社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－DBS－143（9）
2007／7／2

Webページに対する典型的なクエリの発見

甲谷 優† 湯本 高行†† 小山 聡† 田中 克己†

† 京都大学大学院情報学研究科 〒606-8501 京都市左京区吉田本町
†† 兵庫県立大学大学院工学研究科 〒671-2201 兵庫県姫路市書写2167
E-mail: †{kabutoya,oyama,tanaka}@dl.kuis.kyoto-u.ac.jp, ††yumoto@eng.u-hyogo.ac.jp

あらまし 現在，多くのユーザは検索エンジンを用いてWebページを探索し閲覧している．その際の検索式(クエリ)は，Webページ内に潜在的に存在するコンテンツ利用者のニーズを端的に表す要約であると考えられる．したがって，Webページに対する「典型的」なクエリを発見できれば，そのWeb上における位置付けや関連文書，ページ内の利用者ニーズの高い箇所の発見や，さらにはそれを踏まえてのそのページのさらなる内容の充実に役立てられるものと考えられる．本論文では，Webページpに対する典型的なクエリqを，コンテンツ利用者がqで検索してpを閲覧する確率を最も高くするようなものと定義する．本研究では，典型的クエリの候補として実際に使用され得るクエリに限定する．そこで，具体的にどのようなクエリがどのくらいの頻度で実行されているのかという情報を取得するために，検索エンジンのクエリログを利用する．さらに典型的クエリ候補を用いて実際に検索した際にページpを閲覧する確率をqで検索したときのpの順位から推定して利用する方法を提案する．

キーワード 典型的クエリ，クエリログ，ランキング

Discovering Typical Queries for Web Pages

Yutaka KABUTOYA†, Takayuki YUMOTO††, Satoshi OYAMA†, and Katsumi TANAKA†

† Graduate School of Informatics, Kyoto University, Yoshida Honmachi, Sakyo, Kyoto, 606-8501 Japan
†† Graduate School of Engineering, University of Hyogo, Shosha 2167, Himeji-shi, Hyogo, 671-2201 Japan
E-mail: †{kabutoya,oyama,tanaka}@dl.kuis.kyoto-u.ac.jp, ††yumoto@eng.u-hyogo.ac.jp

Abstract Nowadays many users use search engines to find and browse a desirable Web page. Those queries to reach a given page concisely describe the needs of users (contents consumers) existing potentially for the page. Therefore, finding "typical queries" for a given Web page will be useful to discover the interests of Web users to the page, and will be moreover, useful for the author to improve its content. In this paper, we define the most "typical query" q for a given page p as the most frequent query by which consumers reach the page p through a Web search engine, and visit the page p. In our research, candidates for the typical query is assumed order to find the "typical queries" for a given Web page, we need actual query frequency information, and so we use queries owned by a search engine. The probability of actually visiting the page p after finding the page p by the query q is calculated from the rank of p in executing q to a Web search engine.

Key words typical query, query log, ranking

1. はじめに

現在，多くのユーザは検索エンジンの提供するキーワード検索サービスを通してWebページを発見し閲覧している．ユーザは自身の検索意図を何語かのキーワードで表し，それをクエリとしてWeb検索を実行する．それにより得られた検索結果のうち，ユーザは自身の検索意図と合致するものを選択し，それを閲覧していると考えられる．このモデルを鑑みれば，あるページpがあるクエリqによる検索によって発見され閲覧された場合，クエリqはページpがユーザのどのような検索意図と合致するかを端的に表す要約のようなものであると考えられる．

したがって，ページpの作成者にとってページpに対する典型的なクエリqを発見することはきわめて重要なことであると考えられる．なぜなら，qによりユーザがどのようなニーズでpを閲覧したかを知ることができ，さらにそのニーズからpの内容をどのように改善すればより被閲覧数を増やすことができ

るかがわかるからである．

以下に示すような内容のページ [1] を例に取って説明する．

> ... リンク構造（Web ページの被参照構造）を使用した．文脈情報を Web から抽出することで，画像がどのような文脈で使用されているか概覧することが出来，逆に，文脈情報を指定した画像検索が可能になる．さらに，これらの情報に基づき，Web ページそのものがどのような観点（アスペクト，aspect）から参照されているかを把握することが出来る．図は，dolphin というキーワードで検索される画像と，それらの３種の文脈構造を検索し，KWIC (Keyword in Context) 形式で表示したものである．複数の dolphin 画像の周辺にどのような単語や画像が配置されているかによって，各画像の利用の ...

