DEIM Forum 2015 E8-6

探索型情報検索のための未知度提示システム

篠原　拓也†　　北山　大輔 †

† 工学院大学情報学部コンピュータ科学科　〒 163-8677 東京都新宿区西新宿 1-24-2
E-mail: †j111053@ns.kogakuin.ac.jp, ††kitayama@cc.kogakuin.ac.jp

あらまし　検索目標が必ずしも明確でない状態で検索を行う探索型検索において，ユーザは複数の Web ページを閲覧する．しかし，WWW には同様の内容を含む Web ページが多く存在している．そのため，得られる情報量が少ない Web ページを選択してしまうことがあり，効率よく情報収集を行うことができない．本論文では検索結果に Web ページごとの未知度を表示し，ユーザが多くの情報を得られるように支援するシステムの提案を行う．我々は Web ページにユーザが知らない情報が含まれている割合を未知度と定義した．提案手法ではユーザが閲覧した Web ページの単語とまだ閲覧していない Web ページの単語を比較し，未知度を算出する．

キーワード　Web 検索，未知度，閲覧履歴，検索支援

1. はじめに

近年，インターネットの普及や検索技術の向上により，ユーザは求めている情報を見つけやすくなっている．特に，ユーザが検索対象に関して十分な知識を持っていて，明確な情報要求がある場合は求めている情報を容易に発見することができる．一方，ユーザが検索対象の知識に乏しく，検索意図が曖昧な状態で検索を行う場合はユーザが求めている情報を発見することは容易ではない．このような検索を探索型情報検索 [1] [2] という．たとえば，京都に旅行に行きたいと考えた場合，明確な情報要求がある場合は「清水寺　アクセス」などの具体的なクエリで検索を行い，求める情報が見つかったら検索を終了する．しかし，明確な情報要求のない探索型情報検索の場合,「京都　観光」などの曖昧なクエリで検索を行う．この際，明確な目的がないため，さまざまな Web ページを閲覧し，情報を収集する．検索の終了条件は曖昧で，さまざまな情報を収集しながら探索ゴールを少しずつ明確にしていく．また，情報を収集している最中に検索目標を変更することもある．たとえば，京都に旅行に行こうと京都の観光スポットについて調べている時に，いくつかの観光スポットの情報の中から興味のある観光スポットを見つける．その後，今度はその観光スポットのそばにあるお店についての情報を収集する．探索型情報検索において，ユーザは自身がまだ知らない未知の情報を求めて検索を行う．

そこで，本研究では探索型情報検索の際に支援を行うことを目的とし，ユーザの知らない情報が含まれている割合を未知度として表示することによってユーザが効率的に情報収集を行うことを可能にする．未知度提示の概念を図 2 に示す．

以下に本論文の構成について記す．次章では本研究の概要と関連研究について述べる．3 章では，未知度の算出方法を示す．4 章では，既知の単語の予備実験について述べる．5 章では，提案手法のリランキングにより効率情報を集められるかを他の手法と比較して評価する．最後に，まとめと今後の課題について述べる．

