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

2006－DBS－140 (I) (21)
2006／7／12

サーチエンジンのクエリログを利用した同位語・話題語の発見と可視化

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用いて，ある語の同位語や話題語を発見する手法を提案する．同位語とは共通の上位語を持つような語のことである．従来研究にも，同位語や，上位語，下位語を求めるようなものは数多くある．しかし，それらは大量のコーパスを用いるもの，または，膨大な計算処理が必要なものがほとんどであった．それに対して我々は，サーチエンジンのもつメタデータの一種であるクエリログのみを用いて，効率よく同位語を発見し，それらの話題となる語を発見する手法を提案する．ユーザが入力する検索クエリにおいて，ある語とその同位語となる語とはしばしば同様の語を用いて絞り込みをされることに着目し，効率よく同位語を発見する．また，得られた結果を視覚化し分かりやすく提示することも行う．

Discovering and Visualizing Coordinate Terms and Topic Terms Using Search Engine Query Log

Masashi YAMAGUCHI,† Hiroaki OHSHIMA ,†
Satotshi OYAMA † and Katsumi TANAKA †

We propose a method for searching coordinate terms and topic terms by using search engine query logs. " Coordinate terms" are terms which have the same hypernym. There are several researches that acquire coordinate terms, but they need parsed corpora or much time for computation. Our proposing method only needs a query log, which is one of a metadata of Web search engine, and relatively few times for computation to obtain coordinate terms and topic terms. We also visualize obtained coordinate terms and topic terms by using a graph layout algorithm.

1. はじめに

全ての語は他のいくつかの語とそれらとの関係を用いて説明することが出来る．語と語の関係を表す用語で有用なものに，上位語や下位語がある．上位語とはある語の上位概念を表す語であり逆に下位語はある語の下位概念を表す語である．また他にも同位語と呼ばれるものがあり，一般的に「ある語と共通の上位語を持つ語」とされる．本研究では，「同位語同士は共通の話題語を持つ」という考えのもと同位語の発見を行う．ただし，話題語とはある語の話題を表す語である．例えば「京都」の話題語としては「観光」,「地図」などがあり,「巨人」と「阪神」の共通の話題語としては「応援歌」などが挙げられる．

我々が提案する同位語と話題語の発見手法は，サーチエンジンのクエリログのみを用いたものである．クエリログとはユーザが問い合わせに用いた検索語の履歴であり，サーチエンジンは日々この履歴を蓄積している．クエリログの例を表1に示す．クエリログから，ユーザがある語に対してどのような語を追加する事が多いのかが統計的に把握する事が可能である．

クエリログにはユーザの検索目的が如実に反映されているものであり，たとえば「清水寺 拝観料」というクエリからは,「清水寺の拝観料はいくらか?」といったような検索目的を推測することが可能であり，また,「BMW ディーラー」というクエリからは,「BMW のディーラーの情報を得たい」といったような検索目的を推測することが出来る．ここで,「拝観料」は「清水寺」の話題であり,「ディーラー」は「BMW」の話題であると言える．逆に「清水寺 ディーラー」や「BMW 拝観料」というクエリで検索された回数が多くはないことから,「清水寺」の話題語として「ディーラー」は不適であり，同様に「BMW」の話題語として「拝観料」は不適であると言える.「ディーラー」を話題語として持つものに他に「ホンダ」「フェラーリ」があり，

† 京都大学大学院情報学研究科社会情報学専攻
Department of Social Informatics, Graduate School of Informatics, Kyoto University

これらは同位語であると言え，このような共通の話題語を多く持つ同位語を発見することが今回の我々の目的である．

2章でサーチエンジンメタデータについて述べ，3章で関連研究について言及した後，4章以降で同位語や話題語を効率的に求める手法を述べる．

2. Webサーチエンジンメタデータのクイックマイニング

上位語，下位語，同位語などを発見するための研究は，従来にも数多く存在している．しかし，それらのほとんどは，巨大なテキストコーパスや大量に収集したWebページを解析対象にしているものや，語の共起性や相互情報量などの計算などに多くの時間と計算量が必要なものである．

