

信頼度の高いタイトル情報を利用した固有ページ発見

田中 幸一[†], 高須 淳宏[‡], 安達 淳[‡]

[†] 東京大学大学院情報理工学系研究科

[‡] 国立情報学研究所

{ktanka,takasu,adachi}@nii.ac.jp

概要

一般的な情報検索とは異なり、ユーザが既知である対象物の代表的なページの検索を目的とする既知事項検索が近年行われている。既知事項検索では、一般的な情報検索でよく用いられているコンテンツ情報よりも、タイトルやアンカーテキスト、URL などページの名前に関連する情報の利用が期待されている。しかし、それぞれの情報には名前に関連のないものが含まれていることもしばしばあり、有効と思われる情報の選択が必要である。本研究では、複数の情報に出現するものを信頼度の高い情報とみなし、それを用いて固有ページの発見を行う。本提案手法について、NTCIR で作成されたトピックと正解を利用した評価を行い、非常に高い適合性を得たことを報告する。

Finding named pages utilizing reliable title information

Kouichi Tanaka[†], Atsuhiro Takasu[‡], Jun Adachi[‡]

[†] Graduate School of Information Science and Technology

[‡] National Institute of Informatics, The University of Tokyo

{ktanaka,takasu,adachi}@nii.ac.jp

abstract

In recent years, differing from Ad hoc search, a retrieval task aiming at discovering the target page is come to performed. Compared with the former, the latter task is called Known item search. In Known item search, very high precision is required, so not text information which can obtaine high recall but the new information such as URL, title, or anchor text information can be considered. In this article, we focussed on name attachment of a page, and proposed a new method extracting reliable title information from such information. Experiment result showed higher precison than a method using solely information.

1 はじめに

一般的な情報検索とは異なり、あるトピックの代表的なページの発見を目的とする検索が近年行われている。このような検索は、既知事項検索と呼ばれている。前者の例として「zip ファイルの解凍方法が知りたい」という検索要求が挙げられる。この場合、システムへの入力文（本稿ではクエリ

と呼ぶ）を「zip 解凍」など検索要求の一部の語を用いることにより、解凍ツールのダウンロード先やインストール方法などを記述したページの収集を行う。しかし、検索結果の中からユーザ自ら検索要求に適合するページを見つけていかななくてはならないため、クエリの語の選択を注意深く行わなければならない。検索結果に多く含まれてし

まい、ユーザに大きな負担をかける恐れがある。

一方後者の例としては、「日立のホームページを見つけない」という検索要求が挙げられる。この検索では、ユーザは日立という会社については既知であるものの、そのホームページの場所がわからないため、「日立」という名前をクエリに入力することにより目的とするページの発見を行う。一般的な情報検索とは異なり、ユーザはどのページが自分の要求に適合するページなのか容易に判断することができると同時に、適合ページの数はいくつの場合1つなどの限られた数である。そのため、再現率は重視されず極めて高い適合率が要求される。

以上をまとめると、この2つの検索は以下の点で大きく異なっている。

1. 要求する情報の質の違い

一般的な情報検索では、ユーザが要求する情報を表すという抽象的なページの収集が目的であるのに対し、既知事項検索の目的はトピックの代表的なページという極めて具体的なページの発見である。

2. 適合となるページの基準の違い

一般的な情報検索において適合と判断されるページは、要求される情報を含んでいるかページの内容でもって判断される。一方既知事項検索の場合では、ページの内容よりもリンク構造の位置関係などが重視されることが多い。

この既知事項検索を扱うワークショップとして、TREC(Text REtrieval Conference)¹やNTCIR(NII-NACSIS Test Collection for IR Systems)²などがある。

TRECでは、TREC2001においてWeb trackの一環としてhomepage finding taskと呼ばれるタスクが行われた。このタスクは、与えられたトピックが記述されたホームページ（もしくはサイトエンターページ）の発見を目的としている。一般的な情報検索とは異なり、コンテンツ情報のみを用いたグループは結果が伸び悩む一方で、URLのタイプの解析を行ったグループが高い結果を得ている[1]。続くTREC2002では、named page finding taskという若干変更した形で既知事項検索が続行された。このタスクでは、ホームページではなく