例えば，このページが「画像検索」というクエリよりも，「アスペクト」というクエリによる Web 検索により閲覧される頻度が高いとする．このとき，このページはユーザ (コンテンツ利用者) にとって「画像検索」に関する情報に対する検索ニーズよりも「アスペクト」に関する情報に対する検索ニーズに応えていると考えられる．したがって，このページの被閲覧数を増やしたい場合，「画像検索」に関する情報を拡充するよりも「アスペクト」に関する情報を拡充する方が効果的であると考えられる．

本論文では，あるページ p に対する最も典型的なクエリを，p を閲覧するのに利用された頻度で評価する．具体的には，クエリの実行頻度と，そのクエリで Web 検索したときのページ p の順位から決定する．これは，クエリ q で p を閲覧する回数は，q の実行頻度と，q での Web 検索結果から p を選択する確率の両者の積で表されるという仮定に基づく．

最も典型的なクエリの発見に，本研究ではドメインを実際に検索エンジン Yahoo(注1) に実行されたクエリに限定し，そのクエリログを参照できるサービスである Overture のキーワードアドバイスツール(注2) を利用する．このサービスを利用することにより実際にどんなキーワード群からなるクエリが実行されたか，またそのクエリがどれくらいの頻度で実行されたかの情報を取得できる．本研究では，まずクエリログからどれが典型的なクエリになるか，その候補を発見する手法を提案する．

また，本研究ではクエリ q での検索結果からページ p を選択する確率を，q での検索結果における p の順位と負の相関があると仮定する．これは，Konishi [2] らによる，「ユーザは Web 検索においてほとんど上位のコンテンツしか閲覧しない」という調査結果に基づく．本論文では，クエリ q での検索におけるページ p の順位を求めるために，実際に検索エンジンでクエリ q を実行してみてページ p の順位を計測する．

本論文の構成は以下の通りである．まず，2 節にて本研究の関連研究について述べる．3 節では，Web ページが与えられたときの，それに対する典型的なクエリと，典型的なクエリとなり得る候補を評価するための手法について説明する．4 節では，検索エンジンのクエリログを利用することによりクエリ候補を取得し，クエリ候補から典型的なクエリを選択する手法について説明する．5 節では，実際の Web ページからそれを検索するための典型的なクエリを発見する実験と，その結果及び考察について説明する．最後に，6 節ではまとめと今後の課題について言及する．

2. 関連研究

2.1 重要語抽出

Nakagawa らは，言選 Web(注3) というサービスを提供している．これは，Web ページ中に含まれる専門用語を切り出すシステムである．また，Ohsawa らはある文章がどのような内容であるか，どのような話題構造を持っているかを調べるために，重要であろうと考えられる語を複数抽出し，共起するものどうしを線で結ぶことによりグラフ表示し可視化する，KeyGraph [3] と呼ぶ手法を提案している．これらのシステムは単一の語及びそれらの関係を抽出するもので，ページからそれを検索するための複数の語からなるクエリを抽出する本研究とは異なる．またそれぞれに対して目的も異なる．

2.2 Web ページのアスペクト抽出

Zettsu [4] らは，画像やテキストの文脈を，周辺情報から推定するために，画像やテキストの周辺の情報，画像・テキストを含む領域の上位領域，Web ページへのリンク構造を使用した．この研究は，与えられた Web ページが他のページからどのように見られているかという情報を周辺情報 (クエリログ) から取得するという点で，目的は本研究と近い．

しかし，この研究と本研究との違いは，その周辺情報が誰によるものかが異なる．すなわち，アスペクトとは他のコンテンツ作成者からの見られ方を反映した周辺情報であり，本研究の提案する典型的なクエリとはコンテンツ利用者からの見られ方を反映した周辺情報であると位置づけることができる．

2.3 キーワード式生成による質問修正支援

Matsuike ら [5] は，ユーザの実行したクエリにより得られる検索結果を，複数のキーワードを組み合わせたキーワード式により完結に表すことで，ユーザの知識発見や概要理解，及び質問修正を支援するシステムを提案している．ページに対して複数の語からなるクエリやキーワード式を生成するという点は本研究と共通しているが，そのクエリ生成に使っている情報がクエリログではなく検索結果ページである点が本研究とは異なる．また，支援する対象もコンテンツ利用者であり，コンテンツ作成者を対象とした本研究とは異なる．

