞はその逆である。抽出された例として表 4 にクエリ「京都 豆腐 和食」中の各キーワードに対する各格助詞の随伴度を示す。実際に高い値が得られた格助詞を見ていくと、「京都で (グルメを楽しむ)」「京都の (豆腐)」「豆腐を (堪能)」「豆腐と (湯葉)」「和食を (味わう)」「(括弧内はその後続く文章の例) といったように尤もらしいものが取得できていると言える。

本研究ではこの格助詞の中で随伴度の高いものをユーザが期待する役割とみなして検索結果中のページの並び替えに利用する。

3.2.3 各キーワードの名詞の種類の推定

ここで述べる各キーワードの名詞の種類の推定は本手法で直接利用はしないが、いくつかのクエリ中のキーワードで試した結果有意義な結果が得られたので以下に示す。

表 3 中の 6 種類の名詞からはそれぞれ異なる傾向が見られ

表 3 上位 100 ページ中で名詞の後に格助詞の現れる文書数の平均

種類	一般	固有	サ変型	人物名	組織名	地域名	全体
が	14.51	4.48	11.93	16.66	12.33	3.16	10.51
の	45.44	31.58	35.94	48.95	35.14	37.27	39.05
を	22.30	8.26	25.63	9.00	5.45	4.99	12.61
に	14.44	10.76	11.95	10.46	8.85	9.43	10.98
へ	1.45	4.23	2.40	1.46	1.86	2.00	2.23
と	12.90	5.49	11.71	11.94	8.34	4.35	9.12
から	2.40	4.3	1.99	2.96	3.41	2.61	2.95
より	0.52	0.69	0.22	0.80	1.11	0.38	0.62
で	10.33	7.05	10.22	4.73	8.12	6.30	7.46

表 4 「京都 豆腐 和食」中の各キーワードに対する随伴度

種類	京都	豆腐	和食
が	-4.51	3.20	-0.03
の	57.36	-0.1	18.44
を	1.75	63.20	24.38
に	1.65	-0.40	-2.54
へ	1.44	-2.28	-2.28
と	6.12	57.76	4.17
から	0.01	-3.04	1.47
より	0.23	-0.62	-0.62
で	49.79	-0.31	-0.88
推定された役割	の	を	を

表 5 「京都 豆腐 和食」中の各キーワードの推定された分類

種類	京都	豆腐	和食
一般	0.936	0.9038	0.987
固有	0.915	0.9036	0.930
サ変型	0.915	0.9036	0.930
人物名	0.925	0.774	0.940
組織名	0.957	0.775	0.929
地域名	0.981	0.703	0.913
推定された分類	地域名	一般	一般

た。例えば、地域名であれば他に比べ「が」や「と」と共起する文書数が少なく、逆に人物名であれば「が」や「と」と共起する文書数が多い。そのため、格助詞の共起文書数の傾向が 6 種類の名詞のどれと類似しているかを調べることで、そのキーワードがどのような種類の名詞かを推定することが可能と考えられる。そこで予備実験として、類似度の計算手法としてコサイン類似度を用いて「京都 豆腐 和食」中の各キーワードと 6 種類の名詞との類似度を求め、それを基に名詞の種類を分類した。推定された分類を表 5 に示す。京都と和食に関しては正しく推定されたが、豆腐に関してはサ変型名詞と誤って推定された。他の例として「ピエラ アクオス 消費電力」と「温暖化 二酸化炭素」の 2 クエリ中の各キーワードに対する推定結果も図 6 示す。これらのクエリに関しても半数以上のキーワードに対して正しい推定が行われていた。この名詞の種類の利用については今後の課題で述べたいと思う。

4. 提案手法

この章ではまず、これまでの研究で我々が提案してきた「緩和度」「キーワード間の関係性」「元のキーワードを置換するための語の取得」について簡単に説明する。次に、それらの要素と 3. 章で述べたクエリ中の各キーワードが持つ役割を基に新

(注1) : <http://chasen.aist-nara.ac.jp/chasen/distribution.html>

表 6 2 クエリ中の各キーワードの推定された分類

種類	ビエラ	アクオス	消費電力	温暖化	二酸化炭素
一般	0.833	0.944	0.695	0.976	0.954
固有	0.861	0.878	0.514	0.931	0.873
サ変型	0.798	0.924	0.775	0.965	0.951
人物名	0.800	0.941	0.537	0.953	0.928
組織名	0.821	0.936	0.520	0.947	0.911
地域名	0.925	0.935	0.416	0.891	0.848
推定された分類	固有	一般	サ変型	一般	一般

