Vol.2009-DBS-148 No.15
Vol.2009-FI-95 No.15
2009/7/28

Web 検索におけるクエリー属性推定のための クリック・グラフのモデリング

町 永 圭 吾[†1]　内 山 達 也[†1]
Georges Dupret[†2]　藤 田 澄 男[†1]

本論文では Web 検索のクリックスルーデータを用い，PageRank の手法を応用してクリックされた URL の類似性からクエリーの類似性を推定した．この結果少数のシードクエリーから同じファセット語と一緒に使われるトピック語を推定できた．また，URL とクエリーは関係が疎であること，URL をドメイン単位で纏めてしまうとクエリーとの関係が密すぎることに着目し，クエリーに対して URL を構成する各階層との関連を考慮する改善を行い，性能が改善したことを報告する．

Click Graph Modeling for Query Attribute Estimation on Web Search

Keigo Machinaga,[†1] Tatsuya Uchiyama,[†1]
Georges Dupret[†2] and Sumio Fujita[†1]

We apply a PageRank method to click-through data on web search, exploiting clicked URL similarity and estimate relations between queries. As the results of the experiments with small seed query sets, we could extract topic words, which are queried with same facet words. We also combine layered URL node with the method to avoid sparseness between queries and clicked URLs and to relax denseness between queries and clicked URLs' domains. We found the combination improves efficiency.

最近の Web 検索エンジンにおいては，例えば「橋本　地図」といえば橋本駅（市）周辺の地図，「橋本　画像」といえば，どこかの橋本さんの画像，「橋本　株価」といえば，橋本総業の本日の株価ページが，検索結果のトップに現れるのが当たり前のことのように思われてきた．いずれも「橋本」は，まったく別の種類の固有名である．ここでの第 2 の検索キーワードは，検索要求の対象を主題的に絞り込むのではなく，検索対象を，データベースの範囲またはサービスの種類として指定して，絞り込むことを意図しているメタ・キーワードのような役割を果たして，実際にそれぞれに対応した専門データベースの結果が出ている．これらは，トランザクション指向のクエリー[1)] において，そのタスク意図を端的に示す役割をしているようにみえる．例えば，NTCIR-3[2)] の Web タスク検索課題から次のような例を見てみよう．

「サルサ 学ぶ 方法」,「テーピング 方法」,「宮部みゆき 書評 レビュー」,「キュービリック 映画 感想」,「カプサイシン とうがらし 効能」,「ポリフェノール 種類 効果」,「N ゲージ HO ゲージ 意味」,「柴犬 日本犬 特徴」,「グレートバリアリーフ オーストラリア 旅行」

Web 検索をシミュレートした，これらのショート・クエリーの最後の検索語は，検索要求の主要概念をなす名詞句とは独立して，その外延を絞り込もうとしている．これらの検索語は，適切な検索サービスやデータベースに誘導する手がかりとして，クエリーの関連語サジェスチョン機能[3)] などで活用されているが，またファセット・サーチのファセット概念とも重なる．これらを絞り込み用のファセット的なキーワードとみなして，あるクエリーの第一の検索語，主要概念をトピック語と呼び，第 2 の絞り込み用の検索語をファセット語と呼ぶ．例えば「画像」というファセット語は，専ら画像的な情報が求められるトピック語とともに使われるだろう．「ダウンロード」や「インストール」は，それぞれの操作の対象になるトピックと結びつくだろうし，「レシピ」は，調理可能な対象と結びつくだろう (**表 1** を参照)．

一方，検索結果に対するクリックについてみれば，このように比較的，タスク意図の明確な検索では，タイトル，スニペット，URL などから，クリックの是非が容易に判断可能と考えられる．「画像」のあとにクリックされるのは，画像を豊富に用意しているサイトだろうし，「レシピ」の後には，著名なグルメ情報サイト，「音楽」のあとにはミュージック情報サイトが多くクリックされるだろう．このように，サイトまたは URL 側と，クエリー側が共通にもつ検索のタスク意図に関わる属性は，クエリー側では，時には明示的にファセット語で示されるが，時には暗黙的に期待されることがある．表 1 の「嵐 time」は音楽グループと曲名からなるクエリーである．ここで必要なのは，どのようなトピック語には，どのよ

†1 ヤフー株式会社 (Yahoo Japan Corporation)
†2 Yahoo! Inc.

