Vol.2009-DBS-149 No.1
2009/11/20

緩和検索における各ページの話題の共起性に基づくランキング手法の提案

金子恭史[†1]　中村聡史[†1]　田中克己[†1]

Web 検索においてユーザは自分の興味の範囲をうまくキーワードに表せない場合がある．これに対して我々は，ユーザに興味のある例を一つキーワードで示してもらい，それに「緩和度」という値を付加してもらうことで，他にユーザが興味のある語を推定し，それらの語に関する話題も含む検索結果を提供するシステムを提案してきた．このシステムでは，推定された語を基に生成された複数のクエリから得られる検索結果群を統合してユーザに提示しており，その中にはあるクエリからのみ得られる独立したページや，複数のクエリから得られる話題の網羅性の高いページが存在する．本稿ではこれらの差異に基づく 2 つのランキング手法とそれらの有用性を確かめるための評価について述べる．

Ranking Methods based on Co-occurrence of Topics in Each Page in Search Relaxation

Yasufumi KANEKO,[†1] Satoshi NAKAMURA[†1]
and Katsumi TANAKA[†1]

In Web search, users are sometimes unable to express the range of their interests as keywords. For this problem, we proposed the system which enable users to get search results which contains several topics they are interested in as long as they input one of those they interested in and append *Relaxation-value* to the term. There are several queries our system generated in our system. Then, there are some pages we can get by only one of the queries and those we can get by several of the queries in our search results. In this paper, we propose two ranking methods based on the difference and describe the evaluation to show the usefulness of them.

†1 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
Department of Social Informatics, Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto, 606-8501 Japan

1. はじめに

近年，場所や商品やお店など様々な情報を調べる際にキーワード入力型の Web 検索システムが利用されるようになってきている．しかし，キーワードマッチに基づく検索エンジンの多くでは，以下のような問題が存在する．

(1) 入力ミスがあると正しく検索されない（例：打ち間違い，スペルミス）
(2) 表現のばらつきがあると再現率が下がる（例：正式名称と略称）
(3) 各キーワードはそれぞれ何かしらの意図を持って入力されているが，その意図まで検索エンジンに伝えることができない

大手検索エンジンは入力ミスと思われる語に対して正解と思われる語を提示する．また，例えばユーザがある人物を調べる際に，その人物の愛称をキーワードとして入力すると，同時にその人物の本名にもヒットするような検索結果が得られる．このように，(1) と (2) に関してはすでにある程度サポートされている．しかし，(3) に関してはまだ十分なサポートがなされていない．

例えば，あるユーザが京都で豆腐などの食材を扱った和食料理を食べに行きたいと思い，「京都 豆腐 和食」というクエリで検索したとする．ここで，このクエリを分解してキーワードごとの意図を考慮すると，「京都」「豆腐」「和食」という各キーワードについてそれぞれ「京都という場所で」「豆腐などの京都らしい食材」「和食のお店を探したい」という意図を持っていることになるが，キーワード群のみでは情報が少なく，他の意図を持って作られたクエリと重複することも多いため，検索システムがキーワード群のみから正確にユーザの意図を推定するのは難しい．ユーザとしても，自分の持つ意図をうまく検索結果に反映できないのは望ましくはないと考えられる．検索システムにはキーワードとして名詞が利用されることが多いが，最近では「○○とは」「××で」「△△といえば」といったように名詞と定型文を組み合わせて意図する情報を得ようとする検索テクニックも利用されるようになってきており，一見ユーザの意図をうまく伝えられているようにも見える．しかし，これは「○○とは」といったフレーズにマッチするページを見つけてきているに過ぎない．「○○とは」という検索を例にすると，本当にユーザがほしいのは「○○とは」というフレーズを含んでいるかによらず「○○」について説明されているページのはずである．そのためこのような検索テクニックで十分にユーザの意図を検索結果に反映できているとは言えない．

Vol.2009-DBS-149 No.1
2009/11/20

このようなキーワードマッチにおける問題を解決するには，ユーザがより直接的に検索結果に意図を反映できるような検索環境の実現が必要である．その第一歩として我々は，ユーザが各キーワードへの興味の範囲を検索結果に反映できるようなシステムを提案してきた．本稿ではこの提案システムにおける 2 種類の検索結果ランキング手法を提案する．