対して，我々の提案手法は，Webサーチエンジンのメタデータであるクエリログのみを用いるものであり，計算時間も比較的短いものであるため，それらの従来研究とは異なるものである．我々の目的の1つは，Webサーチエンジンが保持しているメタデータには非常に多くの有益な情報が眠っており，その解析により多くの役立つ知識を取得できることを明らかにすることである．

このような，Webサーチエンジンのメタデータのみを解析に用いて知識を取得する手法を，我々は，「Webサーチエンジンメタデータのクイックマイニング」と呼ぶ．Webサーチエンジンメタデータのクイックマイニングには，いくつかのメリットが存在する．まず，知識取得にかかるコストが非常に低いということが挙げられる．例えば本研究の解析対象であるクエリログは日々サーチエンジンが蓄積しているものである．また，適応分野が広いこともその他のメリットとして挙げられる．Web上にはさまざまな分野における文書が存在しており，それらを取得する目的でユーザは日々検索を行っていることから，クエリログの持つ分野は非常に多岐にわたるものである．また，新語が直ちに蓄積されることや，月別のクエリログを利用できることから，常に最新の語を取り扱うことが可能である．

なお本研究に用いたクエリログはOverture[1]が提供する「キーワードアドバイスツール」[2]と呼ばれるWebベースのツールを利用して取得した．ある語に関して問い合わせを行うと，その語を含んだ様々な組み合わせのクエリを前月の検索数と共に取得することが可能である．

3. 関連研究

同位語を発見する従来研究はいくつか存在している．Google Sets[3]というサービスは，複数の語を入力すると，それらの同位語と考えられる語を出力するものである．Google Setsのアルゴリズムは現時点で非公開であるが，Googleが収集した大量のWebページを解析し，同位語のクラスタを発見していると思われる．

同様に同位語発見に関する研究を述べる．Churchら[4]は，相互情報量を用いて，意味的に関連があるような語を発見する手法を提案した．正確には同位語発見を目的とした研究ではないが，発見される語には同位語も含まれている．他の研究のいくつかにおいても，この研究の成果である，相互情報量が高い語どうしは，同位語である可能性が高いことを利用している．Zoubinら[5]によるBayesian Setsはベイズ推定を利用した同位語のクラスタを発見するものである．用いられているアルゴリズムは非常にシンプルで高速であるが，何らかの大規模データを用意することが前提となっている．Linら[6]は類似の語に関するクラスタを生成する手法について提案した．係り受け関係を利用して語同士の類似度を計算することによって，語のクラスタを生成するものである．そのため，大規模コーパスが必要となる．Sinzatoら[7]はHTML文書から同位語を発見する手法を提案した．

また，サーチエンジンクエリログを対象とした従来研究としては，Webページの重要度判定や，クエリ補完に関するものがある．Cuiら[8]は，クエリとWebページの確率的な相関性を基に，適切な補完クエリを求める手法を提案した．Fonsecaら[9]は，クエリログから相関ルールを用いて，関連語を取得し，クエリ補完する手法を提案した．またWenら[10]はサーチエンジンによって得られたページのうちユーザが閲覧した文書を記録し，クエリと文書との関係性からクエリをクラスタリングする手法について提案した．Silversteinら[11]はクエリログのマイニングにより，カイ2乗値を基に語の出現数の偏りから，共起語を求める手法を提案した．いずれもクエリログを対象とした研究であるが，本研究の目的とは異なるものである．

4. 同位語・話題語の発見

4.1 同位語候補集合の取得

検索クエリはしばしばand検索を行う目的で複数の語がスペースで区切られ作成される．本節では，このような絞り込みに使われる語に着目して同位語の候補を取得する手順を述べる．

「トヨタ」を例にあげると「トヨタ カローラ」「トヨタ ディーラー」「兵庫 トヨタ」などのクエリが存在している．このとき，「□ カローラ」「□ ディーラー」「兵庫 □」などといったような，“絞り込みに用いられる語”と“絞り込みの方向”を表す型を絞込型と呼ぶことにすると，トヨタはこれらの絞込型にあてはまる語であると言える．