クエリが名前となるような固有のページの発見が目的とされた。前年のURLのタイプの利用は結果の向上につながらず、コンテンツ情報の結果にアンカーテキストやWebのリンク構造の情報の結果を追加することにより精度が増したという報告が、いくつかのグループによってなされている[2]。そしてTREC2003では、前2年のタスクを混合したhome/named page finding taskとして既知事項検索が行われた。このタスクのクエリは、ホームページのものと固有ページのものの2種類あり、より難易度の高いものになっている。いくつかのグループがクエリのタイプを分類して、それぞれのタイプに適した手法を採用したが、大きな成果は得られなかった[3]。

一方NTCIRでは、NTCIR-4 WEB taskの中でNavigational Retrieval Task 1という名前で既知事項検索が行われた。このタスクでは、既知の対象物（人物、店、施設など）をクエリとし、それらの代表的なページの発見が目的とされている。また、名称ではなく手がかりとなる属性情報や関連情報を用いて検索を行う場合も考えられ、適合するページが1つではなく複数存在し得る点でTRECのhome/named page finding taskとは若干異なっている。結果として、アンカーテキストとリンク情報を用いたグループが最も良く、アンカーテキストを使わずリンク情報のみを使用したものが2番目に、コンテンツ情報のみを扱ったグループが最も悪い結果となっている[4]。

そこで本研究では、既知事項検索で有用とされているアンカーテキスト、タイトル、URLのページに名前に関連する複数の情報から、信頼度の高い情報を抽出して検索を行う手法を提案する。これにより、それぞれの情報を単独で用いた場合と比べて、より高い適合率を実現する。

以下、2章でページのタイトル情報の性質や利用手法について説明し、3章で本研究の提案手法について述べる。4章で評価及び考察を行い、5章でまとめる。

2 タイトル情報

2.1 タイトル

タイトルは著者によるページへの名前付けである。つまり、Jinら[5]によって述べられているように、要求情報に対して要約の役割を持つクエリと

¹<http://trec.nist.gov/>

²<http://research.nii.ac.jp>

極めて類似した性質を持っている。そのため、クエリにマッチングするタイトルが付けられたページはトピックとの関連が強いと言うことができ、既知事項検索では有用な情報の1つと考えられている。

2.2 アンカーテキスト

アンカーテキストは、当該ページへリンクを張る者による名前付けと言う事ができるので、既知事項検索ではタイトルと同様に大きな役割を持っている。まず初めに、アンカーテキストの定義を行う。本稿では、HTML文書の中でハイパーリンクのために表示される”強調されたクリック可能な文字列”，すなわち<A>タグの境界の中で現れる文字列をアンカーテキストと定義する。例えば以下のような形のタグの場合、

```
<A HREF="http://foo.com/">buy furniture</A>
```

アンカーテキストは”buy furniture”であり、それはURL ”http://foo.com/”に存在するページに関連付けられたものである。このアンカーテキストは単語の索引(インデックス)として、テキストと同様に用いられる。アンカーテキスト情報を扱う上で問題となるのが”戻る”や”こちら”などの誘導的な役割を持つアンカーテキストの存在である。これらは、ページの内容とは無関係であるので、一般的な情報検索における”stop word”と同様に除去しなければならない。

Malawong ら [6] は、データマイニングの分野で用いられている frequent item sets によってアンカーテキストで頻出する語の組み合わせをページの索引として用いている。これにより、多数の人物に参照されている部分を抽出することができるが、前述の誘導的な役割を持つアンカーテキストの問題に対処していない。