3. Web ページに対する典型的クエリ

ある Web ページ p が与えられたとき，ユーザは検索エンジンを用いたキーワード検索によって以下のようなステップを通して p に到達する．

(注1)：http://www.yahoo.com/

(注2)：http://inventory.jp.overture.com/d/searchinventory/suggestion/

(注3)：http://gensen.dl.itc.u-tokyo.ac.jp/

(1) ユーザのニーズを表すキーワードをクエリ q として検索を実行

(2) 検索結果リストからその中に含まれている p を表す要素を選択，リンクを辿って閲覧

この一連のステップの頻度，すなわち**クエリ q で検索してかつその検索結果からページ p を閲覧する回数**を PageView(p,q) と表すことにする．このとき，本研究におけるページ p に対する典型的なクエリ q を，PageView(p,q) を最大にするようなクエリ q であると定義する．

このように定義したとき，ページ p に対する最も典型的なクエリ q は，**ページ p に潜在的に存在するインターネット利用者のニーズ**を最もよく表すクエリであると考えられる．ただし，ここではインターネット利用者の行う検索のうち，[6] にて語られている **informational search**(注4)及び **navigational search**(注5) の両者を想定している．

クエリ q の実行頻度を q の Query Frequency 値として qf(q) と定義する．クエリ q で検索した際の結果にページ p が含まれるとき，**ページ p のクエリ q に対する検索結果順位**を rank(p,q) で表す．また，**クエリ q で検索したときの検索結果からページ p を選択する確率**を P$(p|q)$ と表すことにする．このとき，以下の等式が成立する．

$$\mathrm{PageView}(p,q) = \mathrm{P}(p|q) \cdot \mathrm{qf}(q) \tag{1}$$

したがって，P$(p|q)$ と qf(q) が求まれば PageView(p,q) の値が求まることがわかる．qf(q) の求め方は 4 節にて言及する．また，本研究では Konishi [2] らの行った調査結果に基づき P$(p|q)$ は**クエリ q で検索したときのページ p の検索結果順位** rank(p,q) に負の相関があると仮定する．また，各検索結果の閲覧数にスケールフリー性 [7] の存在を仮定し，P$(p|q)$ が rank(p,q) で以下のように表されると仮定する．このとき，

$$\mathrm{P}(p|q) = C^{-\mathrm{rank}(p,q)} \tag{2}$$

ただし，C は 1 より大きい何らかの定数とする．

ここで，クエリ q で検索したとき検索結果順位 $i(i = 1,2,\ldots,n)$ 位のページを p_i と表すことにする．このとき，以下の等式が成立する．