本稿の構成は以下のとおりである．まず 2 章でこれまでの研究について述べ，3 章で関連研究との比較により本研究の位置づけについて述べる．そして 4 章で 2 種類のランキング手法を提案し，5 章でそれらの評価について述べる．6 章でシステムの実装について述べ，最後に 7 章で本稿をまとめる．

2. これまでの研究

我々はこれまで，ユーザが各キーワードへの興味の範囲を検索結果に反映できるシステムを提案してきた[1),2)]．キーワードごとの興味の範囲を取り扱った理由は，我々は日常で他の人に自分の興味のあるものを伝える際に，うまく興味のあるものを包含する言葉を思いつけない場合は「○○などに興味がある」というように例示によって自身の興味の範囲を示すことが多いためである．しかし，一般的な検索エンジンでは調べたいものをはっきりとしたキーワードで表す必要があるため，例示による検索には向いていない．例えば，豆腐に興味があれば「豆腐」で十分な検索結果を得ることができるが，豆腐や湯葉などのように京都のいくつかの食材に興味があるとき,「豆腐」では興味の範囲をうまく伝えきれていない．このときの解決策として,「京都 食材」といったようにより抽象的な語で置き換えることも考えられるが，こうした抽象的な語は興味のない情報も多く含んでしまう可能性が高く，十分とは言えない．他の方法として,「京都 湯葉」「京都 京野菜」といったように具体的なキーワードで順に検索していくことも考えられるが，この場合ユーザが知っているものしか検索できない上に，単語の数だけ再検索を要するために複数の検索結果を閲覧していかなければならないため，ユーザの負担が増えてしまう．一方,「京都 (豆腐 OR 湯葉 OR 京野菜)」のように OR 検索を利用することで一つの検索結果として表示することもできるが，入力の手間は増えてしまう．特に，White らの研究[3)] によると検索クエリとして「-」「site:」などを利用した複雑なクエリを用いるユーザは 8.7% にすぎないことがわかっており，ユーザがこうしたクエリを入力してくれることは期待できない．つまり，多くの検索エンジンでは “○○について調べたい” という検索行動は満たせても “○○などについて調べたい” という検索行動を十分に満たすことができない．

そこで我々は，ユーザがキーワード入力時に各キーワードに対する興味の範囲を手軽に伝えることができるようにし，類似する語にも興味があるという意図が伝えられたキーワードについては自動的にそのキーワードの同位語を複数取得し，その同位語で置換されたクエリ群から得られる検索結果集合を一つの検索結果にまとめてユーザに提示するという検索システムを提案してきた．これにより，ユーザが興味のあるキーワードを一つ例示するだけで自身の興味のあるものを検索することを可能とした．本稿では，このようにキーワードに対して同位語を取得・利用することを「キーワードを緩和する」，キーワードの緩和によって新たな検索結果を得ることを「検索を緩和する」，こうした検索を「緩和検索」とそれぞれ表現することとする．

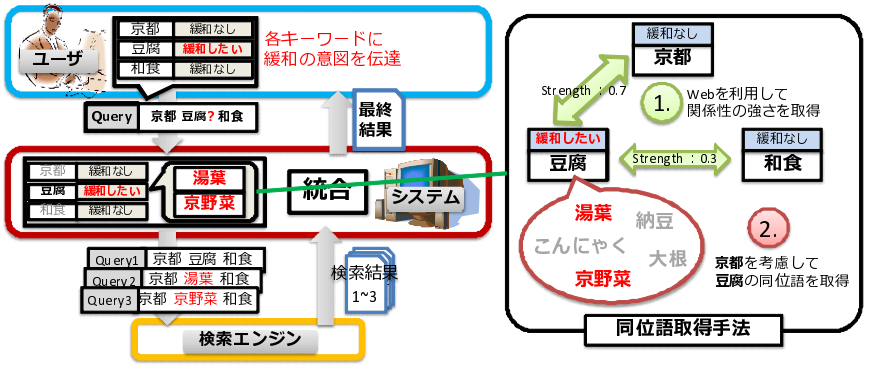

図 1 「京都 豆腐? 和食」というクエリに対する検索結果取得の流れ（左：全体の流れ，右：同位語取得の流れ）

この章ではまず，我々の提案する検索環境において重要な「緩和度」の概念とその伝達手段について述べる．そして，緩和度の付加されたキーワードが与えられた際に「キーワード間の関係性」を取得・利用して複数の検索結果を取得する手法について述べる．全体の流れを図 1 に示す．