Vol.2009-DBS-148 No.15
Vol.2009-FI-95 No.15
2009/7/28

表 1 クリックスルーデータの例

クエリー	クリック URL
オクラ レシピ	http://www.ajinomoto.co.jp/recipe/special/515/S2.asp
たけのこ レシピ	http://allabout.co.jp/gourmet/cookingabc/closeup/CU20060420B/index.htm
チキン南蛮 レシピ	http://cookpad.com/danaehime/recipe/102203/
山本梓 画像	http://yamamoto.sakura88.com/
眞鍋かをり 画像	http://www.g-idol.com/ma/Manabe_Kaori.htm
プーさん 画像	http://kou.chu.jp/deko/puhtml.shtml
無料 音楽	http://www.gyao.jp/music/
youtube 音楽	http://www.youtube-select.net/
ライアーゲーム サントラ	http://cinematicroom.com/soundtrack/0015989/
パイレーツオブカリビアン サントラ	http://www.7andy.jp/cd/detail?accd=C0965609
春のワルツ サントラ	http://www.hmv.co.jp/news/newsdetail.asp?newsnum=608170040
よさこい節	http://www.d-score.com/ar/A02112002.html
exile summer time love	http://music.yahoo.co.jp/shop/c/10/rzcd45590/
嵐 time	http://www.hmv.co.jp/product/detail/2576322

うなファセット語が暗黙的に期待されるかを，Web 検索の利用実態から把握することである．画像を豊富にそろえたサイトがよくクリックされるようなクエリーは，暗黙に「画像」ファセットが期待されていると考えられる．このような手がかりを取得するために，Web 検索のクリック・スルーデータを分析する．クリック・スルーデータは，検索ユーザが，検索結果のリンクをクリックしたときに，リダイレクト・サーバーに記録される，クエリー，クリック URL，タイム・スタンプ，検索順位，ブラウザ識別子などからなるログで，クエリー・リコメンデーション[4] や検索結果への非明示的フィードバック[5] などに利用される．本研究では，クリック・スルーデータを，クエリーと URL のなす 2 部グラフで表現して，グラフ上の属性伝播によって，検索クエリーのファセットとなるような属性を推定する方式を，日本語の商用 Web 検索サービスのデータを使って評価する．

1. 関 連 研 究

Beeferman らは Web 検索のログに対し，クリックした URL の共有率が高いクエリーをマージするステップと，使用したクエリーの共有率が高い URL をマージするステップを繰り返すクラスタリングを行った[6]．この手法はクラスタリングであるため，関連を調べたいクエリー集合を事前に与えるためには工夫を加える必要がある．Craswell らは画像検索のログに対し，遷移先のノードの次数に応じて遷移確率が決まる逆向きランダムウォークを用いて特定のクエリーに関連した画像を抽出した[7]．ここでは画像の関連度が評価されている．Li らは Web 検索のログに対し，特定の意味のクエリー集合（シードセット）からクエリーの文字列から意味を推定し，さらにそれを初期値としてグラフ上の学習を行い，関連クエリーを抽出した[8]．彼らの目的は我々のものと似ているため，あとで比較検討する．

クリックグラフはスパースであるため，URL の代わりに URL のドメイン部を利用する方法が行われている[7][8]．以下では biased-PageRank[9] によるクエリー同士の関連性のスコアリング方法を提示し，クエリーと URL 間の関連性を表すグラフ構造によって，どのように性能が変化するかを報告する．

2. クリックグラフ上のスコアリング

URL の集合を $\mathcal{U}$，クエリーの集合を $\mathcal{Q}$ とし，それらの和集合 $\mathcal{V} = (\mathcal{U} \cup \mathcal{Q})$ をノードとする．クエリーに対してユーザーがクリックした URL の記録から得られるクエリーと URL の関係 $\mathcal{E}$ をエッジとする．このノードとエッジを用い，クリックグラフをグラフ $(\mathcal{V}, \mathcal{E})$ で表す． このグラフは成分 $A_{i,j}$ をクリック頻度とする隣接行列 $A \in \mathbb{N}^{|\mathcal{V}|\times|\mathcal{V}|}$ で表すことができる．クエリーに対してクリックされた URL はクエリーと相互にエッジを持っていることとする．この隣接行列をそれぞれのノードに対して出次数で重みを正規化し，遷移行列 B を得た．