$$\mathrm{rank}(p_i,q) = i \tag{3}$$

このとき，1 度の検索につき常に 1 つの検索結果しか見ないとして以下の等式の成立を仮定する．

$$\sum_{i=1}^{n} \mathrm{P}(p_i|q) = 1 \tag{4}$$

このとき，式 (2)，(3)，(4) より

$$\sum_{i=1}^{n} \mathrm{P}(p_i|q) = \sum_{i=1}^{n} C^{-\mathrm{rank}(p_i,q)}$$

(注4)：目的となる情報のリソースが未知の状態で行われる検索

(注5)：目的となる情報のリソースが既知の状態で行われる検索

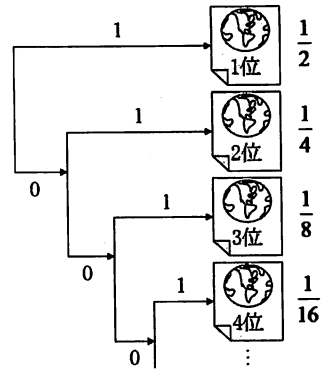

図 1　検索結果順位と閲覧確率の関係

$$= \frac{1 - C^{-n}}{C - 1} = 1 \tag{5}$$

n が十分大であるとすると，式 (5) より $1/C - 1 = 1$ が成立する．ゆえに $C = 2$．これは $C > 1$ を満たす．したがって

$$\mathrm{P}(p|q) = \frac{1}{2^{\mathrm{rank}(p,q)}} \tag{6}$$

が得られる．

このとき，検索結果を「選択する」を 1 で表し，「選択しない」を 0 で表すとすると，P$(p|q)$ のモデルは図 1 のように表される．

4. ページに対する典型的クエリの発見

本節では，具体的にどのようにして与えられたページに対する典型的なクエリを発見するかを説明する．

4.1 従来の手法とその問題点

典型的なクエリを求める対象となるページを p，それに含まれるキーワードの数を n とする．典型的クエリが AND 検索のみによって表されると仮定し，全てのキーワードの組合せに対し tfidf [8] 法のような何らかの評価関数で評価するという手法が考えられる．しかし，この場合 p に対する典型的クエリとなり得る候補の総数は，$O(2^n)$ となり組合せ爆発を起こす．したがって，そのような全てのクエリ候補に対し何らかの評価関数を用いて典型的クエリを求めるということはできない．

また，リファラを用いればページ p を含むサイトへの検索エンジンの流入ワードを獲得できる．しかし，この手法ではページ単位で典型的なクエリを求めることができないうえに，作成したばかりで検索エンジンにインデクスされていないようなページや，他者のサイトに存在するようなページに対する典型的なクエリを求めることはできない．

4.2 クエリログの参照に基づく典型的クエリ発見ステップ

そのため本研究では，典型的クエリの候補を取得するのに検索エンジン Yahoo のクエリログを参照することのできるサービスである Overture のキーワードアドバイスツールを利用する．キーワードアドバイスツールにより，前月の 1ヶ月間に実際にどのようなクエリが利用されたか，またそれぞれの Query Frequency 値を取得することができる．

表 1 「java」でのクエリログ参照結果

検索数	キーワード
47824	java
3492	java tea
789	java jre
615	java プログラミング
447	java eclipse
353	java インストール
289	java 環境変数
114	java curry

キーワードアドバイスツールの入力フォームに対し何らかのキーワードを入力し検索を行うと，結果としてそのキーワードに含まれる全ての形態素を含む実際に利用されたクエリとそれらの Query Frequency 値が得られる．たとえば「java」で検索した場合，表 1 のような結果が得られる．本研究では，Overture のキーワードアドバイスツールによってキーワード「w」の関連検索ワードを検索することを，「w」**でクエリログを参照する**と呼ぶことにする．

クエリログ参照により得られる情報を踏まえて，以下のようなステップによりページ p からそれを検索するための典型的なクエリを発見する．

（1） ページ p から名詞 10 個 $n_i (i = 1, 2, \ldots, 10)$ を抽出
（2） それぞれの n_i をシードとしてクエリログを参照し得られるクエリ候補の集合を Q_i とする
（3） $Q = Q_1 \cup Q_2 \cup \ldots \cup Q_{10}$
（4） Q のそれぞれの要素 q_i に対し評価関数 PageView(p, q_i) により評価
（5） PageView(p, q_i) の値がある閾値を越えるものを p の典型的なクエリとする

4.2.1 シードとなる名詞の抽出

本研究では，MeCab(注6) を用いた形態素解析を行うことによりページ p から名詞のみを抽出する．次に，[8] にある手法を用いて各名詞の特徴量として Term Frequency (TF) 値を計算する．その特徴量により各名詞をランキングしたとき，上位 10 個を典型的クエリ発見のためのシードとする．

4.2.2 クエリログ参照によるクエリ候補獲得

クエリログにあるキーワードを入力し関連するクエリを検索すると，出力として表 1 のように，それぞれの実行頻度 (Query Frequency 値) 順に上位 100 個のクエリ候補が得られる．しかし，Query Frequency 値が 100 位より低いものは取得することができない．そこで本研究では，クエリ候補に含まれる形態素の数と Query Frequency 値がトレードオフの関係にあることに着目し，出力として得られたクエリ候補でさらにクエリログを参照することにより取得することにより Query Frequency 値をより低くするロングテールのものも取得できると考えた．

得られた 10 個の名詞を $n_i (i = 1, 2, \ldots, 10)$ とする．n_i それぞれに対し，以下のステップによりクエリ候補集合 Q_i を取得する．

