Query Networkによる情報発見・収集支援

佐藤進也[†] 原田昌紀[†] 風間一洋[†]

Query Networkとは，Web検索の「行為者」，「検索語」，「検索結果中から選択し閲覧したページ」の三者の相互関係をサーチエンジンのログから抽出し，グラフ構造により表現したものである．検索行為を検索語と閲覧Webページを結び付ける作業，あるいは検索語（で表される検索要求）に最も適合したものとして閲覧ページを推薦する行為と見なせば，Query Networkは複数ユーザによる協調的情報収集の結果としてとらえることができる．本論文では，このQuery Network中の検索語とWebページの関係に特に注目し，情報発見・収集支援に利用する方法を提案する．さらに，この支援法の妥当性を検証し，その実装例であるQuery Network Navigatorを紹介する．

Information Discovery with the Query Network

Shin-ya Sato,[†] Masanori Harada[†] and Kazuhiro Kazama[†]

Query Network is the graph structure representing aggregate relationships between users, queries and web pages in Web search histories. Taking users' search behaviors on the Web as actions making correspondences between queries and relevant Web pages, or in other words, recommendations for relevant Web pages, the Query Network can be thought of an outcome of implicit collaborations among users. In this paper, we propose techniques for making use of relationships among queries and Web pages in the Query Network for information discovery and ascertain their validity. We also presents *Query Network Navigator*, an implementation of the proposed techniques.

1. はじめに

Webは非常に巨大かつ多様性・変化に富んだ情報メディアである．このメディアから必要な情報を効率良く取り出すためには，情報検索の従来手法に加えて，膨大な情報の量，多様さや時間的変化，さらに情報間の相互関係などの情報の性質を考慮した工夫が必要である．Webマイニング[1)]はそのための1つのアプローチであり，Webを解析して特徴を抽出し，それを利用して情報の効率的獲得を狙う．

Webマイニングの主な解析対象としては，Webページの内容，リンク構造，ユーザによる閲覧や検索の履歴などがあげられる．この中でも特に検索履歴は，Web上で生み出された情報が取捨選択を経て利用されている状況，いいかえれば情報流通のダイナミクスを示すものとして非常に興味深い解析対象である．

このような観点から，我々はサーチエンジンのログ解析をすすめている．その解析手法の1つに**Query Network**[2)]がある．Query Networkは，ログに記録されている検索の「行為者」，「検索語」，「検索結果中から選択し閲覧したページ」の三者の相互関係をグラフ構造により表現し可視化したものである．検索行為を検索語と閲覧Webページを結び付ける作業，あるいは検索語（で表される検索要求）に最も適合したものとして閲覧ページを（不特定な他者に）推薦する作業と見なせば，Query Networkは複数ユーザによる協調的情報収集の結果としてとらえることができる．本論文では，このQuery Network中の検索語とWebページの関係に特に注目し，情報発見・収集支援に利用する方法を提案するとともに，その妥当性を検証する．

以下，本論文での議論を次のようにすすめる．まず，2章でQuery Networkの構成方法を示す．次に，3章でQuery Networkを使った情報発見・収集支援法を提案する．さらに，この支援法の効果を高めるためにQuery Networkに求められる条件を明らかにする．4章において，これらの条件が実際に満たされていることを，Webサーチエンジン ODIN☆のログから得られたQuery Networkを使って確認する．5章では，

† NTT 未来ねっと研究所
NTT Network Innovation Laboratories

☆ http://odin.ingrid.org/.

本支援法の実装例として Query Network Navigator を紹介し，利用例を通してその効果を確かめる．さらに，6 章で関連研究との比較を行い，提案手法の特徴を明らかにする．

2. 構成方法

本章では，Web サーチエンジンのログに記録されたユーザの検索・閲覧行為の履歴をもとに Query Network を構成する方法を示す．Query Network には「行為者」,「検索語」,「ページ」の 3 要素すべての相互関係を示す**基本ネットワーク**と，特定の要素に注目し基本ネットワークを縮退することで得られる**派生ネットワーク**がある．これらをそれぞれ 2.1 節と 2.3 節で定義する．

2.1 基本ネットワーク

Web サーチエンジンにおいて，cookie と HTTP リダイレクトのメカニズムを応用することで，ユーザが検索に使用した語に加えて検索結果を閲覧している状況を記録することができる[3]．具体的には，たとえば，「2001 年 10 月 1 日 0 時 0 分 5 秒に `Bobbie` という（cookie で識別される）ユーザが，「グーグル」という語で検索した結果から `http://www.google.com/index.html` というページを選択・閲覧した」という事実を，ログに “`2001/10/01 00:00:05 Bobbie` グーグル `http://www.google.com/index.html`” というレコードとして残すことができる．