$$B_{i,j} = \frac{A_{i,j}}{\sum_j A_{i,j}} \tag{1}$$

Vol.2009-DBS-148 No.15
Vol.2009-FI-95 No.15
2009/7/28

相互にエッジを持っていることとしたので A は対称行列だが，ノードごとに出次数が異なるため B は対称行列ではない．また，ページ間の HyperLink 構造をモデル化した通常の PageRank とは異なり，遷移先がなく除数が 0 になる場合はないので，これを考慮する必要はない．このようなグラフに対して，特定の意味を持つクエリーの集合，シードクエリー $\mathcal{S}$ を想定し，これに対応するベクトル $\boldsymbol{s}$ を作る．$\boldsymbol{s}$ は

$$s_i = \begin{cases} 1/|\mathcal{S}| & (\mathcal{V}_i \in \mathcal{S}) \\ 0 & (\mathcal{V}_i \notin \mathcal{S}) \end{cases} \tag{2}$$

を要素とする長さ $|\mathcal{V}|$ のベクトルである．ただし $\mathcal{V}_i$ は行に対応するクエリーまたは URL とする．以下の実験ではシードとしてクエリーのみを使用しているが，URL を用いることもできるし，クエリーと URL の両方を用いることもできる．このシードのベクトルに対し，各々のクエリーの関連度を推定するため，次の更新式：

$$\boldsymbol{m}^{(k)} = (1-\alpha)B\boldsymbol{m}^{(k-1)} + \alpha\boldsymbol{s} \tag{3}$$

による $\boldsymbol{m}^{(k)}$ の収束値 $\boldsymbol{m}^*$ を求めた．これは確率 $(1-\alpha)$ で出エッジのいずれかをクリック頻度の比に応じてランダムに移動し，確率 α でシードとして与えたクエリー集合のいずれかにテレポートするランダムウォークを表す biased-PageRank とほぼ等価である．

biased-PageRank は，Haveliwala の Topic-Sensitive PageRank[9] で用いられた手法である．通常の PageRank が確率 α で全くランダムにテレポートするランダムウォークを表しており，特定の意味づけを持たない重要度を表すのに対し，Topic-Sensitive PageRank は確率 α で特定のノード集合のうち 1 つにランダムにテレポートするランダムウォークを表しているため，そのノード集合との関連度が考慮される．しかし PageRank と同じく，リンクを多く集めているノードがそうでないノードに比べ高いスコアを持つので，重要度も反映しているといえる．

以下では，$\boldsymbol{m}^*$ を $\mathcal{S}$ に対する類似度スコアとして，スコアの降順でクエリーを列挙し，シードクエリーを除外したものを処理結果として用いている．

類似の手法としては，Li らの方法[8] がある．これは同様にグラフによるスコアの伝播を行っているが，クエリーから URL を経由し，再びクエリーに至るまでのステップに対し，正規化を行っている点が大きく異なる．

$$D_{i,j} = \begin{cases} 1/\sum_j (AA^T)_{i,j} & (i=j) \\ 0 & (i \neq j) \end{cases} \tag{4}$$

$$B' = D^{-1/2}A \tag{5}$$

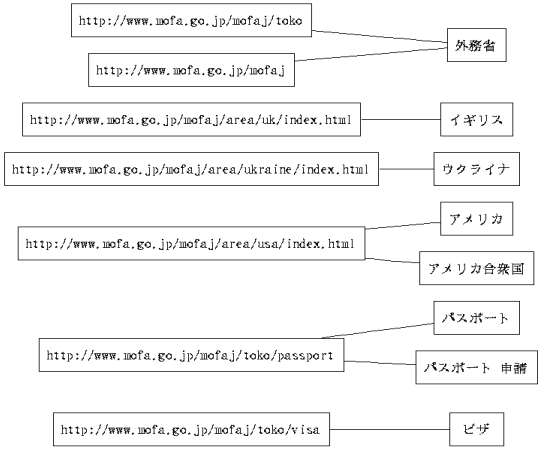

図 1 クエリーと URL にエッジを作る例

$$\boldsymbol{m}^{(k)} = (1-\alpha)B'B'^T\boldsymbol{m}^{(k-1)} + \alpha\boldsymbol{s} \tag{6}$$