提案手法における候補語とはこれらの絞込型に当てはまる語を指す．実験においては，絞込型に含まれている語でクエリログを検索し得られた100件のうち，型に適合する物を候補語としている．

以下に「トヨタ」の同位語候補を得る手順を示す．

まず，「トヨタ」のクエリログを取得する (表1左)．検索回数1位の「トヨタ」は単独クエリであるため，候補語は取得できない．そこで，2位以下のクエリに注目すると，2位は「トヨタ 自動車」であるから，絞込型は「□ 自動車」である．そこで，「自動車」でクエリログを検索し「自動車」を含むクエリログを取得する (表1右)．そのうち，「□ 自動車」の絞込型に適応する物は「トヨタ」以外に「三菱」「日産」「ホンダ」「trend」などが挙げられる．同様に3位以下の「□ レンタカー」「□ 中古車」についても候補語を取得する．

「トヨタ」の場合，絞込型は99個得ることができ，それらに合致する候補語は重複を除いて，2335個得ることが出来た．

すなわち，一般化すると以下のようになる．

(1) ユーザがクエリ p を与える．
(2) サーチエンジンクエリログから p を含むクエリ集合 Q を取得する．
(3) Q のそれぞれの要素において，絞込型を取得する
(4) それぞれの絞込型に含まれている語を全て含むクエリ集合 R を取得し，絞込型に適合する語を候補とする

次節以降では，得られた候補集合を評価する手法として，「cos類似度」および「出現順位加重平均」を用いる方法と，「HITSアルゴリズム」用いる方法を述べる．

4.2 cos類似度

4.1節で述べたように，提案手法は絞込に用いられている語に着目して同位語発見を行うものである．以下，絞込に用いられている語を並列語☆と呼ぶこととする．例えば「トヨタ ディーラー」というクエリにおいて，「トヨタ」の並列語は「ディーラー」であり，逆も然りである．

☆ 並列語は絞込型に含まれている語であるとも言える．

表1 トヨタを含むクエリログ，及び自動車を含むクエリログ (検索回数上位20件のみ，2006年4月)

検索数	キーワード
660026	トヨタ
476901	トヨタ 自動車
143873	トヨタ レンタカー
75200	トヨタ 中古 車
49292	トヨタ レンタリース
31375	ネッツ トヨタ
26890	トヨタ ホーム
25241	トヨタ ネッツ
22739	トヨタ カローラ
20751	トヨタ bb
19092	トヨタ ディーラー
18912	トヨタ レクサス
16402	トヨタ 自動車 ホームページ
16014	トヨタ 紡織
15113	トヨタ クラウン
13597	トヨタ ファイナンス
11835	トヨタ 車体
11668	トヨタ カード
11334	トヨタ 博物館
11154	トヨタ アルファード

検索数	キーワード
2081735	自動車 趣味
513803	自動車 メーカー
476901	トヨタ 自動車
259138	二菱 自動車
252657	日産 自動車
193977	ホンダ 自動車
160289	自動車 保険
119458	自動車
108154	trend 自動車 保険
99555	自動車 試乗 レポート
95075	自動車 税
90130	スズキ 自動車
84869	マツダ 自動車
82947	自動車 教習 所
74659	スバル 自動車
63784	自動車 ドレスアップ
63134	メルセデス ベンツ 自動車
62431	bmw 自動車
59495	自動車 免許
53516	フォルクスワーゲン 自動車

本節においては，語の特徴量として並列語のベクトルを作成し，それらのcos類似度を元に同位語判定を行う手法を述べる．

まず，ある語 p を含むクエリを上位100件取得し，その集合を S_p とする．このとき p における語 a の検索数を加味した出現度 $tf_p^*(a)$ を以下のように定義する．

$$tf_p^*(a) = \sum_{S \in S_p} N(S,a) \tag{1}$$

ただし，

$$N(S,a) = \begin{cases} st(S) & (\text{if } a \in S) \\ 0 & (\text{otherwise}) \end{cases} \tag{2}$$