図 2 クエリログ参照によるクエリ候補獲得

（1） $Q_i \leftarrow n_i$，$queue \leftarrow \phi$
（2） $queue$ に n_i を enqueue
（3） $queue$ から q_{tmp} を dequeue
（4） q_{tmp} で クエリログを参照
（5） 得られたクエリ候補のうち，p を検索結果に含み得る (含まれる全ての形態素が p にも含まれる) もののみを δQ_i とする
（6） $Q_i' \leftarrow \delta Q_i - Q_i$，$Q_i \leftarrow Q_i \cup \delta Q_i$
（7） Q_i' の全ての要素を $queue$ に enqueue
（8） $queue = \phi$ ならば終了
（9） (3) に戻る

このステップの具体例として，ページから典型的クエリ発見のためのシードとして名詞「検索」が抽出されたときのクエリ候補の結果の一部を，図 2 に示す．キーワード「検索」による クエリログ参照では，得られるクエリ候補は Query Frequency 値が 4505 の「高速 道路 検索」までであり，青色の要素「検索 エンジン 情報」「検索 エンジン 情報」は，Query Frequency 値が低いためにキーワード「検索」によるクエリログ参照では取得できない．しかし，「検索」中に含まれるクエリ候補である「検索 エンジン」でクエリログを参照すると取得することができる．

4.2.3 クエリ候補からの典型的クエリ発見

クエリ候補 q がページ p を検索するための典型的クエリであるかを判定するための評価関数 PageView(p, q) は式 (1)，式 (2) 及び式 (5) より以下の等式で表される．

$$\mathrm{PageView}(p, q) = \frac{\mathrm{qf}(q)}{2^{\mathrm{rank}(p,q)}} \tag{7}$$

$\mathrm{qf}(q)$ はクエリログを参照することにより取得可能である．本研究では，$\mathrm{rank}(p, q)$ を取得するために，Yahoo デベロッパーネットワーク(注7)の Web 検索サービスを利用する．クエリ q にて検索し，上位 100 件を取得する．このとき，以下の等式で $\mathrm{rank}(p, q)$ を与える．

$$\mathrm{rank}(p, q) = \begin{cases} i & (i \in [1, 100],\ \text{かつ } p \text{ が } i \text{ 位}) \\ \infty & (p \text{ が } 100 \text{ 位以内に出現しない}) \end{cases} \tag{8}$$

PageView(p, q) が 0 を越えるとき (すなわちページ p がクエリ

(注6)：http://mecab.sourceforge.net/

(注7)：http://developer.yahoo.co.jp/

表 2　実験データとして用いたサンプルページ

	タイトル
p_1	TANAKA Laboratory - 田中研究室 -
p_2	Google の秘密 - PageRank 徹底解説
p_3	秋保温泉 - Wikipedia
p_4	明智光秀
p_5	本能寺の変
p_6	高レベル廃棄物の問題点
p_7	天網恢々疎にして漏らさず
p_8	オブジェクト指向を正しく理解する
p_9	スコティッシュ・フォールド
p_{10}	定石を覚えよう オセロ

q による検索で上位 100 件以内に出現するとき)，クエリ q はページ p の典型的クエリであるとする．

5. 予備実験

本研究では，ページが与えられたとき，提案手法によりどのようなクエリを取得できるか検証するために，実際の Web ページに本提案手法を適用し典型的クエリを発見するという予備実験を行う．

5.1 実験データ

表 2 に実験データとして用いたサンプルページの一覧を示す．ここでは，特にページ p_3 の例を用いて実験手法及び実験結果及びその考察を示す．以下，p_3(注8)の要約を示す．

> 秋保温泉は，宮城県仙台市太白区秋保町湯元（旧国陸奥国、明治以降は陸前国）にある温泉．仙台都心からも近いため，宿泊のみならず，日帰り入浴にも利用されている．同じ宮城県の鳴子温泉，福島県の飯坂温泉とともに奥州三名湯に数えられた．名取川にそって旅館ホテルが建つ．「秋保・里センター」を中心に広がる．初めに温泉街として発展したのは「秋保・里センター」の西側に当たる湯元地区である．この地区には平安時代に起源を有する宿のほか，数百年の業歴を持つ旅館が建ち並んでいる．温泉街には，仙台都心から 20-30 分と近く，高級ホテルのスイートルームに匹敵する部屋（離れ）を有する施設が複数存在し，仙台都市圏で最高の価格とサービスを提供しているため，賓客接待にも用いられている．同様に松島にも高級ホテル・旅館が存在することから，仙台都心にいわゆる高級ホテルが立地出来ない要因ともなっている。