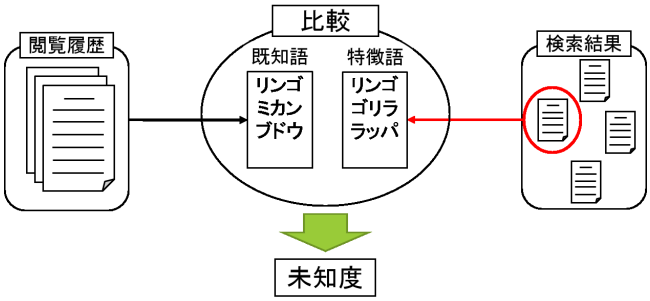

図 1　本研究のアプローチ

2. 本研究の概要と関連研究

2.1 本研究の概要

WWW には同様の内容を含む Web ページが多く存在している．実際,「京都　観光」のクエリで Web 検索を行うと，観光のポータルサイトやおすすめスポットリストが見つかるが，紹介される観光スポットの多くは重複している．このとき，1 度閲覧した Web ページと同様の内容の Web ページを閲覧しても得られる情報が少ないと考えられる．効率よく情報収集を行うためにはユーザにとって未知の情報が多く含まれている Web ページを選択する必要がある．そこで，本研究では検索結果に未知度を表示し，ユーザが効率よく情報を得られるよう支援するシステムの提案を行う．我々は Web ページにユーザが知らない情報が含まれている割合を未知度と定義した．本研究のアプローチを図**??**に示す．ユーザにとって未知の情報か既知の情報かを判断するために Web 閲覧履歴を使用する．ユーザが閲覧した Web ページの単語からユーザが既知の単語を取得する．その後，まだ閲覧していない Web ページにユーザが既知でない単語が含まれる割合を算出し，未知度として表示する．

2.2 関連研究

2.2.1 検索結果の提示手法

検索結果の提示を工夫する研究として湯本ら [3] [4] の研究や北原ら [5] の研究があげられる．湯本らは複数の Web ページを

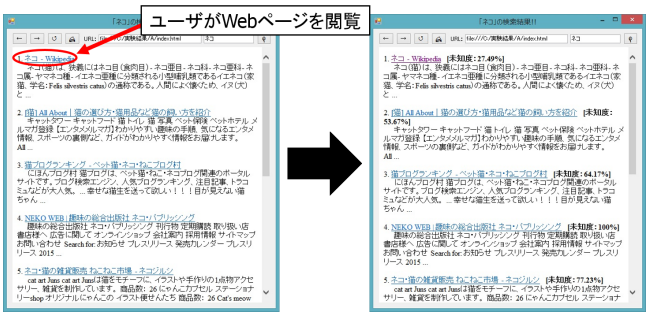

図 2　未知度提示の概念図

統合したページセットを検索の解とみなす統合型検索を提案している．統合型検索によって全容検索や比較検索を行うことを可能にしている．また，ページセットを構成するページ間の関係に着目し，特異箇所を抽出し，利用することによって，統合型検索によって得られた検索解をより効率的かつ効果的に閲覧することを可能にしている．しかし，本研究ではユーザ自身が Web ページを選択する際の支援を行っており，ユーザにとって未知の情報が含まれているかを考慮して情報検索の支援を行っている．

北原らの研究では Web 検索結果と同時に表示されるスニペットだけでは検索結果のリスト内においてクエリに関する内容が取り扱われているかを判断することが困難という問題を解決するために，クエリに関する内容の出現範囲及び局所的な内容の影響度を表す内容密度分布を抽出している．しかし，本研究ではユーザにとって未知の情報が多く含まれている Web ページの発見を支援し，効率よく検索を行えるようにすることが目的であるため，北原らの研究とは目的が異なる．

2.2.2　未知情報がある Web ページの推薦

ユーザにとって未知の情報を推薦する手法として希少な Web ページを推薦する研究がある．多田ら [6] はカテゴリへの所属度から有用性を，カテゴリ内の類似する Web ページの数と本文中の単語から非典型度を算出し，ユーザにとって未知であり検索が難しい希少な Web ページを発見する研究を行っている．本研究もユーザにとって未知である情報の発見を支援しているが，Web ページの希少度は考慮しておらず，典型的か非典型的かにかかわらず未知の情報多く記載されている Web ページの発見を支援している．

2.2.3　閲覧履歴による Web ページの推薦

ユーザの閲覧履歴を用いて検索支援を行う研究として堀ら [7] や佐藤ら [8] の研究があげられる．堀らの研究ではユーザの Web ページ閲覧行動と検索意図には関係があると考え，ユーザの Web 閲覧履歴に出現する単語を用いて検索語を拡張することよって，ユーザが求めている Web ページが検索結果の上位に来るようにしている．

佐藤らの研究ではユーザの閲覧履歴から探索行動を抽出している．また，探索行動からユーザの探索目的の事柄が記載された有用な Web ページも抽出している．

本研究では未知度の高い Web ページほど得られる情報量が多いと考え，ユーザの閲覧した Web ページと内容の異なる Web ページの発見の支援が目的であるため，堀らや佐藤らの研究とは目的が異なる．