たなクエリを生成し、そのクエリの検索結果をマージする手法について述べる。

4.1 緩和度

緩和度とは各キーワードに対してユーザによって付加される 0 から 1 までの連続値であり、値が高いほどそのキーワードをユーザがあまりこだわっていないことを表す。ここでこだわっていない語とは、その語自身でなくても同じような意味合いの語であれば構わない語という意味である。例えば「豆腐など」といった検索意図を持っているとき、キーワードとして「豆腐」しか思い浮かばない場合もあるが、この場合豆腐のみにこだわっているわけではなく、似たような食材についての情報もまとめてほしいと考えている。しかし、キーワードのみではそのような意図を伝えることができない。そこで緩和度という概念を導入し、ユーザが各キーワードに対して緩和度を付加できるようにする。これにより、ユーザはシステムにそのような意図を伝えることができ、それにより必要に応じて似たような意味合いの情報まで含む検索結果を得ることが可能となる。

4.2 キーワード間の関係性

4.2.1 関係の種類

キーワード間には様々な関係が成り立つと考えられるが、本稿では 2 キーワード間で成り立つ以下の 2 つの関係のみを考慮することとする。

従属関係 キーワード A, B が「 A の B 」という形で表せるとき、それらには従属関係が成り立つとする。「京都 紅葉」、「オス スメ スポット」のように修飾関係にあるものや「iPod 価格」、「デジカメ 画素」のように主題と話題の関係にあるものが例として挙げられる。

並立関係 キーワード A, B が「 A や B 」、「 A と B 」という形で表せるとき、それらには並立関係が成り立つとする。「Wii Xbox360」、「パナソニック シャープ」といったように何かを比較する場合や「晩婚化 少子化」、「福利厚生 給料」など要素を列挙する場合に現れることが多い。

4.2.2 関係性の種類と強さの抽出

2 つのキーワード間の関係を取得するために、2 つのキーワード T_1, T_2 と接続助詞 W_c に対して、 T_1 と T_2 を W_c で結んだフレーズがどの程度妥当であるかを表す値を以下の $Strength(T_1, W_c, T_2)$ を用いて求める。

$$Strength(T_1, W_c, T_2) = \frac{DF("T_1 W_c T_2")}{DF("T_1 W_c")} * \frac{DF("T_1 W_c T_2'")}{DF("W_c T_2'")} \quad (1)$$

ここで $DF("X'")$ はクエリ " X " によるフレーズ検索の結果文

表 7 関係性の種類と助詞の対応

関係性の種類	助詞
従属関係	の
並立関係	と・や

書数、 $T_1 W_c T_2, T_1 W_c, W_c T_2$ はそれぞれその順序で語をつなぎ合わせたフレーズを表す。例えば、「京都」と「豆腐」が「の」で結ばれるような関係かを測る場合、評価式の右辺第一項は「京都の」を含むページ中で「京都の豆腐」がどの程度存在するかの割合を表し、右辺第二項が「の豆腐」を含むページ中で「京都の豆腐」を含むページがどの程度存在するかの割合を表す。そのためこれらの積で求められる値は「京都の豆腐」がどの程度一般的に利用されるフレーズであるかを表す。

2 つのキーワード A, B 間の関係を取得する際には W_c として、「の」、「と」、「や」を用いて、 A, B と B, A の 2 通りの順序に対してそれぞれスコアを求める。こうして得られた 6 つの値のうち、最も高い値の得られた助詞に対応する関係性をそのキーワード間の関係性とし、その時の値をその関係性の強さとする (対応は表 7 を参照)。これをキーワードの組み合わせの数だけ行い、各キーワードが関係をもつキーワードを取得する。もしあるキーワードと関係をもつキーワードが複数存在する場合は、その中で最もスコアの高い関係のキーワードのみを利用する。