Kawai ら [7] はサイトのトップページに張られているアンカーテキストをサイトアンカーテキストと定義し、このアンカーテキストのみをページの名前として用いている。一般に、誘導的な役割を持つアンカーテキストは同一サイト内のリンクに含まれているため、外部のサイトからのリンクを用いることにより自動的に除去されている。

2.3 URL

URL にはそのサイトの機関の名称や頭文字が含まれていることが多い。例えば、東京大学のトッ

プページの URL は”http://www.u-tokyo.ac.jp/”であり、国立情報学研究所のトップページの URL ”http://www.nii.ac.jp/”にはその略称の”NII”という文字が含まれている。従って、クエリとのマッチングを行うことにより、関連するページを見つけることができる。また、URL は階層構造をなしているため、root となるエントリーページの URL の長さは一般的に短く、代表的なページの選択にこれを用いる研究も多い。

Craswell ら [8] は、URL のタイプを分類し、それぞれのタイプにおいてエントリーページである確率をテストデータから求め、それを entry page finding task に利用することで高い精度が得られたと報告している。

3 提案手法

3.1 目的

本手法の目的は、タイトルやアンカーテキスト、URL などの複数の情報に出現するものをそのページにおける信頼度の高いタイトル情報とみなし、クエリの固有ページ発見に利用するということである。まず本研究で扱う実験データについて説明する。本実験では、NTCIR-4 Web で行われた Navigational Retrieval Task 1 を用いた。このタスクでは、トピックに対して検索要求を満たす代表的なページである適合ページと、そのトピックを含むサイトのトップページなどが該当する部分適合ページとの2段階の判定が行われている。実験に用いたデータセットは NTCIR-3 で作成された NW100G-01 というデータセットで、これは jp ドメインから収集されたおよそ1千万の Web ページで構成されており、その大きさは100GB ほどである。

本手法では、それらの Web ページのタイトルと URL のみを用いている。また、データセット内のページに対して絶対パスにより関連付けられているアンカーテキストを予め抽出している。なお、このアンカーテキスト集合の大きさは700MB ほどであり、元のテキスト集合のおよそ100分の1である。また、本実験では同一ページへ張られているアンカーテキストをまとめてそのページに対する1つの仮想的なテキストとして扱っている。質問語はトピック製作者により選択された1~3個の単語を使用した。予め優先度の高い順に並べられているため、 i 番目の質問語の重要度を以下のように定

義した.

$$kw_i = 10^{n_q - i} \quad (1)$$

ここで, n_q は質問語の数を表している.

3.2 複数の情報の共起によるスコア付け

本研究では複数の情報に出現する語をそのページの名前として信頼度の高い情報とみなし, それらに高い重要度を与える手法を提案する. 具体的な手順の流れは以下になる.

1. アンカーテキストのスコアの計算
2. 単語のスコアの計算
3. ページのスコアの計算

1のアンカーテキストのスコアの計算では, まずタイトルとアンカーテキストの間の共起を見る. そして以下の2つの点を元にスコアを計算する.

● 2つの情報間の共起数

アンカーテキストとタイトルで共通に現れる語が多ければ多いほど, そのページに対して複数の人間によってより細かいトピックへ分類されていることを表しており, ページの名前として信頼度の高いものであると言う事ができる.

● それぞれの情報の長さ

上のように共起数は重要な要素ではあるが, タイトルとアンカーテキストいずれの文字列も長ければ長いほどより多くの語が共通に出現しやすくなる. 即ち, 1つの単語が持つ影響度が相対的に小さくなってしまふ. そこで, タイトルとアンカーテキストそれぞれの長さをを用いて正規化を行う必要がある.