$$st(S) = (S \text{ の月間検索数}) \tag{3}$$

である．

これを用いて，語 p の特徴ベクトル $V(p)$ を以下のように定義する．

$$V(p) = (tf_p^*(a_1), tf_p^*(a_2), ..., tf_p^*(a_n)) \quad (4)$$

このとき，語 p と q の類似度を以下のように求める．

$$sim(p,q) = \frac{V(p) \cdot V(q)}{|V(p)||V(q)|} \quad (5)$$

このようにして求めた類似度に基づき候補語の評価を行った．結果は 5 節で示す．

4.3 クエリにおける語の出現順位

本節では，語の出現順位を語の特徴量として利用する事について述べる．

ユーザはサーチエンジンにクエリを与える際，絞込の目的で複数の語を組み合わせることが多いが，その順位からはユーザの意図が反映されているものである．例えば，「トヨタ ディーラー」というクエリの場合は，トヨタを 1 番目に指定し，次にディーラーを指定することで，トヨタのディーラーに関する情報を求めようとするユーザの意図が汲み取れる．しかし逆に「ディーラー トヨタ」の順でクエリを与えることは希である．実際 2006 年 4 月のデータによると，「トヨタ ディーラー」の検索数が 19092 件であるのに対し，「ディーラー トヨタ」の検索数はわずか 40 件しか無い．

その他の例を表 2 に示す．

表 2 出現順位による検索数の違い (2006 年 4 月)

検索数	キーワード
19092	トヨタ ディーラー
40	ディーラー トヨタ
3361	ギター 入門
30	入門 ギター
111504	京都 観光
1589	観光 京都
3197	サッカー チケット
474	チケット サッカー
33340	ルイヴィトン 財布
267	財布 ルイヴィトン

ここで，出現順位の傾向を表す値 $wao(p)$ を以下のように定義する．

$$wao(p) = \frac{\sum_{S \in S_p} (st(S) \cdot order(S,p))}{\sum_{S \in S_p} st(S)} \quad (6)$$

ただし

$$order(S,p) = (S \text{ における } p \text{ の出現順位}) \quad (7)$$

である．

$wao(p)$ は 1 以上の値を取り，p を含むクエリ上位 100 件において p が全て 1 番目に出現していれば $wao(p) = 1$ になり全て 2 番目に出現していれば $wao(p) = 2$ となる．

上記の「トヨタ」,「ディーラー」において，この値はそれぞれ

$$\begin{aligned} wao(\text{トヨタ}) &= 1.05104 \\ wao(\text{ディーラー}) &= 2.36592 \end{aligned} \quad (8)$$

となり，明らかに出現順位に偏りがあるといえる．

同様にいくつかの語において，$wao(p)$ の値を表 3 に示した．

表 3 wao の値の例

語 p	$wao(p)$	語 p	$wao(p)$
トヨタ	1.051047	レンタカー	1.589247
ディーラー	2.365919	大学	1.907977
ホンダ	1.026411	京都	1.045451
マツダ	1.061333	ホテル	1.923670
中古車	1.233393	観光	1.940983
情報	2.118529	ルイヴィトン	1.011967

本研究においては，

- $wao(p)$ の値が大きいほど，語 p は話題を表すことが多い．
- 同位語同士であれば $wao(p)$ の値は近い．

という仮説のもとに，同位語，話題語の発見を行う．

4.4 HITS アルゴリズム

候補語の評価に用いる別の方法として，HITS アルゴリズム (Hyperlink-Induced Topic Search) について述べる．HITS アルゴリズムは Kleinberg[12] が提案したアルゴリズムで，“authority”と“hub”という，Web ページの有用性を表す二つの尺度を用いていることが特徴的である．Web ページ p から Web ページ q へリンクされていることを，$p \to q$ と表すとすると，ページ p の“authority”と“hub”は以下のように定義されている．

$$auth(p) = \sum_{q, q \to p} hub(q) \quad (9)$$

$$hub(p) = \sum_{q, p \to q} auth(q) \quad (10)$$

このように，“authority”値は高い“hub”値を持つページからから多く参照されているとき高くなり，“hub”値は高い“authority”値を持つページを多く参照しているとき高くなるというように，再帰的な定義がなされている．

4.1 節で述べたとおり，候補語を取得するために絞込型を用いた．提案手法は HITS アルゴリズムをこの絞込型と得られた候補語との 2 部グラフに対して適用するものである．