ここで A は先ほどの隣接行列であり，D は正規化の行列，$B'B'^T$ はクエリー間の遷移行列である．この方法は D を求める計算量が大きく，処理全体の負担になってしまうという問題がある．また，正規化は経由してきた URL によらず，クエリーから URL を経由して到達したクエリーのスコアの和で正規化するので，クリック総数が大きい URL に影響されやすいと考えられる．

3. URL 階層の考慮

URL はクエリーと強い関連があるが，それゆえに非常に狭い範囲での類似性しか取り出すことが出来ない懸念がある．例えば**図 1** は外務省の URL とそこに辿り着いたクエリーの一例を示している．「イギリス」「ウクライナ」「アメリカ」はこのグラフでは全くリンクしていない．このような場合，URL の代わりに URL のドメイン部を用いる方法が行われている[7][8]．**図 2** は同様に外務省の URL とそこに辿り着いたクエリーの一例を示している．「イギリス」「ウクライナ」「アメリカ」の間にリンクを作ることが出来たが，「ビザ」「パスポート」「外務省」など意味の遠いキーワードも同じ強さの関係になってしまった．そこで我々は，各クエリー間の関連が強いものと弱いもので伝播するスコアに差が出るように，URL の階層構造

Vol.2009-DBS-148 No.15
Vol.2009-FI-95 No.15
2009/7/28

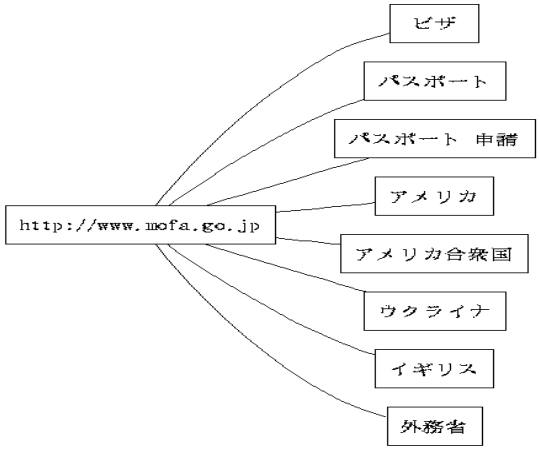

図 2 クエリーとドメインにエッジを作る例

図 3 クエリーと URL の各階層にエッジを作る例

を利用することにした．**図 3** が提案するモデルである．各クエリーは，クリックされた URL の各階層に一様に関連しているとみなす．例えばクエリー「イギリス」であれば，http://www.mofa.go.jp, http://www.mofa.go.jp/mofaj, http://www.mofa.go.jp/mofaj/area, http://www.mofa.go.jp/mofaj/area/uk, http://www.mofa.go.jp/mofaj/area/uk/index.html に一様の重みで関連しているとする．これは厳密なモデルとは言えないが，実装が容易であり，共通する URL 階層の深さを関連度に反映させることができる．これにより例えば「イギリス」「ウクライナ」間は「イギリス」「ビザ」間より属性推定において高い関連度を持つ．

このような拡張はクエリー側にも可能であると考えられる．例えば「イギリス 旅行」「ウクライナ 旅行」といったクエリーから「旅行」ノードを作成することも考えられる．今回は実験により効果を検証するため，URL 側の階層構造のみを考慮した．

以降，クリック・スルーデータの URL 部分について，URL をノードとしたグラフを用いるものを URL ノード方式，URL の各階層をノードとしたグラフを用いるものを階層ノード方式，URL のドメイン部を用いるものをノードとしたグラフをドメインノード方式と呼ぶ．

4. 評　　価

4.1 実験方法

複数のシードセットを用意し，実装した手法の有効性を検証した．検索のクリックスルーデータは Yahoo! JAPAN の検索 1 日分から，クリック頻度の高いクエリーと URL のセット 100 万件を選択したものである．**表 2** はクリックスルーデータに関する情報を示している．図の 100K.set は，クリック頻度の高いクエリーと URL のセット 10 万件を用いたもので，計算時間のかかる比較手法との比較に用いた．