同様に URL との共起も考慮して, アンカーテキストのスコアを以下のように計算する.

$$Score(D_j) = \frac{Co(D_j, T_p) \times 1.0 + Co(D_j, URL_p) \times 0.5 + Co(T_p, URL_p) \times 0.7}{\sqrt{|D_j|} \times \sqrt{|T_p|}} \quad (2)$$

ここで, T_p はページ p のタイトル, D_j はページ p にリンクしているアンカーテキスト, URL_p はページ p の URL を表している. $Co(a, b)$ は a, b 両者に共通に出現する単語数を表し, $|D_j|, |T_p|$ はそれぞれの長さとする. URL との共起係数は呼び実験にて最もよい結果を示した値を用いている. また1つ

の語については最も係数が高い共起を見ることにしているため, (2) 式はアンカーテキストとタイトルが完全に一致する場合最大値 1 をとる.

次に単語のスコアについて説明する. 単語のスコアは統計的なスコアとアンカーテキストのスコアを元に以下のように計算した.

$$Score(w_i, D_j) = Stat(w_i) \times (1 + Score(D_j)) \quad (3)$$

統計的スコア $Stat(w_i)$ は, 本稿では単語の重要度の計算式としてよく用いられている Okapi BM25 を用いている [9].

$$Stat(w_i) = tf_d \times \frac{\log\left(\frac{N - n + 0.5}{n + 0.5}\right)}{1.25 \times (0.25 + 0.75 \times \frac{dl}{avdl}) + tf_d} \quad (4)$$

ここで, tf_d は仮想テキスト内における語 w_i の出現頻度, N は全ページ数, n は仮想テキスト内で語 w_i が少なくとも 1 回以上現れるページ数, $dl, avdl$ はそれぞれ文書長, 平均文書長を表している.

最後にページのスコアを計算する. ページのスコアは仮想テキスト内の全ての質問語のスコアの総和として以下のように求める.

$$Score(p) = \sum_{t \in q} \sum_{D_j \rightarrow p} Score(t, D_j) \quad (5)$$

ページの順位付けには, このスコアと質問語の重要度を用いて行っている.

3.3 他のアンカーテキストのスコアの伝播

データセットの中には, 日本語のページにもかかわらず, タイトルが英語であるページが多く存在する. 例えば日立のトップページである "http://www.hitachi.co.jp/" では「HITACHI: ホーム」というタイトルが付けられている. 一方, このページへリンクしているアンカーテキストの多くは「日立」となっている. 本手法の場合, この2つの情報において共起している語がないため, このページに対して「日立」という名前はふさわしくないと判断することになってしまう. この問題を解決するために, 他のアンカーテキストのスコアを伝播させる手法を提案する. もし, このページに対して「HITACHI」というアンカーテキストが付けられている場合, 一致するアンカーテキストの存在により著者によって付けられた「日立」は, このページの名前として信頼度の高いものであると言うことができる. そこで, 全てのアンカーテ

キストに対して、重みを上方修正する。新たなスコアは元のスコアとそのページへのアンカーテキストの中で最大となるスコアとして以下のように計算する。

$$Score_{new}(D_j) = \frac{1}{2} (Score_{old}(D_j) + MAXSCORE(p)) \quad (6)$$

これにより、日立と無関係のページに「日立」というアンカーテキストが存在する場合、「日立」の固有ページとしてこのページの順位が上昇してしまう恐れがある。この影響の有無について、実験結果をもとに考察する。

4 実験結果

4.1 ベースライン

本手法の有効性を示すための比較対象として、以下の2つのベースラインの手法を設定する。

コンテンツ情報にのみを用いたベースライン

ページ内のテキストに対して形態素解析器 MeCab[10] により単語毎に分解し、全ての単語について Okapi BM25 で統計的なスコアを与える。ページ内の質問語のスコアの総和をページのスコアとし、質問語の重要度と併せてページの順位付けを行っている (BLC)。

アンカーテキスト情報のみを用いたベースライン

アンカーテキストに現れる単語についてコンテンツ情報の場合と同様に Okapi BM25 でスコアを与える。そしてページに対する仮想テキスト内の質問語のスコアの総和をそのページのスコアとし、質問語の重要度と併せてページの順位付けを行っている (BLA)。