この各レコードを，Query Network の最小単位グラフ（**図 1**）に対応させる．これは，ユーザによる検索・閲覧を語と Web ページを結び付ける仲介行為と見なし，グラフによって表現したものである．

ユーザ（`Bobbie`）の隣に表示されている数字は時刻に対応するもので，あらかじめ決めておいた時刻からの経過（秒）を示している．Query Network では，同一ユーザによる行為でも検索語あるいは閲覧 Web ページが異なる場合には独立したものとして扱い，時刻を識別子としてそれらを分離する．

ログ中の複数のレコードから複数の最小単位グラフが得られるが，レコードにまたがって同じ語を使った検索や同一 Web ページの閲覧がある場合には，その語やページに対応するノードを共有させることで最小単位グラフを連結する．たとえば，ユーザ `John` が「google」で検索し `http://www.google.com/index.html` を閲覧したとする．この閲覧ページは図 1 のものと同じなので，この 2 つの事実があったことを示すグラフは**図 2** のようになる．

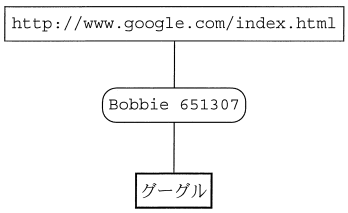

図 1 Query Network の最小単位グラフ
Fig. 1 Atomic graph of the Query Network.

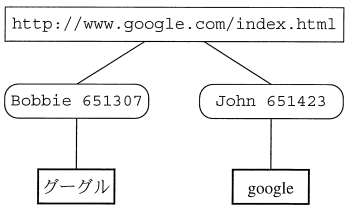

図 2 連結された最小単位グラフ
Fig. 2 Unified atomic graphs.

このように，順次ログのレコードから最小単位グラフを生成し，最小単位グラフどうしを適宜連結させることにより得られるグラフが Query Network である．一般に Query Network は複数の連結成分からなり，それぞれの連結成分においては意味的なまとまり（一貫性）が認められる[2]．連結成分が全体として意味的なまとまりを持つうえで，検索語の多義性，様々な話題を網羅するページ，ユーザの閲覧ミスなどが障害となりうる．しかし，「いま “java” といえば多くの人は（インドネシア共和国にある島のことよりむしろ）プログラミング言語のことだと思う」というような語の使用に関するトレンド（あるいは偏り）があること，適切な検索語や Web ページを選択するためにユーザの知的判断がなされていること，さらに，サーチエンジンが検索結果に各ページの内容を引用するなどしてユーザが適切なページを選択できるよう支援していることなどがこの問題をおおむね解消し，連結成分に意味的一貫性を与えていると考えられる．

なお，本論文では，ノードとリンクを構成要素に持つ（抽象的な）構造あるいはその表現方法を “グラフ” と呼び，具現化されたグラフを “ネットワーク” と呼ぶことにする．

2.2 実　　例

2001 年 10 月 1 日から n 週間の期間 I_n $(n = 1, 2, 3)$ にサーチエンジン ODIN に寄せられた検索リクエストのうち，検索語に 1 語のみを用いているレコードだけ