2.2.4　リランキングによる検索支援

リランキングによる検索支援を行う研究として高見ら [9] や倉門ら [10] の研究があげられる．高見らはスニペットに着目し，ユーザが選択した特定のスニペットを基準に，そのスニペットと類似するように他のスニペットを再生成する手法を提案している．さらに，再生成したスニペットを類似度でリランキングし，情報検索の支援を行っている．しかし，高見らはスニペットに着目しているのに対し，本研究では Web ページに含まれている情報に着目しており，Web ページ内にユーザが知らない情報が含まれている割合をリランキングの基準としている．

倉門らの研究では Wikipedia に基づいたリランキング手法を提案している．リランキングのために Wikipedia から利用できる素性として，「inlink」，「outlink」，「リンク共起」，「カテゴリ」の 4 つがあると考え，それぞれを利用し，様々な手法でリランキングを行っている．しかし，本研究ではリランキングの基準として Wikipedia を用いていない．また，倉門らは Web ページの評価の際，クエリとの関連度を Wikipedia を利用したモデルで算出し，関連度の高い見出し語を多く含む Web ページを良いサイトと評価しているが，本研究では関連する単語が含まれる数ではなく含まれる割合を Web ページの評価に用いている．

3.　閲覧履歴に基づく未知度の算出

3.1　既閲覧 Web ページを用いた単語の既知度の算出

未知の Web ページを発見するためには，まずユーザが何を知っているのかを把握する必要がある．本研究では，既知情報を単語の集合で表現し，各単語に対して，既知である可能性を示す既知度をあたえる．

単語の既知度を算出するために形態素解析を行い名詞を抽出する．なおこの時，ユーザが閲覧した Web ページのテキストに出現する半角記号を全角スペースに置換する．これは形態素解析した際に，半角記号が名詞と判断されてしまうためである．また，名詞が連続して出現する場合は複合名詞とみなし，連続する名詞をまとめてひとつの名詞として抽出する．さらに，名詞でも単体では意味を持たない接続名詞，非自立名詞および代名詞などは抽出しない．形態素解析には辞書やコーパスに依存せず，辞書に登録されていない未知の単語に対して品詞を推測できるという特徴を持つ MeCab [11] を用いる．

次に，抽出した名詞の中からストップワードに登録されている名詞と検索クエリに使用した単語を除外する．ストップワードには “A” などのアルファベット 1 文字や “一覧” や “リンク” などの多くの Web ページに出現するが Web ページの特徴を表さない単語が登録されている．ストップワードや検索クエリに使われた単語を除外し，残った名詞を Web ページの特徴語とする．

既閲覧 Web ページの特徴語をユーザにとって既知の単語と考える．しかしこの時，出現回数に応じてユーザのその単語に対する記憶への留めやすさが異なると考えた．すなわち，出現回

数が多いほど記憶に残るが，出現回数が少ない特徴語は，ユーザの記憶に残りにくく既知の単語にはならないと考えた．そこで，既閲覧 Web ページの特徴語において出現回数が N 回より少ない特徴語は既知の単語として扱わない．つまり，出現回数が N 回以上の特徴語を既知の単語とする．既知の単語に対して次式を用いて既知度を算出する．

$$KW_i = \log_{10}(tf_i^h + 1) \tag{1}$$

tf_i^h はある単語 i が既閲覧 Web ページ集合に出現する回数の総和である．式 (1) によって出現回数が多い特徴語ほど既知度が高くなるようにする．

3.2 検索結果ページの未知度の算出

ユーザが入力したクエリに対して検索結果から Web ページを取得する．3.1 節と同様の方法で取得した Web ページの特徴語を抽出する．その後，Web ページごとの未知度を次式を用いて算出する．

$$Unknown = \frac{\sum_{u \in U}(tf_u^c \times UW_u)}{\sum_{k \in K}(tf_k^c \times KW_k) + \sum_{u \in U}(tf_u^c \times UW_u)} \times 100 \tag{2}$$

$$UW_i = \log_{10}(tf_i^c + 1) \tag{3}$$

tf_i^c は検索結果の Web ページ内の単語 i の出現回数である．K は検索結果の Web ページ内の既知の単語であり，U は既知ではない単語である．UW は Web ページに出現する未知の単語に対する重みである．重みはその単語の出現回数が多いほど大きくなる．また，未知度は%で表示するため最後に値を 100 倍する．