「トヨタ」の場合の例を挙げる．

(1) 絞込型は 99 個得られ，それぞれから得られた候補語は重複を除いて，2335 個である．

(2) それぞれの絞込型からそれに合致する全ての候

補語にリンクを張り2部グラフを作成する．

(3) 作成した2部グラフに対し，HITSアルゴリズムを適用する．

グラフを図1に示す．例えば候補語である「ホンダ」は絞込型「□ ディーラー」や「□ 中古車」からリンクされている．また，「□ カローラ」からは「東京」「次期」「70」へリンクしている

Kleinbergは“authority”同士は存在を認めあおうとせず，“hub”によって間接的につながりを持つと想定している．同位語同士が同時にクエリに含まれるのは比較的少数であるといえることから，作成する2部グラフはKleingbergのモデルに合致しているといえ，HITSアルゴリズムを適用するのにふさわしいリンク構造であると言える．また，HITSアルゴリズムを使う他の利点としては，候補語を含むクエリログを取得する必要が無いことが挙げられる．

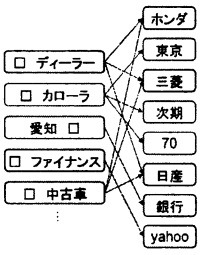

図1 絞込語と候補語の2部グラフ

本研究では，上位の“authority”値を持つ語を同位語，上位の“hub”値を持つ絞込型を話題語とする．

5. 実装システムにおける実験

5.1 「cos類似度」および「出現順位加重平均」を用いる手法

得られた候補語それぞれに対し，cos類似度を求め，出現順位加重平均 wao の差 δ が一定値以内の語を同位語と判定した．なお今回の実験では $\delta = 0.3$ を用いている．表に「トヨタ」と「ルイヴィトン」の結果を示す．なお，表中の*が付与された語は，$wao(\text{トヨタ}) = 1.05$ と $wao(\text{ルイヴィトン}) = 1.01$ それぞれとの $wao(p)$ の値の差が $\delta = 0.3$ 以上あるため，同位語ではないと判定したものである．cos類似度を計算する際には，対象となる語の並列語ベクトルを作成するために，今回の実験においてはWebアクセスが必要となる．実験一回あたり数千語の並列語ベクトルを作成する必要があるため，十分な量の実験は行えていない．cos類似度と出現順位加重平均について，ある程度の可能性は示すことが出来たと考えるが，有効性については今後議論していきたい．

5.2 「HITSアルゴリズム」を用いる手法

表5から表9にそれぞれトヨタ，ルイヴィトン，京都，ギター，サッカーについて，実験を行った結果を示す．なお表には“hub”値上位20件の絞込型と“authority”値上位20件の候補語のみ挙げている．

表8のギター場合，5番目に“authority”値が高い「ゴルフ」は「ギター」の同位語ではないため不正解である．しかし，「練習法」や「入門」，「レッスン」といった話題語を共通に持つ語である点で共通している．10番目の「パソコン」，12番目の「バイク」，18番目の「野球」も同様である．

表9のサッカーの例においては，候補語のうち上位の“authority”値を持つものは概ね正解であると言える．しかし，話題語となる絞込型は地方のスポーツ協会を表す物であり，情報検索の際に役立つ話題語であるとは言い難い．

本稿に掲載していない語についても，全く無関係の絞込型，候補語が上位に来る例があり，単純なHITSアルゴリズムの適用だけでは限界がある．

また，高い“authority”値を示した語には，同位語ではなく元の語の上位語にあたる語が含まれていることもわかる．表6のルイヴィトンの例においては，2番目に“authority”値が高いものとして「ブランド」が挙がっている．また，「トヨタ」には「車」，「サッカー」には「スポーツ」，「ギター」には「楽器」がそれぞれ高い“authority”値を示している．これらの例から，上位語も同様の絞込をされうることが分かる．

これらの問題に対処するための手法の改良については，今後の課題とする．

6. 同位語・話題語の可視化

得られた結果を可視化したものを図2示す．HITSアルゴリズムを適用した際に上位の候補語と絞込語を，それぞれ同位語，話題語として提示する物である．同位語それぞれが持つ話題語に対し，張力をもったリンクを張りレイアウトすることで，リンクされていない同位語と話題語は互いに遠い位置に配置される．