図中の「シード・評価データ」は，1M.set では,「中古」,「株価」,「予約」,「レシピ」,「画像」,「ファッション」,「ダウンロード」の 7 種類のファセット語に対して，これらの語を空白で分割された最後の要素に持つクエリーの数である．100K.set では同様に「レシピ」「画像」について収集したクエリーの数である．100K.set の場合は，十分な数が収集できたこの 2 語を対象とした．**表 3** は「レシピ」,「予約」,「ファッション」に関するシードデータを例示したものである．このようなデータセットを擬似的な正解とみなして 2 等分のクロスバリデーションによる評価を行った．このときシードには，ファセット語がついているものと，ファセット語を削除したものの両方を用いた．例えば「野菜嫌い」であれば「野菜嫌いレシピ」と「野菜嫌い」の両方のクエリーに対してシードとしての値を付与する．このよう

Vol.2009-DBS-148 No.15
Vol.2009-FI-95 No.15
2009/7/28

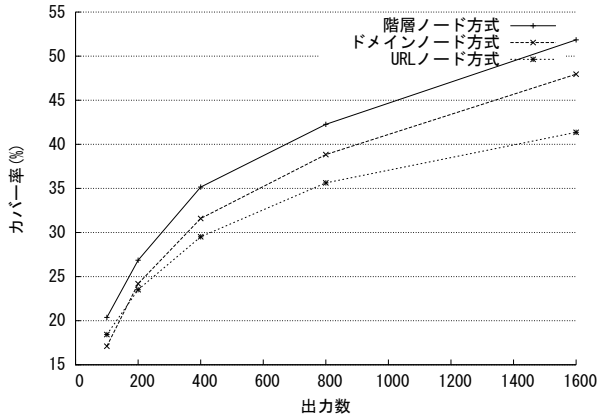

図 4 出力数を変化させたときのカバー率の変化

にしたとき，片方のみを用いるよりよい結果が得られた．評価データに対してはファセット語を除いたもののみを評価対象とし，ファセット語が付属した語はシステムが出力したとしても無視した．これは，ファセット語が付属しないような語も抽出できるかどうかを評価する実験に近づけるためである．以下のいずれの実験でも，パラメーター α は Haveliwala にならって 0.25 に設定した[9]．

表 2 実験データのリスト

	1M.set	100K.set
クエリー ノード数	403,574	51,664
URL ノード数	686,055	74,370
クエリー・URL エッジ数	1,000,000	100,000
各 URL 階層 ノード数	1,054,717	111,962
クエリー・各 URL 階層 エッジ数	2,464,690	202,497
ドメイン ノード数	225,728	41,603
クエリー・ドメイン エッジ数	932,192	94,845
シード・評価データ (クエリー数)	4,344	354

4.2 実験結果

表 4 は「レシピ」に関するシードを与えたとき，提案手法でスコアが高かったクエリーを例示している.「ビシソワーズ」「梅ジュース」など関連のあるクエリーが取得できていることがわかる．**表 5** は前述の Li らが採ったスコア伝播方法との比較である．この方法は処理時間がかかるため，100K.set を用いて実験した．条件をそろえるため，各手法でスコアが高いものから 800 件を出力結果とした．カバー率を，出力結果に含まれる正解の割合とする．

表 3 シードクエリーの例

レシピ	予約	ファッション
野菜嫌い	ゴルフ	ストリート
金目鯛	スカイライナー	チカーノ
酢味噌	キッザニア	フランス
りんご酢	じゃらん	セレブ
タイ風カレー	jr 新幹線	80 年代
食パン	北斗星	40 代
イタリアン	車検	流行
エビフライ	嵐 time	サーフ系
オムライス	海外 ホテル	結婚式
ひやむぎ	海外 ホテル	通販

表 4 「レシピ」シードに対する提案手法でスコア上位 20 件のクエリ

クエリー	スコア
ビシソワーズ	0.0899
梅ジュース	0.0874
シエスパ	0.0736
料理レシピ	0.0707
ポテトサラダ	0.0644
アムウェイ	0.0586
簡単レシピ	0.0563
ドライカレー	0.0559
らっきょの漬け方	0.0554
生姜焼き	0.0523
今日の料理	0.0517
梅ジャム	0.0498
ホットケーキミックス	0.0475
大根サラダ	0.0471
温泉卵の作り方	0.0468
お弁当レシピ	0.0468
ゴーヤ	0.0465
温泉卵	0.0461
梅ジュースの作り方	0.0454
cookpad	0.0435

Vol.2009-DBS-148 No.15
Vol.2009-FI-95 No.15
2009/7/28

表 5 100K.set に対する各手法でスコアが高いクエリ 800 件の評価用データに対するカバー率によるラベルの伝播方法の比較（どちらも URL のノード作成はドメインノード方式）