3.3 ユーザの指定に合わせたスコア算出

本研究では Web ページの全単語に対する未知の単語の割合から未知度を算出している．未知度が高い Web ページほどユーザは既知ではない新しい情報を得ることができる．しかし，未知度が高い Web ページほど今まで閲覧した Web ページの内容とは関係のない内容となってしまう．ユーザの求めている情報によっては今まで閲覧した Web ページの内容とある程度関係している．つまり，未知度が高すぎない Web ページを求めている場合がある．そこで，ユーザが理想の未知度の値を指定し，それを基に次式を用いてスコアを算出し，リランキングを行う．図 3 は検索結果をリランキングしたものである．

$$Score = 100 - |X - Unknown| \tag{4}$$

X はユーザの指定した未知度の値を表し，ユーザの指定した未知度に近いほど値が高くなるようにスコアを算出する．

本研究ではユーザが新しい Web ページを閲覧するたびに，閲覧した Web ページから既知の単語を取得し，未知度の再計算を行い，リランキングする．図 4 が全体のフロー図である．

4. 既知の単語の予備実験

本実験の目的は Web ページを閲覧した際，出現回数に応じて単語に対する記憶への留めやすさが異なるかの確認である．また，既知とする単語の出現回数の閾値 N の決定も行う．

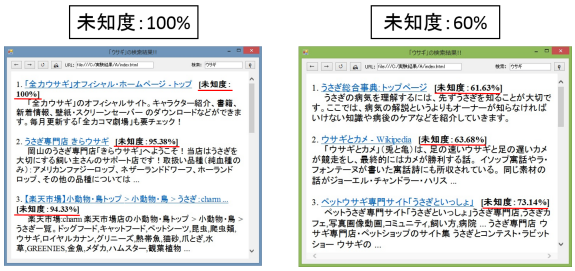

図 3　リランキング

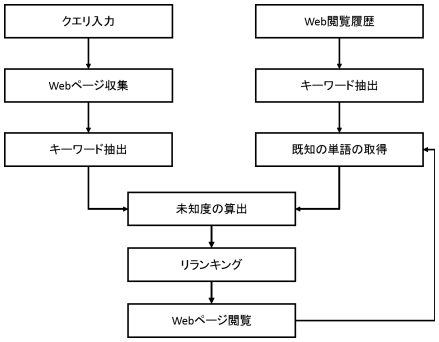

図 4　全体のフロー図

4.1 実験概要

被験者に Web ページを閲覧してもらった後に，単語を 30 個提示し，どの単語が閲覧した Web ページに出現したかを答えてもらい正解率を算出した．被験者に提示する単語はその Web ページに出現する単語と出現しない単語を合わせた単語集合から 30 個ランダムに取り出した．単語集合は Web ページに出現する単語が 7 割，出現しない単語が 3 割の割合となるようにした．被験者は 10 人で各被験者 5 つの Web ページを閲覧してもらった．

4.2 予備実験の評価

結果を表 1 に示す．2〜や 3〜はその数字以上の出現回数の単語のみの正解率である．つまり，2〜は出現回数が 1 回の単語を除外した正解率，3〜は 2 回以下の出現回数の単語を除外した正解率である．正解率を比較してみると，出現回数が少ない単語を除外し，出現回数が多い単語のみになるにつれて正解率は高くなった．このことから，単語の出現回数は Web ページ閲覧時の記憶の残りやすさに関係があるといえる．

出現回数が 2 回以上の単語の正解率と 3 回以上の単語の正解率の間で 0.11 と大きな差があった．また，出現回数が 3 回以上の単語の正解率は 0.7 を超えており，多くの場合で記憶に残っていると考えられる．以上から本研究では既知の単語を取得する際の閾値 N を 3 とし，出現回数が 3 回以上の単語をユーザの既知の単語とする．