7. おわりに

本研究で利用したクエリログは，Overtureのキーワードアドバイスツールであるが，このクエリログは，厳密にはユーザの入力クエリを忠実に保存している物ではないと思われる．語の順序は正確に保存されてい

るものの，一部の語はスペースを間に挟まず，つなげて入力されたにもかかわらず切り分けられて保存されている．例えば，「京都大学」は「京都」と「大学」に分割されている．我々は，この問題に最大限対処したが，忠実に保存されているクエリログを用いることで，何らかの改善が期待できる．

また，HITS アルゴリズムの問題点として，単一のコミュニティしか発見できないことはよく知られており，従来から指摘されている．提案手法においてもクエリと関係の無い語の authority 値が高くなる現象がみられた．また，多義語の場合も同様に，単一の意味においての同位語が上位になった．今後はこのような問題の解決を目指す．

今回の実験において，サーチエンジンクエリログに有益な情報が眠っていることを示した．本稿では，同位語及び話題語の発見に焦点を当てたが，その他にも様々な有益な情報を得ることが出来ると思われる．今後はある語の同位語だけでなく，上位語，下位語についての発見手法も模索していきたいと考えている．

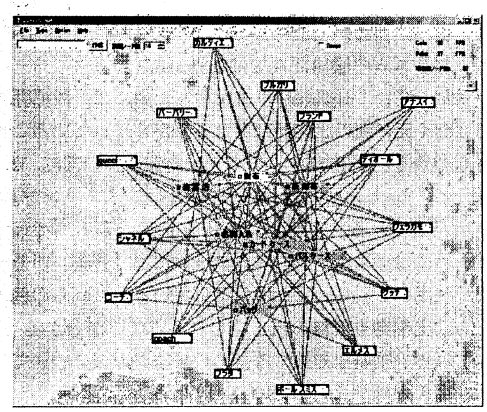

図 2　同位語・話題語の可視化

謝　　辞

本研究の一部は，文部科学省 21 世紀 COE 拠点形成プログラム「知識社会基盤構築のための情報学拠点形成」（リーダー：田中克己，平成 14〜18 年度），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，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）によるものです．ここに記して謝意を表すものとします．

表 4　「トヨタ」と「ルイヴィトン」それぞれの実験結果（cos 類似度と wao を用いた手法）

候補語 p	類似度	wao(p)
日産	0.9630	1.03
*ドレスアップ	0.9155	1.95
光岡	0.9072	1.05
マツダ	0.9071	1.06
いすゞ	0.9024	1.09
スバル	0.8763	1.11
フォルクスワーゲン	0.8745	1.02
ヒュンダイ	0.8503	1.01
amg	0.8435	1.15
*メーカー	0.8395	2.19
suv	0.8347	1.17
マセラッティ	0.8336	1.02
bmw	0.8252	1.06
アルピナ	0.8230	1.15
*メンテナンス	0.8160	2.05
アウディ	0.8067	1.02
ケーターハム	0.8014	1.01
*趣味	0.8002	2.00
gm	0.7939	1.32
*整備士	0.7931	2.01

候補語 p	類似度	wao(p)
gucci	0.9151	1.02
miumiu	0.8697	1.01
グッチ	0.8361	1.08
フェンディ	0.7787	1.03
miu	0.7778	1.14
ロエベ	0.7709	1.03
アナスイ	0.7654	1.01
シャネル	0.7533	1.03
プラダ	0.7509	1.02
クレージュ	0.7307	1.13
コーチ	0.7292	1.19
*クロコ	0.7284	1.54
フェリージ	0.7075	1.04
クレイサス	0.7000	1.03
dakota	0.6982	1.11
*コードバン	0.6889	1.39
*がま口	0.6808	1.45
furla	0.6804	1.02
フェラガモ	0.6733	1.08
オーストリッチ	0.6658	1.12

参 考 文 献

1) “Overture”. http://inventory.jp.overture.com/.
2) “キーワードアドバイスツール”. http://inventory.jp.overture.com/.
3) Google Sets (http://labs.google.com/sets).
4) Kenneth Ward Church and Patrick Hanks: “Word association norms, mutual information, and lexicography”, Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, pp. 76–83 (1998).
5) Z. Ghahramani and K. Heller: “Bayesian sets”, Proceedings of the Nineteenth Annual Conference on Neural Information Processing Systems (NIPS2005) (2005).