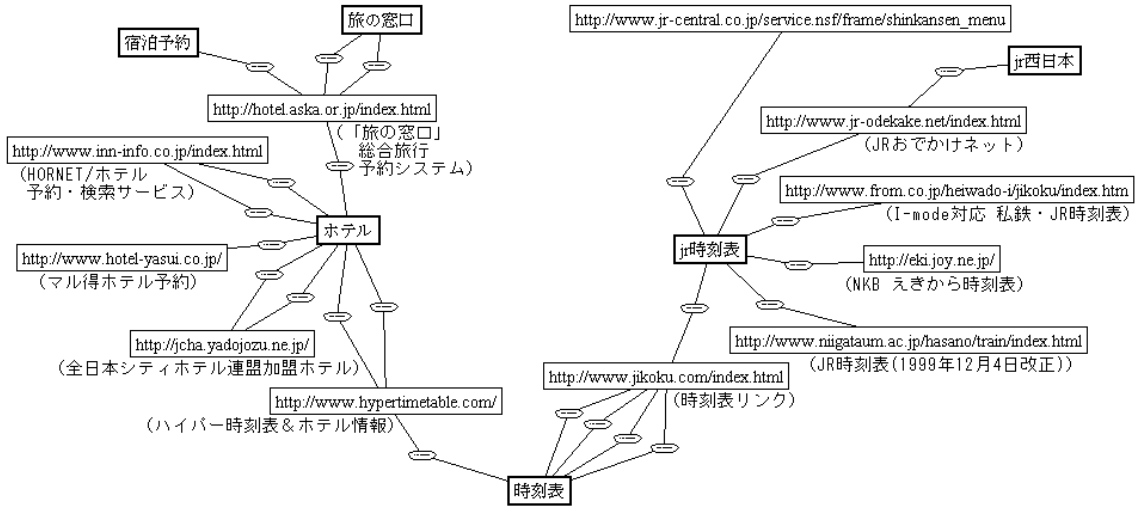

図 3 Query Network の例
Fig. 3 Example of the Query Network.

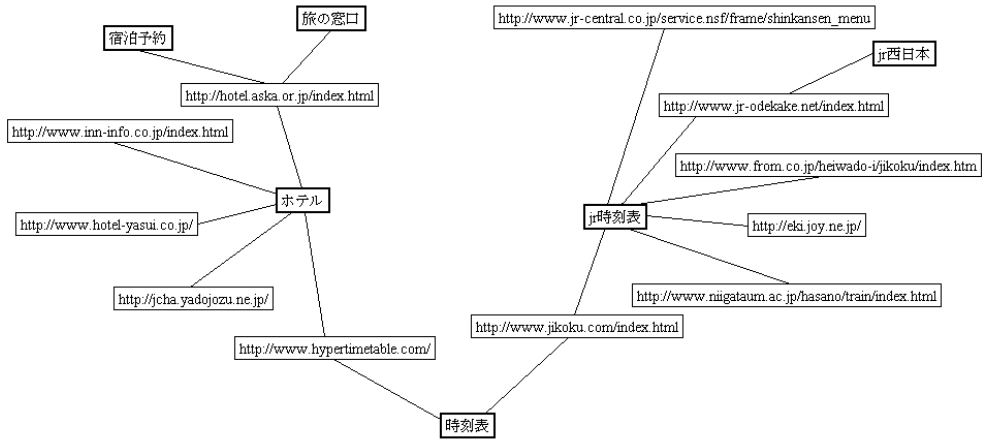

図 4 検索語と Web ページからなる派生ネットワーク
Fig. 4 Derived network of queries and Web pages.

を抽出し☆，検索語の大文字は小文字に，全角は半角に正規化して構成した Query Network を $N(I_n)$ とする．例として，$N(I_1)$ の部分ネットワークを**図 3** に示す（以下これを N_0 と呼ぶ）．本図では検索語と Web ページ（URL）の可視性を高めるためにユーザに対応する部分を小さく表示している．また，Web ページにタイトルがある場合には URL の下部に表示した．

以降，本論文ではこの $N(I_n)$ を用いて議論をすすめる．

☆ Web 検索の大半は 1 語のみによる[4] ので，検索行為を分別するうえでこれは比較的緩い条件である．

2.3 派生ネットワーク

基本ネットワークから，特定の要素間の関係を抽出することで得られるネットワークを派生ネットワークと呼ぶ．

その一例が，基本ネットワーク $N(I_n)$ から検索語と Web ページの関係を抽出して得られる派生ネットワーク $N_{q+p}(I_n)$ である．これは，検索語と Web ページをノードするグラフで，$N(I_n)$ において 1 人以上のユーザによって結び付けられているものの間にリンクを張ることで得られる（$q+p$ は「query と page」の意）．**図 4** は，N_0 に対応する $N_{q+p}(I_1)$ の部分ネットワークである．

もう 1 つの例として，検索語をノードとし，

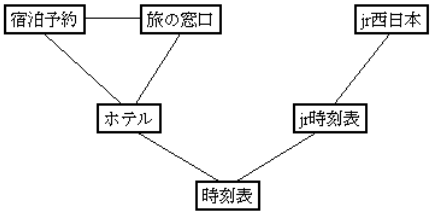

図 5 検索語のみからなる派生ネットワーク
Fig. 5 Derived network of queries.