正解ラベル	比較手法	提案手法
レシピ (134)	54.48% (73)	66.42% (89)
画像 (220)	53.64% (118)	57.27% (126)
マクロ平均	54.06%	61.85%

表 6 各方式でスコアが高いクエリ 800 件の評価用正解クエリーに対するカバー率によるグラフモデルの比較

正解ラベル	ドメインノード方式	階層ノード方式	URL ノード方式
中古 (57)	49.12% (28)	52.63% (30)	36.84% (21)
株価 (191)	41.88% (80)	44.50% (85)	37.70% (72)
予約 (60)	56.67% (34)	61.67% (37)	63.33% (38)
レシピ (1374)	29.91% (411)	37.85% (520)	27.87% (383)
画像 (2235)	20.09% (449)	23.00% (514)	19.60% (438)
ファッション (52)	28.85% (15)	26.92% (14)	21.15% (11)
ダウンロード (375)	45.33% (170)	49.33% (185)	42.93% (161)
マクロ平均	38.84%	42.27%	35.63%

$$\text{カバー率} = \frac{\text{出力中のテストが仮定する正解の数}}{\text{テストが仮定する正解の数}} \tag{7}$$

評価用データは正解として含めたいクエリーをくまなく保有してるのではないため，再現率ではなくカバー率とした．評価用正解クエリーが，真の正解クエリー母集団のグラフ上での分布を反映しているとみなして，同一のシードセットに対するそれぞれの手法のカバー率を比較することによって，どの手法が，シードセットの反映するクエリー属性を持つクエリー集合を同定するのに適しているかを評価できると考えられる．表 5 にあるように 2 つのファセット語に対するカバー率のマクロ平均は，提案手法が，比較手法に対して 14.4%改善した．**表 6** はクエリーと URL 間の関係を表すグラフの作成方法を比較した結果である．システムは同様に，各手法でスコアが高いものから最大 800 件を出力結果としている．この結果，7 つのファセット語に対するカバー率のマクロ平均は階層ノード方式が最もよく 42.27%，ドメインノード方式で 38.84%，URL ノード方式で 35.63%だった．階層ノード方式と，ドメインノード，URL ノードの各方式との差は統計的に有意である (t-test，$\alpha = 0.05$)．一方ドメインノード方式と URL ノード方式の差は統計的に有意でなかった．**図 4** は出力件数を変化させた場合のカバー率の変化である．100 件，200 件，400 件，800 件，1600 件のいずれの場合でも階層ノード方式が最もよいカバー率を示した．

4.3 考　　察

表 6 から,「レシピ」「画像」などシード数・評価数が多い実験セットの方がカバー率が低いことがわかる．しかし実際に，上位にランクされたクエリーを目視してみると，他に比べて悪いという印象はない．ここから，ファセット語との共起によって集められた評価用正解クエリー以外にも，正解とされるべきクエリーが多数，ランク結果に含まれていることが想定される．評価用正解クエリー以外の，正解とされるべきクエリーが上位を占めてしまった場合にも，カバー率は下がるからである．グラフ生成方式の比較では，URL ノード方式がもっともカバー率が悪かったが，このような理由で，URL ノード方式の評価値が，不当に低くなっていないか，検討してみることにする．

実際，URL ノードの場合，シードからのパスがなくスコアが 0 となる評価用正解クエリーが多く確認された．例えばレシピとの共起クエリーである 1,376 件のうち半数の 688 件をシード，残りの半数の 688 件を評価用正解クエリーとして使用した場合を見てみよう．クリック・グラフ上の全クエリー 403,574 件中の 7,879 件のクエリーが，非ゼロのスコアが得られた．そこに評価用正解クエリー 688 件のうち 363 件 (52.76%) が含まれた．残りの評価用正解クエリーはスコアが 0 であった．URL ノード方式では，グラフが疎であるために，明らかに意味的に近いクエリーに対してスコアを伝播できていないことがわかる．一方，上記 URL ノード方式でスコアが 0 にならなかったクエリー数と同数の 7,879 件を，階層ノード，ドメインノードの各方式のクリック・グラフで学習した結果によるクエリーランク結果上位から取得する．階層ノード方式では 547 件 (79.51%)，ドメインノード方式では 512 件 (74.42%) の評価用正解クエリーが含まれていた．ここから，階層ノード方式のクリックグラフが，より多くの評価用正解クエリーに対してスコアを伝播できていることがわかる．これは，図 4 で，出力数の全範囲で階層ノード方式が優れていることによっても裏付けられる．