2.1 緩 和 度

我々の提案する緩和検索の環境を実現する上で，ユーザの各キーワードへの興味の範囲を区別する必要がある．それを区別するための値として我々はキーワードごとに「緩和度」という値を設定する．緩和度は 0 から 1 の連続値で，この値が高いキーワードほどそのキーワードを多くの同位語で緩和するものとする．

我々が緩和度という概念を導入するのはユーザの検索環境をより快適にするためであり，そのためにはこの緩和度をユーザが手軽に各キーワードに付加できるような仕組みが求め

られる．緩和度をキーワードに設定する方法としてスライドバーや複数の入力ボックスを利用することも考えられるが，キーボードとマウスを持ちかえる手間が増えるという問題があるため，キーボードのみで入力できるインタラクション手法を提案してきた．提案した手法は，入力速度によって緩和度を伝える方法や,「豆腐?」「豆腐??」といったようにキーワードの後ろの “?” の数で緩和度を伝える方法，さらにその応用として「豆腐など」「豆腐かな」といったようにキーワードの後ろに特定の助詞を付加させることでその助詞に応じた緩和された検索結果を返す方法などである．本稿では，簡単化のため緩和度は “?” で表現することとする．例えば「京都 豆腐? 和食」というクエリは豆腐について緩和することを意味する．

2.2 元のキーワードを置換するための語の取得手法

例えば,「豆腐などの食材に興味がある」という場合は湯葉や京野菜にも,「FinePix などのデジカメに興味がある」という場合は LUMIX や IXY にも興味があると推測できる．このようにユーザが自身の興味のあるものを例示した場合，他に興味のあるものはその同位語であることが多い．そこで，我々のシステムではキーワードを緩和する際をそのキーワードの同位語を自動的に取得して利用する．その際にそのキーワードだけではなくそのキーワードと関係性のあるキーワードも考慮することで取得する同位語の精度の向上を試みている．ここではまず本稿で扱う 2 種類の関係性と，その関係性の強さの取得手法について述べる．そして，その関係性の強さを利用した同位語取得手法について述べる．

2.2.1 キーワード間の関係性の種類

本稿では 2 語間で成り立つ以下の関係性を考慮する．

従属関係 「京都 紅葉」,「iPod 価格」のように，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．A という主題に対して興味の対象を限定するための話題として B が用いられることが多いと考えられる．

並立関係 「XBox360 PS3」,「晩婚化 少子化」のように，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．A と B を比較する場合や，A や B などに興味がある場合にクエリに含まれることが多いと考えられる．

2.2.2 キーワード間の関係性の強さの取得手法

2 つのキーワード間の関係性の強さを取得するために，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．

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

ここで $DF(\text{“}X''\text{)}$ はクエリ “X'' によるフレーズ検索の結果文書数，$T_1 W_c T_2$，$T_1 W_c$，$W_c T_2$ 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」（T_1）と「豆腐」（T_2）が「の」（X）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している．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している．

2 つのキーワード A，B 間の関係性の強さを取得する際には 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，最も高い値をそのキーワード間の関係性の強さとする．これをキーワードの組み合わせの数だけ行い，各キーワードに対してそれぞれ最も強い関係性をもつキーワードを取得する．

2.2.3 キーワード間の関係性を考慮した同位語の取得手法

元のキーワードを置換するための語の取得には，大島らの手法[4] 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，2.2.2 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．例えば豆腐の同位語を取得する場合，背景語を用いない場合は納豆，こんにゃく，大根といった一般的な同位語が得られるのに対して,「京都」を背景語とすることで湯葉や京野菜といった元のクエリに沿った同位語を優先して取得することができるようになる．本稿では，あるキーワードを緩和したとき，そのキーワード自身と取得された同位語をまとめて「緩和語」と表現することとする．

2.3 クエリ群の生成・検索結果集合の取得

得られた同位語で元のクエリ中のキーワードを網羅的に置換していき，その組み合わせの数だけクエリを生成する．そしてそれらのクエリから得られる検索結果を結合し，各ページにスコアリングを行い，そのスコアに応じてランキングの行われた検索結果をユーザに提示