秋保温泉の歴史及び周辺観光地や温泉街等の地理に関する情報が，p_3 に対するコンテンツ利用者のニーズであると考えられる．

5.2 実験結果

5.2.1 シードとなる名詞

4.2.1 項にある手法を用いて，p_3 から抽出された典型的クエリ発見のためのシードとなる名詞 10 個とそれぞれの TF 値について，表 3 に示す．

表 3　p_2 から抽出された名詞

名詞	TF 値	名詞	TF 値
温泉	31	ホテル	8
仙台	22	県	8
秋保	21	編集	8
湯	14	号	8
宮城	10	年	8

表 4　p_3 に対する典型的クエリ候補

クエリ候補 q	qf(q)	rank(p_3, q)	PageView(p_3, q)
秋保 温泉 街	72	2	18.000
秋保 温泉 温泉 街	16	1	8.000
秋保 日帰り 温泉	339	6	5.297
秋保 温泉 日帰り	1329	8	5.191
秋保 温泉 宴会	36	3	4.500
秋保 温泉 高級 旅館	30	3	3.875
秋保 温泉 観光 協会	88	5	2.750
秋保 温泉 アクセス	39	4	2.438
秋保 温泉 日帰り 湯	33	4	2.063
秋保 温泉	29765	14	1.817

表 5　$p_i (i = 1, \ldots, 10)$ に対する典型的なクエリ

対象ページ p	典型的クエリ q
p_1	研究 開発 データベース
p_2	pagerank
p_3	秋保 温泉 街
p_4	明智 光秀
p_5	織田 信長 本能寺の変
p_6	地層 問題
p_7	万物 復帰
p_8	指向
p_9	突然 変異 種
p_{10}	オセロ 定石

5.2.2 クエリ候補からの典型的クエリ発見

4.2.2 項にある手法により，表 3 に示されたそれぞれの名詞をシードとして，クエリログを参照することによりクエリ候補を得る．さらに，4.2.3 項にある手法によって典型的クエリを発見する．p_3 に対する典型的クエリ候補の一部 (PageView(p_3, q) 値上位 10 個) を表 4 に示す．また，p_1 から p_{10} それぞれに対する典型的なクエリを，表 5 に示す．

5.3 考　　察

表 3 の結果を見ると，純粋な出現頻度に基づく手法ではページが与えられたときにそれに対する典型的なクエリを求めることができないことが予測される．「年」「号」のような語は一般過ぎる語としてストップワードに指定できるかもしれないが，「ホテル」が p_3 の主眼でないことは tf 値では評価できない．

表 4 は，400 近くの数もあるクエリ候補から厳選された僅か 10 個であり，ほぼ全て p_3 を表すのに相応しいクエリであるように感じられる．ただ，「秋保 温泉 日帰り」「秋保 日帰り 温泉」の両クエリは秋保温泉への日帰り旅行に関する情報を目的としたクエリであるとみなすと，p_3 を検索するための典型的クエリとは言い難い．また，「秋保 温泉 高級 旅館」は秋保温泉の旅

(注8)：http://ja.wikipedia.org/秋保温泉

館に関する情報を求めたクエリで，これも p_3 を表す典型的クエリとは言い難い．また，「秋保 温泉 観光 協会」は秋保温泉の観光協会のページを navigational search によって検索するためのクエリであると思われるが，これはただ単に p_3 内にその全ての形態素が含まれてしまったがためのノイズであると思われる．

表 4 に含まれる全てのクエリは，「秋保」「温泉」という 2 語を含むことがわかった．しかし，本提案手法における典型的クエリの評価 (PageView(p,q)) では，これらの AND を取った「秋保 温泉」が最も典型的なクエリとは判断されない．これは，PageView がコンテンツ利用者の検索ニーズを適切に表しているとするならば，p_3 が秋保温泉に対するコンテンツ利用者の限定されたニーズのみに応えているという結論に至る．実際に，p_3 から秋保温泉の歴史や地理に関する情報を得ることはできても，秋保温泉旅行のツアーに関する情報を得ることはできない．