4.3 元のキーワードを置換するための語の取得

元のキーワードを置換するための語の取得には、2. 節でも紹介した大島らの手法 [3] を用いる。この手法はある語 X を入力すると、 X の同位語が出力として得られる。さらに入力として、 X とは別に「背景語」というものを設定することができ、この背景語を設定することで X の同位語のうち背景語に関する語を優先的に取得することが可能となっている。そこで、4.2.2 節より、各キーワードが他のどのキーワードとの関係性が強いかが分かっていることを利用し、ユーザが緩和度を高く設定したあるキーワード X に対する同位語を取得する際に、背景語として X と関係性の強い語を利用することでよりそのクエリに沿っていると考えられる同位語を取得することができる。

4.4 検索結果の生成

本システムが最終的に返す検索結果はシステムが複数のクエリを生成し、それらによって得られる検索結果集合中のページを並び替えることで生成される。その際重要な点として、どのようなページを最終的な検索結果に含めるかという点と、何を基にページを並び替えていくかという点が挙げられる。この 2 点を踏まえてどのようにして最終的な検索結果を生成していくのかを表したイメージを図 2 に示す。

本手法では、図 2 に示すように置換クエリと格助詞クエリの 2 種類のクエリを用いる。置換クエリは 4.3 節で取得した同位語を基に、元のクエリ中のキーワードを置換することで生成する。格助詞クエリは 3.2.2 節で取得した最も随伴度の高い格助詞を元のクエリに付加させていくことで生成する。例えば、「京都 豆腐 和食」というクエリでユーザが豆腐にあまりこだわっていない場合、置換クエリは元のクエリも含めた「京都 豆腐

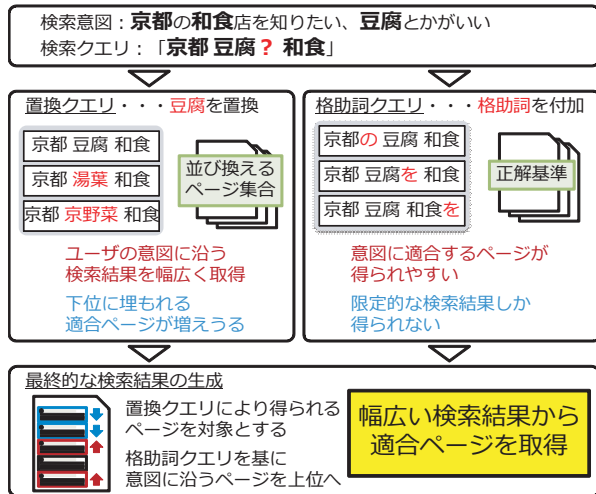


図2 各検索結果の利用法

和食」「京都 湯葉 和食」「京都 京野菜 和食」となり、格助詞クエリは「京都の 豆腐 和食」「京都 豆腐を 和食」「京都 豆腐 和食を」となる。ここで、格助詞クエリを生成する場合に1キーワードずつしか付加させていないのは、本研究ではキーワードごとに格助詞の随伴度を求めており、どの格助詞の組み合わせが最も妥当かまでは考慮できていないためである。ただし、これらの格助詞クエリは元のクエリに比べてキーワードの役割を限定しており、これらを集約することで十分にユーザの期待する役割を検索結果に反映できると考えられる。置換クエリは適切な置換語を用いることにより元のクエリよりも多くの話題を含む検索結果を取得することができる。しかしページ集合が大きくなるためうまくページを並び換えなければ下位に多くの適合ページが埋もれてしまう危険性がある。一方で格助詞クエリはキーワードの役割を限定しているため、ユーザの意図に沿ったページが得られやすいが、格助詞を伴うようなページしか得られないため元のクエリに比べて得られるページは限定されてしまう。

この2種類のクエリを効果的に利用するために、本手法では置換クエリにより得られるページ集合を最終的に提示するページ集合とし、このページ集合を格助詞クエリにより得られるページ集合を基に並び換える。このような手法をとるのは、格助詞クエリにより得られるページ集合は適合性が高いと考えられ、このページ集合と類似するようなページをより大きなページ集合から見つけて上位に並び換えることで、多くの適合ページを上位に含む検索結果を生成できると考えたためである。

4.4.1 語の重みづけ

ここでは格助詞クエリに基づく語の重みづけについて述べる。はじめに、各格助詞クエリの検索結果を取得し、その検索結果中の上位100ページのタイトルとスニペットに対して、そのクエリ中の格助詞付きのキーワードを含む文章を抽出する。例えば、「京都の 豆腐 和食」というクエリに対して、「京都の」を含む文章を抽出する。そして、それらの文章中に現れる語の出現頻度を取得し、その対数をその語の重みとする。