する．このランキングの際のスコアリング手法については 4 章で述べる．

3. 関 連 研 究

3.1 クエリ修正

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語の置換によってクエリ修正を行う研究も多い[5)6)]．我々の研究ではクエリ修正をする際に大島らが提案した手法[4)] を用いて Web 上から同位語を取得して利用している．このとき，野田らが提案している手法[7)] を利用して各キーワード間の関係性の強さを測り，緩和するキーワードと関連のある他のキーワードを求めて利用することで精度の向上を図っている．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[8)] や HTML 構造[9)] や大規模辞書[10)] を利用したものなど多く存在するが，Web を利用することで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．

3.2 検 索 意 図

ユーザの検索意図を扱う研究は昔から盛んに行われてきた[11),12)]．Kang らは検索クエリを分類し，その分類に応じた検索アルゴリズムの提案を行っている[13)]．しかし複数の解釈ができるようなクエリをキーワードのみを利用して分類することには限界がある．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[14)15)] 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[16)]．この手法では 1 つの検索結果のランキングを行うのに対して，我々は複数の検索結果を統合し，その中に含まれる各ページに対して検索意図を考慮したランキング手法を提案している点でことなる．

3.3 複数検索結果の統合

複数の検索結果を統合して利用する検索システムとしてメタサーチエンジンが有名である．メタサーチエンジンはあるクエリに対する複数の検索エンジンの検索結果を統合してユーザに提示するもので，多くの研究が存在する[17)18)19)]．我々の研究は複数の異なるクエリから得られる検索結果の集合を結合してユーザに提示するものであり，同じクエリから得られる検索結果の集合を結合するメタサーチエンジンとはこの点で大きく異なる．

4. ランキング手法

ユーザが単純に検索範囲の拡張を目的として緩和検索を行った場合，ある検索結果のページばかりが偏って出現してしまっては緩和検索を行う意味がないため，各検索結果中のページが均等に出現する方が好ましいと思われる．一方で，ユーザが緩和語同士の関係がどのようなものかまで理解・比較したいという目的を持って緩和検索を行った場合，単一の緩和語を扱ったページよりも複数の緩和語をまとめて取り扱ったページの方が好ましいと思われる．このようにユーザの検索目的に応じて望まれるランキング手法は異なる．本章では，これらの目的に応じて提案した 2 種類のランキング手法について述べる．

4.1 単純な検索範囲の拡張を目的としたランキング手法

ユーザがいずれかの緩和語に関するページを分け隔てなく眺めていきたいという意図で緩和検索を行った場合，ある緩和語に関するページばかりが偏って並べられていては緩和を行う意味がない．そこで，各検索結果中のページが均等に出現するように並べていくことを考える．ここでは元の検索結果における順位は信頼できるものと仮定し，各検索結果中のページを元の検索結果における順位の高いものから順に並べていくこととする．

また，例えばある検索結果の 1 位のページと他の検索結果の 1 位のページがあった場合，どちらを優先するべきかという尺度も必要になる．そこで，元のクエリはユーザが直接入力したものであるためユーザの興味と最も適合すると仮定し，元のクエリから得られる検索結果と類似した検索結果の得られるクエリほどユーザの興味に近いクエリであると見なすこととする．本稿ではあるクエリ q から得られる検索結果を R_q，元のクエリから得られる検索結果を R_{origin}，その検索結果中の上位 100 件のタイトルとスニペット中に出現する語の文書頻度ベクトルを $V_{DF}(R_q)$，2 つのベクトル v_1，v_2 のコサイン類似度を $Cosine(v_1, v_2)$ とし，以下のようにあるクエリ q と元のクエリ q_{origin} の類似度 $QS(q)$ を定義する．

$$QS(q) = Cosine(V_{DF}(R_q), V_{DF}(R_{q_{origin}})) \tag{2}$$

なお，あるページに対して，そのページが得られるクエリ（以後そのページのヒットクエリと呼ぶ）が複数あった場合は，その中で最もスコアの高くなるヒットクエリのみを考慮することとする．

以上の点を満たすように，ページ p のヒットクエリの集合を Q'，あるクエリ q' におけるページ p の順位を $Rank_{q'}(p)$ として，以下のようにして $Score_{base}(p)$ を定義し，これを基にランキングを行う．

$$Score_{base}(p) = \max_{q' \in Q'} \{ \frac{1}{Rank_{q'}(p) + (1 - QS(q')} \} \tag{3}$$