表 5 「トヨタ」の実験結果
(HITS アルゴリズムを用いた手法)

絞込型 p	$hub(p)$
□ ディーラー	0.06488
□ 専門店	0.06162
□ 純正 部品	0.04719
□ リコール	0.04261
□ 自動車	0.03917
□ 中古 車	0.03194
□ 部品	0.03018
□ 純正 ナビ	0.02962
□ 中古 車 販売 店	0.02714
□ レンタカー	0.02251
□ ファイナンス	0.02056
□ 自動車 ホームページ	0.02002
□ 部品 共販	0.01993
□ 自動車 株価	0.01608
□ 株価	0.01576
□ 中古 車 販売	0.01536
□ レンタ	0.01280
中古 車 □	0.01239
□ 生産 方式	0.01230
□ ハイブリッド	0.01217

候補語 p	$auth(p)$
日産	0.01345
ホンダ	0.01262
toyota	0.01187
マツダ	0.01118
スズキ	0.01063
bmw	0.00963
ダイハツ	0.00944
車	0.00913
スバル	0.00886
中古 車	0.00864
三菱	0.00854
ポルボ	0.00751
自動車	0.00675
アウディ	0.00649
バイク	0.00641
フォルクスワーゲン	0.00582
ポルシェ	0.00579
三菱 自動車	0.00574
ベンツ	0.00539
メルセデス ベンツ	0.00497

表 6 「ルイヴィトン」の実験結果
(HITS アルゴリズムを用いた手法)

絞込型 p	$hub(p)$
□ 長 財布	0.05052
□ 財布	0.04980
□ パス ケース	0.04432
□ 名刺入れ	0.04366
□ 直営 店	0.04149
□ カード ケース	0.04134
□ 小銭 入れ	0.03978
□ バッグ	0.03861
□ 携帯 ストラップ	0.03721
□ キーホルダー	0.03379
□ 財布 新作	0.03331
□ 偽物	0.03133
□ 偽者	0.02815
□ 新作 財布	0.02814
□ ネックレス	0.02661
□ 財布 メンズ	0.02436
□ ストラップ	0.02377
□ バック	0.02263
□ バッグ 新作	0.02203
□ サングラス	0.02190

候補語 p	$auth(p)$
グッチ	0.01410
ブランド	0.01348
シャネル	0.01272
コーチ	0.01262
エルメス	0.01196
gucci	0.01147
ブルガリ	0.01125
プラダ	0.01104
バーバリー	0.01102
ポール スミス	0.00993
coach	0.00989
ディオール	0.00906
フェラガモ	0.00872
カルティエ	0.00860
アナスイ	0.00787
ビビアン	0.00772
クロエ	0.00733
サマンサ タバサ	0.00721
dior	0.00633
革	0.00624

表 7 「京都」の実験結果
(HITS アルゴリズムを用いた手法)

絞込型 p	$hub(p)$
□ 市 旅行	0.04622
□ 市 観光 ガイド	0.04595
□ 市 エリア ガイド	0.04545
□ 市 ホテル	0.04464
□ 市役所	0.04146
□ 市 観光 協会	0.03921
□ ビジネス ホテル	0.03600
□ 市 バス 時刻 表	0.03584
□ 土産	0.03136
□ 風俗	0.02765
□ 観光	0.02735
□ グルメ	0.02693
□ 市	0.02692
□ 市 バス	0.02690
□ 観光 地図	0.02519
□ 駅	0.02518
□ ラーメン	0.02423
□ 桜 名所	0.02320
□ 観光 おすすめ	0.02183
□ 信用 金庫	0.02177