4.4.2 ページの並び替え

すでに述べたように、本手法では置換クエリによる検索結果中のページを並び替えてユーザに提示する。まずは各ページのタイトルとスニペット中に含まれる語の重みの総和を求め、その値の大きいものから順に順位を付けていき、各ページに重みによる順位 $Rank_{Weight}$ を設定する。また、各ページの元の検索結果における順位を $Rank_{Original}$ とする。ここで、 $Rank_{Original}$ は元のクエリとどの程度適合するかの尺度を、 $Rank_{Weight}$ はユーザの意図とどの程度適合するかの尺度を表しているとみなし、両方の順位を考慮してページの並び換えを行うことにする。具体的には各ページ p に対して以下のようにしてページのスコア $Score(p)$ を算出する。

$$Score(p) = \frac{1}{Rank_{Origin}(p)} + \frac{1}{Rank_{Weight}(p)} \quad (2)$$

最後に、このスコアに基づきページを並び替えたものを最終的な検索結果としてユーザに提示する。

5. 実装

これまで提案してきた手法を基にC#でシステムのプロトタイプを実装した。検索結果はYahoo!Japanの提供しているAPI^(注2)によって得られるものを利用している。

ユーザは入力ボックスにキーワードを入力することで検索を行うことができる。もしキーワードの中にこだわっていないものがあれば、ユーザはそのキーワードに「?」を付加することでその意図をシステムに伝えることができる。システムはこの「?」の数だけ置換クエリを生成し、そのクエリによる検索結果中のページを並び替えてユーザに提示する。

図3は「京都 豆腐 和食」というクエリと豆腐にこだわってはいないという意図を入力として与えている例である。この入力に対してシステムはまず「豆腐」に対して「湯葉」と「京野菜」という置換語を取得し、それを基に置換クエリを生成しページ集合を取得する。同時に各キーワードに期待される役割を推定し、その推定された意図に基づきページ集合を並び替えてユーザに提示するようになっている。

6. まとめと今後の課題

6.1 まとめ

本稿では、検索意図を推定する上で重要と考えられるクエリの構造の概念について述べた。また、クエリの構造の一つとして各キーワードの役割に注目し、その推定手法を提案した。さらに、推定された各キーワードの役割を基に、検索結果中のページを意図に応じて並び替える手法を提案した。

6.2 今後の課題

今後の課題としてまず、この手法の評価が挙げられる。3.2.2節の伴いやすい格助詞の抽出精度、3.2.3節のキーワードの種類の推定精度、4.4節の最終的な検索結果中のページがユーザの意図に沿って並び替えられているかなど、評価すべき項目が多く残ってしまっている。また、格助詞クエリは実際に高い適

(注2) : <http://developer.yahoo.co.jp/webapi/search/>

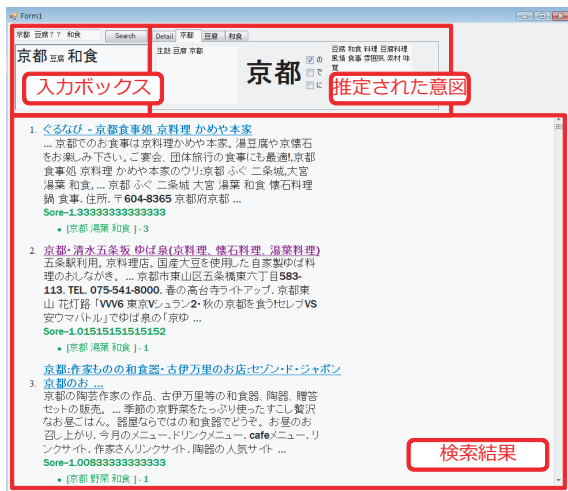


図 3 クエリの構造

合性を持っているのかどうかなど、本稿で仮定していたことは実際に成り立つのかを確かめる必要もある。

2つ目として、クエリの構造を考える上でキーワード間のつながりを考慮していかなければならないと考えている。例えば、本研究ではキーワードごとに伴いやすい格助詞を求めたが、これは格助詞の組み合わせとして推定するのが望ましい。例えば、「京都の」「豆腐を」「和食を」とそれぞれで最もらしい格助詞が推定されたとしても、「豆腐を」と「和食を」が共起しやすいというわけではない。「京都の豆腐を和食で(楽しみたい)」といったようにキーワードのつながりも考慮して格助詞の組み合わせという形で推定することで、より検索意図に近づくことができると考えられる。キーワード間のつながりは本研究において非常に重要な部分なので今後これをうまく考慮できる手法を考えていくつもりである。