4.2 理解や比較を目的とした検索におけるランキング手法

あるキーワードを緩和して，そのいずれかに関するページがほしい場合は前節で述べた元

Vol.2009-DBS-149 No.1
2009/11/20

の検索結果での順位に基づくランキングで十分だと思われる．しかし，緩和検索では緩和によって得られた同位語をまとめて取り扱ったようなページ（以後，複数話題ページと表現する）の方が望まれる場合も多いと考えられる．本稿でいう複数話題ページには例えば以下のようなページが含まれる．

- ポータルサイトなど特定の情報をまとめて取り扱ったページ
- 複数の商品の情報を比較しているようなページ
- ブログでの旅行記のように関連のある物事を列挙しているようなページ

一方で，元の検索結果における順位を考慮した場合，各検索結果の上位のページのみが上位に現れることになる．各検索結果の上位にはそのクエリに特化したページが現れやすく，他のクエリとも関連する複数話題ページが現れるとは限らない．例えば,「デジカメ FinePix?」というクエリで検索すると，システムは「デジカメ FinePix」の他に「デジカメ LUMIX」などのクエリも生成するが，こういった商品を指すクエリの場合，上位には各メーカーの公式サイトが並んでいる場合が多く，そのサイトに他社の商品の情報は載っていない場合が多い．理解を目的とした場合，商品の違いなどまで一通り把握したいという意図があると思われる．このような場合ユーザは，単独の商品だけを扱ったページよりも，複数の商品を取り扱った複数話題ページを閲覧したいと考えるだろう．

こうした複数話題ページを得るための手法として，我々はページのヒットクエリ数に注目した．我々のシステムは同位語を利用した置換によってクエリ群を生成するため，その各クエリは「デジカメ FinePix」「デジカメ LUMIX」といったように異なる話題を扱うことになる．そのため例えば FinePix だけを取り扱ったページは「デジカメ FinePix」のクエリでしか得られないが，FinePix や LUMIX をまとめて扱っているページはどちらのクエリからも得られる．そこで，このクエリの共起が利用できると考えられるため，ヒットクエリ数の高いページから順に並べ替えることで複数話題ページを上位に並べることを試みた．実際にどの程度の効果が見られたかについては次節の予備実験で述べる．

4.2.1 予 備 実 験

ヒットクエリ数の高いページを上位に並び換えることで実際に複数話題ページが上位に出現するのかを確かめるために予備実験を行った．まず,「京都 豆腐？ 和食」といったように緩和するキーワードを一つ指定したクエリとそのクエリの検索目的の組を 20 通り準備した（表 1）．その各クエリに対して，緩和の際に利用する同位語の数を変えていった場合の上位 1 件，3 件，5 件の複数話題ページの出現率を測定した（図 2）．なお，図 2 の「全語」とは自動取得された同位語をすべて利用した場合の結果を表している．この実験結果から，緩和

表 1 評価に用いた検索目的とクエリ

	検索目的	クエリ
1	京都に京都っぽい食材を扱った和食料理を食べに行きたい．	京都 豆腐? 和食
2	ナシゴレン等のインドネシア料理のレシピがほしい．	インドネシア ナシゴレン? レシピ
3	リンゴとかを使ったダイエットがあるらしいので知りたい．	ダイエット リンゴ?
4	北海道で白い恋人とかお土産を買わなくてはならない．	北海道　白い恋人?
5	桜井翔など嵐のメンバーの出演しているドラマなどが見たい．	嵐 桜井翔? ドラマ
6	レンタカーが必要になったのでトヨタなどで調べてみる．	レンタカー トヨタ?
7	各社の携帯音楽プレーヤーについて調べたい．	携帯音楽プレーヤー アップル?
8	松井秀喜などメジャーリーグの日本人選手について知りたい．	メジャーリーグ 松井秀喜?
9	LUMIX くらいしか知らないけど，デジカメについて調べたい．	デジカメ LUMIX?
10	インフルエンザに効く薬ってタミフル以外にはないのだろうか．	インフルエンザ タミフル?
11	液晶テレビを調べてみよう．VIERA とかあったような …．	液晶テレビ VIERA?
12	ノート PC の VAIO とかについて調べてみよう．	ノート PC VAIO?
13	京都に観光に行くので金閣寺などを調べておきたい．	京都 観光 金閣寺?
14	イタリアのベネチアなど有名な都市に行ってみたい．	イタリア ベネチア?
15	サイパンあたりに観光に行ってみたい．	サイパン? 観光
16	フランスのエッフェル塔などの名所を知りたい．	フランス エッフェル塔?
17	忘年会の出し物を調べたい．例えば手品とか．	忘年会　手品?
18	世界中の名言やことわざなどに興味がある．	世界 名言?
19	PS3 の評判とか動向とかを知りたい．	PS3 評判?
20	トランプのいろいろなルールを知りたい．	トランプ ババ抜き?