一方，精度重視でランク結果の上位のみを用いるようなタスクを想定して,「レシピ」,「画像」の上位 100 件について目視で評価してみた．ここでの精度は，上位 100 件中の目視で正解と判定されたクエリーの割合である．

$$\text{上位 100 件精度} = \frac{\text{上位 100 件中の目視正解数}}{100} \tag{8}$$

表 7 のように，今度は URL ノード方式が最もよく，階層ノード方式は，わずかに低かっ

Vol.2009-DBS-148 No.15
Vol.2009-FI-95 No.15
2009/7/28

表 7 各方式でスコアが高いクエリ 100 件に対する目視での精度評価によるグラフモデルの比較

正解ラベル	ドメインノード方式	階層ノード方式	URL ノード方式
レシピ (1374)	26% (26)	43% (43)	46% (46)
画像 (2235)	41% (41)	50% (50)	53% (53)
マクロ平均	33.5%	46.5%	49.5%

た．一方，ドメインノード方式は大きく劣っていた．グラフを拡張することによって，関連するクエリーを網羅的に集めることができる反面，上位精度を，わずかに損なうことがわかる．階層ノード方式の上位ランク結果を見ると，非常に一般的なクエリーが，ノイズであるにもかかわらず高いスコアを与えている例が幾つかあった．これは，ドメインノード方式でも同様の問題がある．本研究では，特定のファセット属性をもち得るクエリーを網羅的に列挙することを目的としているので，網羅性に優れかつ上位精度も満足できるレベルにある階層ノード方式の方が優れているといえる．

5. おわりに

本研究では Web 検索のクリックスルーデータを用い，PageRank の手法を応用してクリックされた URL の類似性からクエリーの類似性を推定した．この結果少数のシードクエリーから同じファセット語と一緒に使われるトピック語を推定できた．また，URL とクエリーは関係が疎であること，URL をドメイン単位で纏めてしまうとクエリーとの関係が密すぎることに着目し，クエリーに対して URL を構成する各階層との関連を考慮する改善を行った結果，カバー率で測った推定効果を改善することができた．本研究では，クエリーをシードとして与え，同じファセット属性を持つクエリーを推定したが，次のステップでは，URL をシードとしたり，URL 側の属性を推定する実験を行いたい．また，今回は URL 階層を用いてグラフを拡張して効果が認められたが，クエリー側の構造を利用してグラフを拡張することも試してみたい．拡張によりノード数も増えグラフが密になるが，クリック・グラフは，Web グラフに比べてもともとコンパクトなため，処理は可能であると思われる．

参考文献

1) Broder, A.: A taxonomy of web search, *ACM SIGIR Forum*, Vol.32, No.2, pp.3–10 (2002).
2) : NTCIR. http://research.nii.ac.jp/ntcir/index-ja.html.
3) : Yahoo!検索 ヘルプ - 「関連検索ワード」とは. `http://help.yahoo.co.jp/help/jp/search/web/web-17.html`.
4) Dupret, G. and Mendoza, M.: Recommending Better Queries from Click-Through Data, *Proceedings of the 12th International Symposium on String Processing and Information Retrieval(SPIRE 2005),LNCS 3246*, Springer, pp.41–44 (2005).
5) Joachims, T., Granka, L., Pan, B., Hembrooke, H. and Gay, G.: Accurately interpreting clickthrough data as implicit feedback, *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM New York, NY, USA, pp.154–161 (2005).
6) Beeferman, D. and Berger, A.: Agglomerative clustering of a search engine query log, *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM New York, NY, USA, pp.407–416 (2000).
7) Craswell, N. and Szummer, M.: Random walks on the click graph, *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM New York, NY, USA, pp.239–246 (2007).
8) Li, X., Wang, Y. and Acero, A.: Learning query intent from regularized click graphs, *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM New York, NY, USA, pp. 339–346 (2008).
9) Haveliwala, T.: Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search, *IEEE Transactions on Knowledge and Data Engineering*, Vol.15, No.4, pp.784–796 (2003).