4.2 評価手法

本実験での評価手法として、Navigational Retrieval Task 1 で用いられた WRR (Weighted Reciprocal Rank) と DCG (Discounted Cumulative Gain) をそのまま用いた。WRR は検索トピックごとに最初に出現した正解の順位の逆数を求め、それらを全トピックにわたり平均したものである。一方、DCG は決められた順位 (本実験では 10 位) 以内に出現する正解の順位についての累積スコアの平均を表している。どちらも詳細については [11] を参照されたい。本稿では、適合ページのみを正解とする

表 1: それぞれの手法の WRR, DCG の結果

手法	WRR1-0	WRR1-1	DCG3-0	DCG3-2
BLC	0.1132	0.1990	0.5859	1.0062
BLA	0.5565	0.6456	2.1744	2.6467
提案手法 他アンカーなし	0.5852	0.6555	2.2037	2.7073
提案手法 他アンカーあり	0.5953	0.6670	2.2698	2.7703

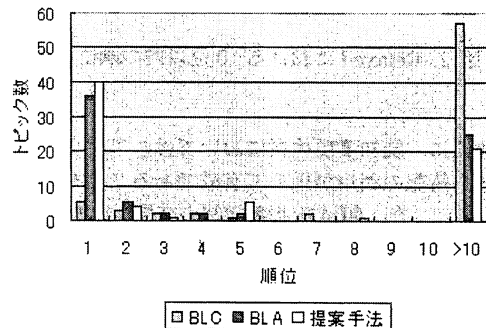


図 1: Rigid における 10 以内の順位の様子

もの (Rigid) を WRR1-0 及び DCG3-0、適合ページと部分適合ページを正解とするもの (Relaxed) を WRR1-1 及び DCG3-2 と定義している。

4.3 実験結果及び考察

2つのベースライン及び提案手法の結果を表 1, 10 位以内の順位の様子を図 1, 2 に示す。

表 1 より、アンカーテキスト情報の利用によって WRR, DCG いずれも大きく上昇していることがわかる。図 1, 2 で示されている実際の順位においても、BLC では 8 割以上のトピックが 10 位以上であり、10 位以内に入ったトピックの間で大きな特徴は見られなかった。BLC では、質問語を多く含んでいるページほど上位に順位付けされることになるが、既知事項検索ではトピックの代表的なページであることが重要であり、そのようなページは必ずしも多くの質問語を含んでいるとは限らないことが要因の 1 つであると考えられる。一方、BLA と提案手法では 10 位以内に順位付けされているトピックの数が大幅に増加した上、大部分のトピックが 1 位に順位付けされている。これは高い適合率が要求される既知事項検索の要件を満たしている。

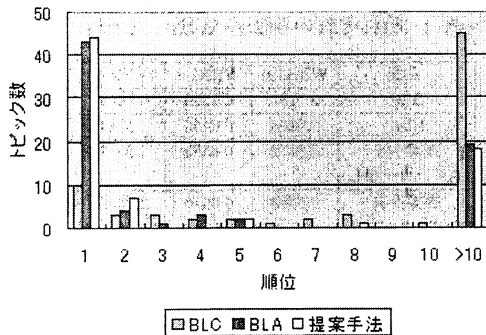


図 2: Relaxed における 10 位以内の順位の様子

ゆえに、既知事項検索においては、アンカーテキスト情報の利用が極めて有効であることを示している。また、BLA と提案手法の結果を比較すると、提案手法の方がより上位に順位付けされている。図 1 では 5 位、図 2 では 2 位に順位付けされたトピックの数が増えているが、これは BLA で 10 位以上に順位付けされたトピックによる影響である。次に 3.3 節で提案した他のアンカーテキストを伝播させたときの結果について考察する。表 1 を見ると、これを用いることにより WRR, DCG いずれも上昇している。これは、前述した「日立」というアンカーテキストが、日立とは無関係のページよりも日立のトップページに対して付けられているものが圧倒的に多い。そのため、前者の不適合であるページの順位よりも適合したページの順位を上げる働きが強く、結果として精度が向上することになったと考えられる。