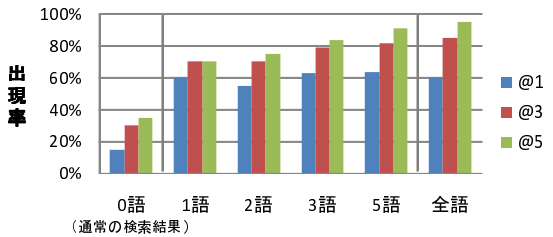

図 2 複数話題ページの出現率 (同位語利用数別)

されていない通常の検索結果（同位語の利用数が 0 の場合）に比べて，同位語を 1 つ以上利用して緩和された検索結果は上位に複数話題ページを高い確率で含むことが確認できる．

表 2 実験に用いた検索意図

	検索意図
1	検索範囲の単純な拡張（緩和した語のいずれかを話題として含むページに辿りつければいい）
2	理解・比較（緩和した語とその同位語との関係等まで理解したい，比較したい）

表 3 実験結果：各検索意図ごとの選択された割合

	M_{rank}	M_{cov}
検索意図 1	95.0%	5.0%
検索意図 2	7.5%	92.5%

5. ユーザ実験

5.1 実験内容と結果

実際にユーザが意図に応じて異なるランキング手法を好むのかを確かめるためにユーザ実験を行った．被験者は普段検索エンジンをよく利用している 20 代の学生 5 名である．ここで，4.1 の元の順位に基づくランキング手法を M_{rank}，4.2 のヒットクエリ数に基づくランキング手法を M_{cov} と表すこととする．評価に用いたのは表 1 に示す 20 個のクエリで，被験者には各クエリに対する M_{rank} と M_{cov} の 2 つの検索結果の上位 5 件を横に並べて同時に提示し，表 2 に示す各検索意図に対して，それぞれの場合にどちらの検索結果が好ましいと感じるかを選んでもらった．このときの検索結果は予め適切な同位語によって緩和された検索結果であり，緩和の際に利用された同位語や生成されたクエリ群は一目でわかるようになっているが，2 つの検索結果はそれぞれどちらのランキング手法によるものなのかは伏せて左右ランダムに配置して提示した．その結果，表 3 に示すような結果が得られた．

この結果から，実際にユーザは検索意図 1 の場合は M_{rank} を，検索意図 2 の場合は M_{cov} の手法を非常に高い確率で好むことが確認できた．今回提案した 2 つの手法がそれぞれの意図に特化した大きく性質の異なるものであったことから，より詳細な分析をするには比較手法を増やしていく必要はある．しかし，相対的ではあれ大きな差が現れたことから，検索意図に応じてユーザは異なるランキングを好むことを示しており，我々が提案した 2 つのランキング手法がそれぞれ想定した意図に対して有効であったことを示していると言える．

5.2 考察と今後の方針

5.2.1 単純な検索範囲の拡張を目的とした手法

緩和したキーワードに関するいずれかの話題を含むページが見つかれば十分な場合であれば，元の順位に基づき均等に各検索結果のページを並べていくだけでも効果のあることが実験により確認できた．

しかし，実際に生成された検索結果を眺めてみると前後のページとの統一性があまり見られないという問題が生じていた．この問題に関しては検索結果単位ではなくページ単位で類似度を計算するようにし，同時に類似度の影響が大きくなるようにスコアリングの定義を修正することで改善を図る予定である．また，本稿では元のクエリとの検索結果に類似しているものほどよいという仮定のもとランキングを行った．しかし，例えば 2 位のページが閲覧されるのが 1 位のページに満足できなかったためだった場合はあまり類似していないページの方が望ましいかもしれない．そのため，ページの類似度についてもユーザの検索目的を細分化して考えていく必要がある．