5. 評価実験

本実験の目的は，提案システムのリランキングによりユーザが効率よく検索を行えるかの確認である．

5.1 実験概要

評価を行うため，適合性フィードバック，適合性フィードバッ

表 1　出現回数と正解率

	特徴語数	1〜	2〜	3〜	4〜	5〜
ページ 1	363	0.48	0.71	0.83	0.93	1.00
ページ 2	283	0.48	0.64	0.64	0.70	0.75
ページ 3	307	0.46	0.59	0.80	0.79	0.72
ページ 4	199	0.47	0.63	0.67	0.73	0.79
ページ 5	567	0.43	0.45	0.58	0.72	0.76
平均		0.46	0.60	0.71	0.77	0.80

表 2　被験者ごとの手法と検索キーワードの組み合わせ

	降順	昇順	提案
ウサギ 飼い方	B，E，H	C，F，I	A，D，G
エボラ出血熱	A，F，G	B，D，H	C，E，I
クリスマス　起源	C，D，I	A，E，G	B，F，H
三大紅茶とは	C，D，I	A，E，G	B，F，H
金閣寺　歴史	A，F，G	B，D，H	C，E，I
オブジェクト指向	B，E，H	C，F，I	A，D，G

クの逆順位，提案手法の 3 つで比較を行う．システムに検索キーワードを与え，検索結果の上位 50 ページを取得した．検索エンジンには Bing を用いた．被験者には検索結果上位 10 ページのみを提示し，その中から Web ページを選択してもらった．被験者が選択した Web ページを基にそれぞれの手法で検索結果のリランキングを行い，また上位 10 ページのみを被験者に提示した．これを被験者が 5 ページ閲覧するまで繰り返し，5 ページ閲覧したらその検索キーワードについての検索を終了とした．被験者が検索キーワードについての検索を終了したら単語集合を提示し，その単語について知識を得ることができた単語を選択してもらった．被験者に提示する単語集合はその検索キーワードの検索結果 50 ページに出現する単語の出現回数上位 50 個とした．検索キーワードは全部で 6 種類 (「ウサギ　飼い方」,「エボラ出血熱」,「クリスマス　起源」,「三大紅茶とは」,「金閣寺　歴史」,「オブジェクト指向」) 用意し，被験者は各手法 2 種類ずつ検索を行った．被験者ごとの各手法と検索キーワードの組み合わせを表 2 に示す．提案は提案手法，降順は適合性フィードバック，昇順は適合性フィードバックの逆順位のことである．

5.2　実験結果と考察

結果を表 3 に示す．被験者によって全体的に多く選択する被験者や少ししか選択しないなど差が見られた．そこで本研究では被験者の選択した単語数の平均との差を評価に用いる．各被験者の平均との差を表 4 に示す．

選んだ単語数の差を手法ごとにまとめたものを表 5 に示す．数値は数値は被験者が選択した単語数の平均との差の平均であり，数値が高いほど多くの単語について知識を得ることができたことを表している．6 個の検索キーワードのうち最も高い値となったのが提案手法が 3 個，適合性フィードバックが 2 個，適合性フィードバックの逆順位が 1 個と提案手法が最も多かった．

知識が得られた単語数の合計を見てみると適合性フィードバックが最も多く，提案手法は 2 番目に多かった．しかし，知識を得ることのできた単語を比較してみると,「金閣寺　歴史」の場合，適合性フィードバックで得られ，提案手法では得られなかった単語としては「拝観」,「北山」,「室町時代」など金閣寺そのものを表す単語が多くトピックに偏りが見られたが，提案手法の場合で得られ，適合性フィードバックでは得られなかった単語としては「舎利」,「銀閣寺」,「夕佳亭」など幅広く知識を得ることができていた．

表 3　知識を得ることのできた単語数

	A	B	C	D	E	F	G	H	I
ウサギ　飼い方	23	14	9	4	14	7	18	21	14
エボラ出血熱	22	18	18	10	21	11	23	26	17
クリスマス　起源	13	14	14	12	13	3	9	17	14
三大紅茶とは	13	17	14	7	12	6	21	17	14
金閣寺　歴史	16	12	18	5	7	3	16	14	13
オブジェクト指向	16	35	13	18	12	7	19	26	8
平均	17.17	18.33	14.33	9.33	13.17	6.17	17.67	20.17	12.83