$N_{q+p}(I_n)$ 中で少なくとも 1 つの Web ページによって結び付けられているものの間にリンクを張って得られる派生ネットワーク $N_q(I_n)$ が考えられる．図 4 のネットワーク（さらに遡れば N_0）から得られる検索語の派生ネットワークは**図 5** のようになる．

派生ネットワークにより，特定要素に焦点をあてた関係の解析と応用が可能になる．たとえば，検索語の派生ネットワーク $N_q(I_n)$ には，一種のシソーラスとしての応用が考えられる．

さらに，派生ネットワークは基本ネットワークを理解するための道具としても利用できる．派生ネットワークの特徴にはそのもととなる基本ネットワークの特徴が反映されているので，派生ネットワークを調べることが基本ネットワークの理解につながる．たとえば，図 4 の派生ネットワークの構成により，図 3 の基本ネットワークは巨視的には木構造をなしていることがはっきりと分かる．

3. 情報発見・収集の支援

前章の例を見て分かるように，Query Network から検索語や Web ページどうし，あるいは検索語と Web ページの間の関係を読み取ることができる．本章では，この相互関係を情報発見・収集支援に利用する方法を提案する．

Query Network の一要素である行為者も，たとえば検索語や Web ページの人気の指標として，情報の取捨選択の助けになると考えられる．しかし，行為者はそれ単独では検索語や Web ページに比べて理解性が低い（何を意味しているかが分かり難い），関係把握の対象を絞り込むことにより相互関係を単純で把握しやすいものにできる，という理由から，本論文では Query Network の 3 要素のうち検索語と Web ページの 2 者に焦点をあてる．

具体的な相互関係の利用方法としては，次の 2 つを提案する．その 1 つはネットワークのつながり（トポロジ）を利用するものであり，もう 1 つは時間経過にともなうネットワークの成長を利用するものである．

3.1 トポロジの利用

2.2 節で例として示した Query Network N_0 では，検索語「時刻表」から出発して右方向に順次リンクをたどることで鉄道（の時刻表）に関する情報を収集できる．これがネットワークのトポロジを利用した情報収集である．

この情報収集を効率的なものにするために Query Network に求められるのが，関連性の高い情報どうしをまとめ，かつ，関連性の低いものを分離して配置するという性質である．この 2 つの（性質を持っているという）条件により情報探索に方向性を持たせることができる．

N_0 ではこの 2 条件が満たされており，左右にそれぞれ宿泊関連情報と交通機関関連情報がまとまっているうえ，それらの間には（検索語「時刻表」を介する以外）リンクが張られていない．その結果，単純に右方向にリンクをたどりさえすれば宿泊情報の領域（左側）に迷い込むことなく交通機関の情報だけを収集できる．

この性質はこの例だけにとどまらず，ほぼ Query Network 全体に認められる．本論文では，この主張の根拠となる次の 2 つの事実を示す．1 つは，検索語どうしのネットワーク N_q 上での近接性（何ホップで到達可能か）と意味的な関連性に相関があることであり（4.1 節），もう 1 つは，派生ネットワーク N_{q+p} にはループが少なく，木構造の集合体になっているという事実である（4.2 節）．これらの事実は，関連性の高い情報どうしは互いに木構造上の近くに配置され，関連性の低いものどうしは，N_0 でトピックが左右に分かれていたように，枝分かれによって分離されている☆ということを示している．

3.2 時間経過にともなう成長の利用

時間の経過にともないサーチエンジンにおける検索履歴の蓄積量は増加する．そして Query Network もまた履歴の量に依存して成長する．たとえば，検索語「狂牛病」を含む Query Network の連結成分は，10 月 1 日 11 時の時点では 17 ノード，12 時間後には 49 ノード，そして 1 週間後には 154 ノードからなる大きなネットワークへと成長している．この急速な成長は，狂牛病への関心の高まりから多くの関連情報が生産され（検索による）取捨選択を経て多くの人々に利用されているという状況を反映したものと考えられる．つまり，Query Network の成長から，Web で情報が生み出され利用されている状況を読み取ることができ

☆ あるいは，非連結性によって分離されている場合もある．

る．これが，ネットワークの成長を利用した情報収集である．

ネットワークの成長を調べる（観察する）ということは，具体的には，時間経過にともなってネットワークにノード（検索語，Web ページ）が “付着” していく様子を追うことである．付着を追う視点をどこ（どの範囲）におくかによって情報収集のスタイルも変わってくる．たとえば，ネットワーク中のある特定箇所に視点を固定しその近傍を観察するという方法は，特定の話題に注目してその変化を追うというスタイルに対応する．4.3 節では，この方法で $N_q(I_1)$ の成長を観察し，実際に関連情報を発見できることを示す．

4. Query Network の性質

前章で提起した Query Network に期待される 3 つの性質，すなわち，ネットワーク上での検索語の近接性と意味的な関連性に相関があること，N_{q+p} にはループが少なく，木構造の集合体になっているということ，ネットワークの特定箇所に注目しその近傍を観察することで関連情報を収集できることを，それぞれ本章の各節で示す．