5.2.2 理解や比較を目的とした手法

この手法は予備実験でも十分に複数話題サイトを上位に並べることができており，ユーザ評価でも理解や比較を目的とする場合に高い確率で選ばれたことから有用なものだと言える．ただし，今回はクエリ中のキーワードの一つだけを緩和した場合の実験しか行っていないため，複数のキーワードを緩和した場合にも同様の結果が得られるかを実験する必要がある．また，例えばあるページのヒットクエリが「湯葉」「生湯葉」といった非常に近い意味合いの語からなるクエリの組だった場合と,「FinePix」「LUMIX」といった特定の場合しか共起しないような語からなるクエリの組だった場合で，同じ扱いをするべきではないと思われる．この問題に対しては，生成された各クエリ間におけるクエリの共起の際の重要度をうまく定義していく必要がある．

6. 実　　装

今回提案したランキング手法を踏まえて，Web アプリケーションとして Ruby によってシステムを実装した[⋆1]．各クエリの検索結果の取得には Yahoo! JAPAN の提供するウェブ検索用 API[⋆2]を利用している．図 3 は「京都 豆腐?? 和食」というクエリで検索した場合の実行例である．

6.1 モジュール

入力フォーム このシステムでは，キーワードに「?」を付加することでそのキーワードに緩和度を設定することができ,「?」の数を増やすほど多くの話題を含む検索結果が得ら

⋆1 http://www.dl.kuis.kyoto-u.ac.jp/kaneko/
⋆2 http://developer.yahoo.co.jp/webapi/search/websearch/v1/websearch.html

れるようになっている．また，検索オプションの部分からランキング手法を変更することができる（初期設定では M_{rank}）．

ページ集合　各ページはタイトル，スニペット，URL，ランキングにおけるスコア，そのページが得られたクエリの一覧によって構成されており，どのようなクエリから得られたページなのか確認できるようになっている．

クエリリスト　この検索において生成され利用されたすべてのクエリが表示されている．ユーザはこのうち特定のクエリの検索結果に興味を持った場合，そのクエリの右側の虫眼鏡ボタンを押すことでその検索結果を確認することができる．

同位語リスト　緩和度の付加されたキーワードに対して自動抽出された同位語が表示されている．赤色が現在利用されている同位語で灰色が利用されていない同位語である．ユーザはここに並べてある同位語の中から自分の興味のあるものを選択し，ReSearch ボタンを押すことでその同位語により緩和された検索結果を得ることができる（図 4）．これにより，ユーザは自分の興味の範囲を容易にシステムに伝えることが可能となっている．

6.2 インタラクション

ユーザは入力ボックスのみで検索することが可能であり，緩和したい場合も緩和したいキーワードに「?」をつけるだけで緩和された結果を得ることができる．緩和された結果を得たユーザは，右側の各リストによりどのような同位語やクエリ群によって緩和された結果なのかを一目で把握することができる．ユーザはページのリストからよさそうなページを選ぶこともできるが，同位語リストに他に興味のある同位語があればその話題も含む結果を得ることができ，クエリリストの中に興味のあるクエリがあればその検索結果にジャンプすることもできるといったように，興味の遷移にも対応できるようになっている．

7. ま と め

本稿では，緩和検索において単純な検索範囲の拡張を目的としたランキング手法と理解・比較をサポートするランキング手法を提案した．また，2 つのランキング手法がそれぞれ異なる検索目的において好まれることをユーザ評価によって示した．本稿では，Web 検索においてユーザが求める範囲をより柔軟に指定できる検索システムにおいて，クエリの共起性に基づくランキング手法を提案した．また，クエリの共起性に基づくランキングを行うことで複数の話題を含む有用なページを上位に並び換えることができることを確認した．

今後の課題はより踏み込んだランキング手法の提案とその評価である．キーワード間だけではなく同位語間に成り立つ関係性も考慮し，それをクエリ生成やランキングに反映していこうと考えている．また，各ランキング手法をユーザに指定してもらうインタラクションの実現や，可能であれば最適なランキング手法の自動推定も行いたいと考えている．その上でシステムを再実装し検索結果だけではなくインタラクションも含めてシステム全体の評価を行っていく予定である．