4.4 コンテンツ情報とアンカーテキスト情報の併用

前節で述べたように、アンカーテキスト情報を用いることによりトピックの絞込みが十分に行われ適合率の高い結果を得ることができた。図 1, 2 より上位 10 位に入らなかったトピックの数は BLA, 提案手法どちらも 20 個ほど存在したが、このうちほぼ半数のトピックについては適切なアンカーテキストを抽出できずに順位付けを行うことができなかったものである。一方、BLC では質問語を含むページは必ず存在しており、全てのトピックについて順位付けを行うことができた。そこで 2 つの

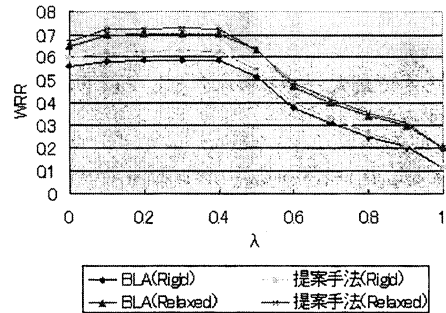


図 3: コンテンツ情報とアンカーテキスト情報を併用した結果

情報の長所を利用するために、各々の結果を混合させることを考える。混合の方法としては、各々のスコアに隔たりがあること、評価手法が上位 10 位以内の正解の順位を元に求められていることを考えて、それぞれの結果の順位を用いることにした。まず、あまりに低い順位のページは無視するために、それぞれ上位 30 位まで順位付けを行い、それぞれの順位の逆数をパラメータ λ で以下のように線形結合させた。

$$rank_{combi} = \frac{1}{\lambda \times \frac{1}{rank_{content}} + (1 - \lambda) \times \frac{1}{rank_{anchor}}} \quad (7)$$

コンテンツ情報の結果は BLC を用いて、アンカーテキスト情報の結果は BLA と提案手法の 2 通りの結果を用いて実験を行った。そのときの WRR の結果を図 3 に示す。

図 3 より、2 つの情報を併用することによって、元々の結果よりも上回っていることがわかる。 $\lambda = 0.3$ 即ちコンテンツ情報とアンカーテキスト情報の利用比が 3:7 のときに最も高い値を得ている。DCG についても同じ値のときに最大となった。

最後に、一般的な情報検索システムの結果と既知事項検索システムの結果について比較を行う。表 2 に提案手法による実験結果と Google³ による実験結果を示す。ここでは上位 10 位以内で最初に出現した適合ページの順位を表しており、10 以上のものについては × としている。提案手法が上回ったトピックでは、施設の地図情報や製品の説明ページなどユーザが知りたい情報を含むページが公式のトップページよりも上位に来ている。即ち、この

³<http://www.google.co.jp/>

2つの検索の目的は大きく異なるため、それぞれの目的に合ったシステムが必要であると言える。

5 まとめ

本稿では、既知事項検索において、複数のタイトル情報で共通に出現する情報を信頼度の高いタイトル情報とみなし、それを用いて固有ページの発見を行う手法を提案した。また、アンカーテキストとタイトルが異なる言語で書かれている場合にも対処するために、他のアンカーテキストによりタイトルの信頼度を高めてアンカーテキストのスコアの再計算を行った。結果として、ベースラインと比較してWRR1-0が最も大きく上昇し、高い適合率が要求される機知事項検索において非常に有効であると言えることができた。

今後の課題としては1つ目に単語のマッチング方法の改良が挙げられる。本手法では一致するかしないかで判断しているが、表記の揺れや類似語などに対処することができない。タイトル情報をより効果的に利用するためには、シソーラスを用いてより柔軟なマッチングを行うことが必要である。2つ目として、タイトルの拡張が挙げられる。ページによっては著者によるタイトルが存在しないものもあるため、ページ内の見出し語などテキストの一部をタイトルに追加することにより、適合性を落とさずに再現性を高めることができるのではないかと考えている。本実験で用いたトピックはほぼ日本語であるため、全て英数字であるURLとの共起関係はほとんど利用されなかった。しかし、日本語ドメイン化が進むにつれて、URLの効果的な利用が精度を上げる上で大きなウェートを占めてくるのではないかと考えている。