表 5 からは，p_1，p_2，p_3，p_4，p_5,p_{10} では比較的うまくそれらのページに内在するコンテンツ利用者のニーズを表した「典型的なクエリ」を発見できたが，p_6，p_7，p_8，p_9 ではうまく発見できていないことがわかる．これは，これらのページ (p_6，p_7，p_8，p_9) に十分な権威 (被閲覧数や被リンク数等) がないため，いずれのクエリでもそれらが上位にランクされることがなく，そのために検索結果のノイズとして p_6，p_7，p_8，p_9 が含まれたときのクエリが選ばれてしまったためであると考えられる．また，計測する順位の値も 100 位までに限定してしまったために，100 位以内に入るだけの権威のあるページでなければ本提案手法により典型的なクエリを求めるのは不可能である．したがって，本研究で提案した手法では十分権威のあるページには適用できるが，そうでない (作成されて間もないために検索エンジンにインデクスされていなかったり，上位にランクされることなのないような) ページには適用が困難であると言える．

6. まとめと今後の課題

本研究では，ページ p に内在するインターネットユーザのニーズを抽出するために，それを検索するための典型的クエリ q を発見する手法を提案した．具体的には，クエリ q の利用頻度 $\mathrm{qf}(q)$ と，クエリ q で検索したときのページ p の検索結果順位 $\mathrm{rank}(p,q)$ から発見する．

予備実験を行ったところ，ページ p が作成されて十分に時間が経過しておらずまだ検索エンジンにインデクスされていない，もしくは十分に権威 (被リンク数や被閲覧数等) を得られていない状況では，$\mathrm{rank}(p,q)$ の正確な値を得ることができず (上位 100 位以内に出現せず)，そのために典型的なクエリを発見できない．

また，今回の実験の結果以外にも実際には利用されていないが，ユーザのニーズを表すという意味でより典型的なクエリの存在が考えられる．なぜなら，ユーザが実際に利用するクエリに含まれるキーワードの数は多くても 5 個程度であり，それ以上の数のキーワードを含むクエリは滅多に実行しないからである．

これらを踏まえ，以下のような方向性でさらに研究を進めていく予定である．

（1） $\mathrm{rank}(p,q)$ が得られなかった場合の予測値 $\mathrm{rank}'(p,q)$ の計算手法を提案する．ただし，上位 100 件については既に調べページ p が含まれないことは確認できているはずなので，少なくとも $\mathrm{rank}'(p,q) \geqq 101$ が成立する．

（2） 検索エンジンにインデクスされているようなページを対象に，本提案手法の評価実験をリファラ情報と比較することにより行う．

（3） クエリログ参照により得られるクエリ候補は，実際にユーザが実行したクエリであるためにキーワードの数が 5 個を越えるものはほとんどない．しかし，今回の提案手法により得られたクエリ候補に，さらにキーワードを追加することで得られる別のクエリ候補も想定されるはずである．そのようなキーワードの数の多いようなものを見つけるために，どれが追加すべきキーワードなのかを評価する手法を提案する．

（4） 今回提案したクエリログ参照の手法は，幅優先探索である．各クエリ候補に対し評価関数は計算できているので，典型的クエリ発見のためのより効率的なクエリログ参照の手法を提案する．

謝辞 本研究の一部は，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)，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」における計画研究「情報爆発時代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，A01-00-02，課題番号 18049041)，および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」における計画研究「情報爆発に対応する新 IT 基盤研究支援プラットフォームの構築」(研究代表者：安達淳，Y00-01，課題番号：18049073) 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] 是津, 木俵, 田中：“Web ページのアスペクトの発見”.

[2] S. Nakamura, S. Konishi, A. Jatowt, H. Ohshima, H. Kondo, T. T., O. S. and T. K.: “Trustworthiness Analysis of Web Search Results”, Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL2007) (2007).

[3] Y. Ohsawa, N. Benson and M. Yachida: “KeyGraph: automatic indexing by co-occurrence graph based onbuilding construction metaphor”, Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on, pp. 12–18 (1998).

[4] K. Zettsu, Y. Kidawara and K. Tanaka: “Discovering aspect-based correlation of Web contents for cross-media information retrieval”, Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on, **2**, (2004).

[5] Y. Matsuike, S. Oyama and K. Tanaka: “Approximate intensional representation of web search results”, Lecture notes in computer science, pp. 607–608 (2005).

[6] A. Broder: “A taxonomy of web search”, ACM SIGIR Forum, **36**, 2, pp. 3–10 (2002).

[7] A.L., R. Albert：“Emergence of Scaling in Random Networks”, Science, **286**, 5439, p. 509 (1999).

[8] G. Salton: “Automatic Information Organization and Retrieval.”, McGraw Hill Text (1968).