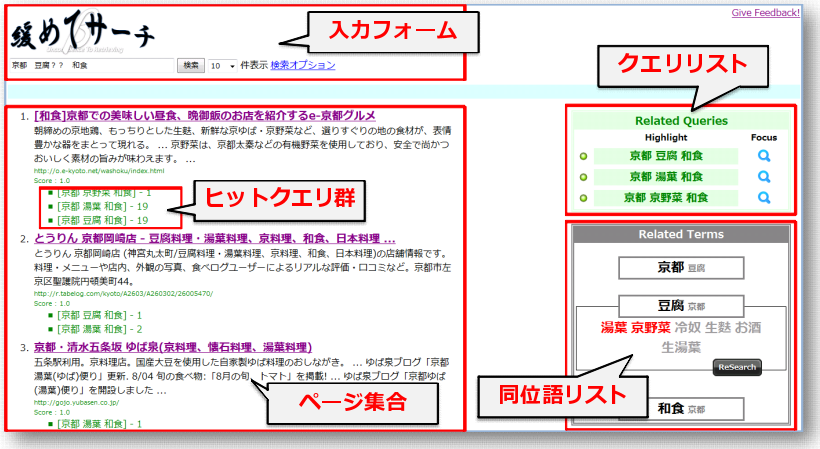

図 3　システム実行例

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）によるものです．ここに記して謝意を表すものとします．

参 考 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) Y.Kaneko, S.Nakamura, H.Ohshima and K.Tanaka: “Query Relaxation Based on Users’ Unconfidences on Query Terms and Web Knowledge Extraction”, pp. 71–81 (2008).

Vol.2009-DBS-149 No.1
2009/11/20

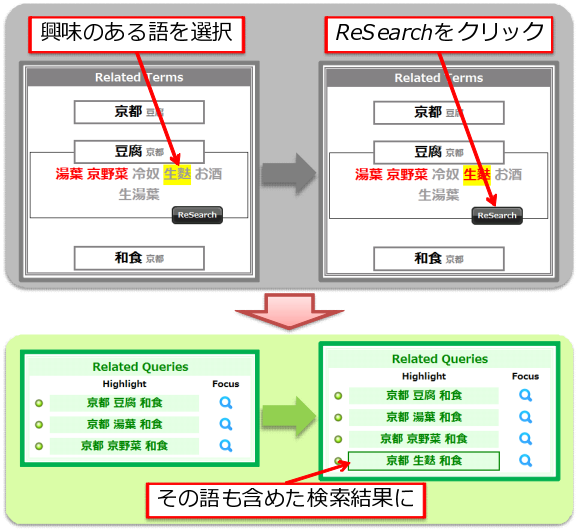

図 4 同位語リスト中に興味のある語があれば，それを選択することで検索結果に反映可能

3) R.White and D.Morris: “Investigating the querying and browsing behavior of advanced search engine users”, Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrievalACM, p. 262 (2007).
4) 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).
5) 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).
6) 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).
7) 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.
8) 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.
9) 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).
10) Z.Ghahramani and K.Heller: “Bayesian Sets”, ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, **18**, p. 435 (2006).
11) A.Broder: “A taxonomy of web search”, ACM Sigir Forum, Vol.36ACM, p.10 (2002).
12) 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).
13) 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 informaion retrievalACM New York, NY, USA, pp. 64–71 (2003).
14) J.Rocchio, et al.: “Relevance feedback in information retrieval”, The SMART Retrieval System: Experiments in Automatic Document Processing, pp. 313–323 (1971).
15) 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).
16) 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.
17) C.Dwork, R.Kumar, M.Naor and D.Sivakumar: “Rank aggregation methods for the web”, Proceedings of the 10th international conference on World Wide WebACM New York, NY, USA, pp. 613–622 (2001).
18) J.Aslam and M.Montague: “Models for metasearch”, Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrievalACM New York, NY, USA, pp. 276–284 (2001).
19) M.Renda and U.Straccia: “Web metasearch: rank vs. score based rank aggregation methods”, Proceedings of the 2003 ACM symposium on Applied computingACM New York, NY, USA, pp. 841–846 (2003).