表 4　平均との差

	A	B	C	D	E	F	G	H	I
ウサギ　飼い方	5.83	-4.33	-5.33	-5.33	0.83	0.83	0.33	0.83	1.17
エボラ出血熱	4.83	-0.33	3.67	0.67	7.83	4.83	5.33	5.83	4.17
クリスマス　起源	-4.17	-4.33	-0.33	2.67	-0.17	-3.17	-8.67	-3.17	1.17
三大紅茶とは	-4.17	-1.33	-0.33	-2.33	-1.17	-0.17	3.33	-3.17	-1.83
金閣寺　歴史	-1.17	-6.33	3.67	-4.33	-6.17	-3.17	-1.67	-6.17	0.17
オブジェクト指向	-1.17	16.67	-1.33	8.67	-1.17	0.83	1.33	5.83	-4.83

表 5　手法ごとの結果

	降順	昇順	提案
ウサギ 飼い方	-2.67	-3.33	0.83
エボラ出血熱	14.99	6.17	15.67
クリスマス　起源	3.51	-13.01	-10.67
三大紅茶とは	-4.49	-2.01	-4.67
金閣寺　歴史	-6.01	-16.83	-2.33
オブジェクト指向	21.33	-5.33	8.83
合計	26.66	-34.34	7.66

6.　まとめと今後の課題

本論文では，ユーザの探索型情報検索の支援を目的とする，未知度提示システムを提案した．そして，未知度に応じたリランキングにより効率的に情報を得られるのかを他の手法と比較し評価した．実験の結果，提案手法では効率的に幅広く情報を得ることができることを確認した．

今後の課題として時間経過を考慮した既知度の変更があげられる．得た情報は時間がたつことに忘れていくと考えられるので，情報を得てから時時間がたつごとに既知度が低下させる必要がある．また，関連する項目を検索した際の未知度の確認も行う必要がある．関連していて似た単語が多く出てくる場合でも未知度が正しく機能するかを確認する．そのほかの課題としてキーワードレベルだけではなく，トピックレベルで既知度や未知度の判定を行うことがあげられる．また，未知度を提示することによってユーザの検索行動にどのような変化が起こるのかを確認する必要がある．

謝　　辞

本研究の一部は，平成 26 年度科研費若手研究 (B)(課題番号 : 24700098) 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] http://rerank-lab.org/exploratory-search/
[2] White, R.W. and Roth, R.A. (2009). Exploratory Search: Beyond the Query Response Paradigm. San Rafael, CA: Morgan and Claypool.
[3] 湯本高行，田中克己．“Web ページ集合を解とする全容検索”．情報処理学会論文誌 : データベース Vol48 No.SIG(TOD34)．(2007)
[4] 湯本高行．“統合型検索のための Web ページ間の意味的な関係に基づく検索結果の提示手法”．DEWS2008 B6-1
[5] 北原沙緒里，田村航弥，波多野賢治．“Web テキストにおける内容密度分布の抽出とその評価”．DEIM Forum 2011 F1-2
[6] 多田亮平，湯本高行，新居学，佐藤邦弘．“カテゴリに対する所属度と典型度を考慮した希少な Web ページの発見”．情報処理学会研究報告 Vol.2012-DBS-155 No.13．(2012)
[7] 堀幸雄，今井慈朗，中山堯．“ユーザの Web 閲覧履歴を用いた検索支援システム”．情報知識学会誌 Vol.17 No.12．pp.95-100．(2007)
[8] 佐藤大樹，菅原俊治．“ブラウザの閲覧履歴に基づく探索行動抽出手法”．電子情報通信学会 AI2010-6．pp.31-35．(2010)
[9] 高見真也，田中克己．“類似性を考慮したスニペットの再生成による検索結果のリランキング”．日本データベース学会 Letters．pp.109-112．(2007)
[10] 倉門浩二，大石哲也，長谷川隆三，藤田博，越村三幸．“Wikipedia のリンク共起とカテゴリに基づくリランキング手法”．情報処理学会研究報告 Vol.2010-DBS-150 No.12．(2010)
[11] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237 (2004.)