3つ目として、本稿では利用しなかったキーワードの名詞の種類推定結果を今後利用していきたいと考えている。例えば、クエリ内のある2キーワード間に4.2.1節で述べた従属関係が成り立っているとする。このとき、その2キーワードが「京都観光」の場合はどこで何をしたいか示しているのに対し、「京都嵐山」の場合は場所のみを示していることになる。これまではこの2つを区別できなかったが、名詞の種類を考慮することで「(地域名)+(サ変型)」と「(地域名)+(地域名)」というように区別することができ、それぞれに対して異なる処理を行うことも可能となる。意図推定においてこの名詞の種類の利用は大きな可能性を持っていると思われるので、今後これの具体的な利用手法を考えていきたいと考えている。

謝辞 本研究の一部は、京都大学グローバル COE プログラム「知識循環社会のための情報学教育研究拠点」、文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」、計画研究「情報爆発に対応するコンテンツ融合と操作環境融合に関する研究」(研究代表者：田中克己、課題番号 1809041)、文部科学省研究委託事業「知的資産の電子的な保存・活用を支援するソフトウェア技術基盤の構築」、異メディア・アーカイブの横断的検索・統合ソフトウェア開発

(研究代表者：田中克己)、および、NICT 委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」によるものです。ここに記して謝意を表すものとします。

文 献

- [1] 金子恭史, 中村聡史, 大島裕明, 田中克己: “緩和度付き検索語の意味関連分析による検索意図推定とそのクエリ入力インタフェース”, Journal of the DBSJ, **7**, 1.
- [2] 野田武史, 大島裕明, 小山聡, 田島敬史, 田中克己: “主題語からの話題語自動抽出とこれに基づく Web 情報検索”, DBSJ Letters, **5**, 2.
- [3] H. Ohshima, S. Oyama and K. Tanaka: “Searching Coordinate Terms with Their Context from the Web”, LECTURE NOTES IN COMPUTER SCIENCE, **4255**, p. 40 (2006).
- [4] 山口雅史, 大島裕明, 小山聡, 田中克己: “サーチエンジンのクエリログを利用した同位語の発見”, DBSJ Letters, **5**, 2.
- [5] K. Shinzato and K. Torisawa: “A simple WWW-based method for semantic word class acquisition”, AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4, **292**, p. 207 (2007).
- [6] Z. Ghahramani and K. Heller: “Bayesian Sets”, ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, **18**, p. 435 (2006).
- [7] R. Jones, B. Rey, O. Madani and W. Greiner: “Generating query substitutions”, Proceedings of the 15th international conference on World Wide WebACM New York, NY, USA, pp. 387–396 (2006).
- [8] R. Kraft and J. Zien: “Mining anchor text for query refinement”, Proceedings of the 13th international conference on World Wide WebACM New York, NY, USA, pp. 666–674 (2004).
- [9] I. Kang and G. Kim: “Query type classification for web document retrieval”, Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrievalACM New York, NY, USA, pp. 64–71 (2003).
- [10] U. Lee, Z. Liu and J. Cho: “Automatic identification of user goals in Web search”, Proceedings of the 14th international conference on World Wide WebACM New York, NY, USA, pp. 391–400 (2005).
- [11] J. Rocchio, et al.: “Relevance feedback in information retrieval”, The SMART Retrieval System: Experiments in Automatic Document Processing, pp. 313–323 (1971).
- [12] B. Tan, A. Velivelli, H. Fang and C. Zhai: “Term feedback for information retrieval with language models”, Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrievalACM Press New York, NY, USA, pp. 263–270 (2007).
- [13] 田馳, 手塚太郎, 小山聡, 田島敬史, 田中克己: “質問キーワードの近接性と密度分布に基づくウェブ検索の改善手法”, DBSJ Letters, **5**, 1.