候補語 p	$auth(p)$
大阪	0.01363
名古屋	0.01223
広島	0.01190
札幌	0.01182
横浜	0.01164
神戸	0.01123
福岡	0.01120
仙台	0.01080
岡山	0.01061
新潟	0.00986
金沢	0.00936
静岡	0.00903
長野	0.00863
鹿児島	0.00851
函館	0.00838
熊本	0.00822
千葉	0.00808
長崎	0.00789
東京	0.00734
浜松	0.00710

表 8 「ギター」の実験結果
(HITS アルゴリズムを用いた手法)

絞込型 p	$hub(p)$
□ 練習 法	0.03489
□ 入門	0.02952
□ レッスン	0.02624
□ 初心者	0.02586
□ 弾き 方	0.02559
yamaha □	0.02467
□ コード 表	0.02464
ヤマハ □	0.02387
□ 上達	0.02214
ジャズ □	0.02201
□ 譜面	0.02126
□ 教室	0.02075
□ 修理	0.02050
中古 □	0.01964
□ 練習	0.01927
□ アーティスト	0.01919
□ コード譜	0.01845
□ コード 一覧	0.01823
□ スコア	0.01800
□ 弾き語り	0.01735

候補語 p	$auth(p)$
ピアノ	0.00960
ベース	0.00920
ドラム	0.00636
ウクレレ	0.00455
ゴルフ	0.00437
フルート	0.00419
バイオリン	0.00417
楽譜	0.00392
トランペット	0.00381
パソコン	0.00365
音楽	0.00364
バイク	0.00361
楽器	0.00335
アコギ	0.00331
サックス	0.00329
テニス	0.00316
無料	0.00256
野球	0.00250
キーボード	0.00227
エレキ ベース	0.00219

表 9 「サッカー」の実験結果
(HITS アルゴリズムを用いた手法)

絞込型 p	$hub(p)$
神奈川 県 □ 協会	0.04898
千葉 県 □ 協会	0.04813
福岡県 □ 協会	0.04681
埼玉 県 □ 協会	0.04592
静岡 県 □ 協会	0.04589
大分県 □ 協会	0.04441
東京 都 □ 協会	0.03896
大阪 □ 協会	0.02963
□ 団体	0.02902
千葉 県 □	0.02587
日本 □ 協会	0.02432
静岡 県 □	0.02432
兵庫県 □	0.02373
愛知 県 □	0.02341
□ ルール	0.02225
神奈川 県 □	0.02159
埼玉 県 □	0.02116
大阪 府 □	0.01917
□ 用品	0.01870
□ 大会	0.01825

候補語 p	$auth(p)$
テニス	0.00874
バスケット ボール	0.00854
バレーボール	0.00813
ゴルフ	0.00735
バドミントン	0.00712
ラグビー	0.00712
ソフトボール	0.00693
野球	0.00678
卓球	0.00665
ハンドボール	0.00595
陸上 競技	0.00591
観光	0.00583
不動産	0.00554
バスケット	0.00535
スポーツ	0.00517
看護	0.00506
トラック	0.00505
体操	0.00499
サッカー	0.00497
体育	0.00487

6) Dekang Lin: "Automatic retrieval and clustering of similar words", Proceedings of the 36th annual meeting on Association for Computational Linguistics, pp. 768–774 (1998).
7) Keiji Shinzato, Kentaro Torisawa: "A simple www-based method for semantic word class acquisition", Proceedings of the Recent Advances in Natural Language Processing (RANLP05), pp. 493–500 (2005).
8) H. Cui, J. Wen and W. Ma: "Probabilistic query expansion using query logs", Proceedings of the eleventh international conference on World Wide Web, ACM Press, pp. 325–332 (2002).
9) B. F. Federal: "Using association rules to discover search engines related queries".
10) J.-R. Wen, J.-Y. Nie and H.-J. Zhang: "Clustering user queries of a search engine", World Wide Web, pp. 162–168 (2001).
11) C. Silverstein, M. Henzinger, H. Marais and M. Moricz: "Analysis of a very large altavista query log", Technical Report 1998-014, Digital SRC (1998).
12) J. M. Kleinberg: "Authoritative sources in a hyperlinked environment", Journal of the ACM, **46**, 5, pp. 604–632 (1999).