参考文献

- [1] D. Hawking, and N. Craswell: Overview of the TREC-2001 Web Track, Proceedings of TREC2001, 2001.
- [2] N. Craswell, and D. Hawking: Overview of the TREC-2002 Web Track, Proceedings of TREC2002, 2002.
- [3] N. Craswell, and D. Hawking: R. Wilkinson, and M. Wu: Overview of the TREC2003 Web Track, Proceedings of TREC2003, 2003.
- [4] K. Oyama, K. Eguchi, H. Ishikawa, and A. Aizawa: Overview of the NTCIR-4 WEB Navigational Retrieval Task 1. Working Notes of NTCIR-4, Tokyo, 2-4, June, 2004.
- [5] Rong Jin, Alex G. Hauptmann, and ChengXiang Zhai: Title language model for information retrieval. Proc. of the 25th annual international ACM SIGIR conference on research and development in information retrieval, pages 42-48, Tampere, Finland, August 2002.
- [6] J. Malawong, and A. Rungsawang: Finding Named Pages via Frequent Anchor Description. Proceedings of Trec-11, 2002.
- [7] H. Kawai, K. Tateishi, and T. Fukushima: Navigation Retrieval with Site Anchor Text: Working Notes of NTCIR-4, Tokyo, 2-4, 2004.
- [8] N. Craswell, D. Hawking, and S. E. Robertson: Effective Site Finding using Link Anchor Information. Proc. of the 24th annual international ACM SIGIR conference on research and development in information retrieval, pages 250-257, New Orleans, Louisiana, USA, September 2001.
- [9] S. E. Robertson, and S. Walker: Okapi/keenbow at TREC-8. Proceedings of the Eighth Text REtrieval Conference(TREC-8), NIST Special Publication 500-246, pp.151-161, 2000.
- [10] MeCab. <http://chasen.org/~taku/software/mecab/>
- [11] K. Eguchi, K. Oyama, E. Ishida, N. Kando, and K. Kuriyama: Overview of the Web Retrieval Tasks at the Third NTCIR Workshop, NII Technical Report, No.NII-2003-002E, Jan., 2003.

表 2: 提案手法と Google による各トピックにおける適合ページの検索順位

トピック ID	質問語	適合ページの URL	提案手法の順位	Google の順位
007	マスタートードシード 代理店	http://www.mustardseed.co.jp/	1	1
012	JPNIC	http://www.nic.ad.jp/	1	1
014	ギガバイト マザーボード	http://www.gigabyte.co.jp/	1	1
015	ゼンリン	http://www.zenrin.co.jp/	1	1
018	Becky メールソフト	http://www.rimarts.co.jp/becky-j.html	1	4
039	全日本プラスチック日用品フェア	http://www.jpmm.or.jp/jpf/index.html	1	2
041	アエロフロート	http://www.aeroflot.gr.jp/	1	1
049	SUICA JR	http://www.jreast.co.jp/suica/index.html	1	1
051	かっぱ寿司	http://www.kappa-creat.co.jp/	1	1
053	モスバーガー メニュー	http://www.mos.co.jp/menu/index.html	1	1
062	可視化情報学会	http://www.vsj.or.jp/	1	1
063	国立天文台 観望会	http://www.nao.ac.jp/pio/kanboukai.html	1	1
072	東京大学経済学部	http://www.e.u-tokyo.ac.jp/index-j.html	1	1
076	東京商船大学	http://www.tosho-u.ac.jp/	1	1
077	フィンランド	http://www.finland.or.jp/	1	4
079	シーメンス	http://www.siemens.co.jp/	1	1
081	株式会社データ通信システム	http://www.dts.co.jp/	1	1
088	角藤 九州工学部 近畿大学	http://www.fsci.fuk.kindai.ac.jp/kakuto/arira.html	3	×
089	瀬崎薫	http://www.mcl.iis.u-tokyo.ac.jp/	1	1
090	小林市	http://www.city.kobayashi.miyazaki.jp/	1	1
091	国土地理院	http://gsi.go.jp/	1	1
092	遺伝子組み換え食品	http://fsic.co.jp/bio/index.html	2	2
094	国立オリンピック記念青少年総合センター	http://www.nyc.go.jp/	1	1
101	川崎市役所	http://www.city.kawasaki.jp/topmenu.htm	×	1
102	JA 長野県農業情報ネットワーク	http://www.janis.or.jp/agrinet/	1	1
104	世界銀行	http://www.worldbank.or.jp/	1	1
105	新生銀行 口座開設	http://www.shinseibank.co.jp/	1	1
115	上野公園	http://www.tokyo-park.or.jp/park/05/05-park.htm	2	3
116	代々木公園	http://www.tokyo-park.or.jp/park/67/67-park.htm	×	1
117	青年海外協力隊	http://www.joca.or.jp/	5	3
124	an	http://www.linos.co.jp/ans/	×	1
125	ぐるナビ	http://www.gnavi.co.jp/	1	1
127	国会会議録 検索	http://kokkai.ndl.go.jp/	1	1
131	秋葉 PC ホットライン	http://www.watch.impress.co.jp/akiba/	1	1
141	日本映画データベース	http://www.jmdb.ne.jp/	1	1
146	TOEFL	http://www.cieej.or.jp/	8	1
147	公認会計士 試験	http://fsa.go.jp/notice/noticej/kaikeishi-menu.html	5	3
149	TOEIC	http://www.toeic.or.jp/	1	1
150	銀行業務検定協会	http://www.kenteishiken.gr.jp/	×	1
154	靖国神社	http://www.yasukuni.or.jp/	1	1
157	浅草名所七福神	http://www.asakusashichihukujin.jp/	×	1
158	明治神宮	http://www.meijijingu.or.jp/	1	1
175	関東自動車学校	http://www.ktds.co.jp/	×	1
177	ブックオフ	http://www.bookoff.co.jp/	1	1
178	日経ビジネス	http://nb.nikkeibp.co.jp/	×	1
183	ダイソー	http://www.daiso-sangyo.co.jp/	2	2
191	永谷園 お茶漬け	http://www.nagatanien.co.jp/shouhin.index.html	×	5
198	神戸新聞	http://www.kobe-np.co.jp/	1	1
204	日本相撲協会	http://www.sumo.or.jp/	1	1
207	ゴルフ ルール	http://golf-gtpa.or.jp/rule.book/index.html	1	1
215	ラクロス	http://www.lacrosse.gr.jp/	1	1
230	小沢一郎	http://ozawa-ichiro.jp/	×	1
237	上田研 早稲田大学 理工学部	http://www.ueda.info.waseda.ac.jp/index-j.html	×	1
265	伊勢丹百貨店	http://www.isetan.co.jp/	1	2
269	安土城 博物館	http://www.azuchi-museum.or.jp/	1	1
274	日本未来科学館	http://www.miraikan.jst.go.jp/	2	1
278	花祭り 鬼	http://www.vill.tsutgo.aichi.jp/g.hanamatari.htm	×	3
279	百万石文化 加賀 金沢	http://www.city.kanazawa.ishikawa.jp/dentou/index.html	×	1
283	プラネタリウム	http://www.wmn.ne.jp/wmn-u/planeta/data1/index.html	×	1
284	コスモホール 稲美町	http://www.kakogawa.or.jp/inamitown/sisetu/library.htm	×	2
285	国立演芸場	http://www.ntj.jac.go.jp/	1	2
294	ホテル日航東京	http://www.hnt.co.jp/	1	1