4.1 検索語間の関連性

4.1.1 関 連 度 γ

語間の意味的な関連度を測るためには，CLASSI システム[5] で用いられている相関係数（correlation coefficient）c を応用する．その基本的なアイデアは，2 つの語 q_1，q_2 の関連性を（コーパスなどの）文書集合 U における出現の依存（非独立）性から推定するというものである．

相関係数 c は，独立性検定の χ^2 値☆から導かれるもので，**表 1** にあるような分割表（x は U に属する文書で語 q_1，q_2 がともに出現するものの数．y，z，w も同様）を考えたとき，

$$c(q_1,q_2)=\frac{(xw-yz)\sqrt{|U|}}{\sqrt{(x+y)(z+w)(x+z)(y+w)}}$$

表 1 2 つの語の出現による U の分割表
Table 1 Corsstable for testing occurrence dependency of two terms in documents in U.

	q_1 が出現する	q_1 が出現しない
q_2 が出現する	x	y
q_2 が出現しない	z	w

☆ 観測値と独立性を仮定したときの期待値との隔たりを示す値．表 1 の分割表の例では，観測値 i $(i=x,y,z,w)$ に対する期待値を e_i とすると，$(e_x-x)^2/e_x+(e_y-y)^2/e_y+(e_z-z)^2/e_z+(e_w-w)^2/e_w$ で与えられる．この統計量は自由度 1 の χ^2 分布に従う．

で与えられる．q_1 が出現するという事象と q_2 が出現するという事象の独立性が高ければ c は小さくなるので，この値の大きさをもって q_1 と q_2 の関連度とすることができる．

関連度を表すという c の性質を保ちつつ，関連性の把握（比較）を容易にするため，本論文では最大値が 1 となるように定数倍した γ を用いる．

$$\gamma(q_1,q_2)=\frac{c(q_1,q_2)}{\sqrt{|U|}}$$

以下，U として ODIN の索引に含まれる Web ページの集合（およそ 4,230 万 URL）を用いて γ を計算する．例として，Web ディレクトリサービス Open Directory☆☆のカテゴリ名をいくつか選んでそれらの間の γ を計算した結果を**表 2** に示す．

関連性が低いと思われる「投資信託」と「アウトドア」の γ の値は，関連性のある「アウトドア」と「スポーツ」のものより低くなっている．さらに，「スポーツ」における「相撲」と「野球」，「野球」における「読売ジャイアンツ」と阪神タイガース」と，カテゴリの階層が深くなっていくのに従い γ の値が大きくなっており，適切に関連性を表しているのが分かる．

4.1.2 近接性と意味的な関連性

この γ を使って，ネットワーク上の検索語の近接性と意味的な関連性の関係を調べる．

$N_q(I_1)$ は数多くの検索語からなるため，すべての関係を調べあげるには多大なコストがかかる．そこで，計算量を減らすために以下のような手順で処理対象を制限し解析を行う．まず，$N_q(I_1)$ の連結成分を無作為に 10 選びその全ノードの集合を Q とする．Q の任意の要素 q_1，q_2 について，両者を結ぶ $N_q(I_1)$ 上の最短経路長 h（同じ連結成分に属さない場合は，便宜的に ∞ とする）と $\gamma(q_1,q_2)$ を計算し，h ごとに γ の平均をとり，h との関係を調べる．

この計算を**図 6** のネットワークを例にとり，具体的に説明する．図中の q_a などは検索語を表している．このネットワークでは q_a と q_b，q_a と q_c，q_b と q_c，q_c と q_d の最短経路長は 1 であり，q_a と q_d，q_b と q_d

表 2 γ の値の例
Table 2 Some γ values.

語の組	γ
投資信託，アウトドア	0.007
アウトドア，スポーツ	0.123
相撲，野球	0.167
読売ジャイアンツ，阪神タイガース	0.220

☆☆ http://dmoz.org/

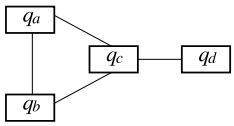

図 6 N_q の連結成分の例
Fig. 6 Example of a connected component of N_q.

表 3 最短経路長 h と γ の平均値
Table 3 Relationships between shortest path length h and average of γ.

サンプル $\setminus$ h	1	2	3	4	∞
Q_1	0.133	0.034	0.032	0.009	0.001
Q_2	0.138	0.083	0.049	0.016	0.001
Q_3	0.112	0.040	0.002	–	0.005

の最短経路長は 2 となっている．よって，γ の平均は，

$h = 1$ の場合：

$$\frac{\gamma(q_a, q_b) + \gamma(q_a, q_c) + \gamma(q_b, q_c) + \gamma(q_c, q_d)}{4}$$

$h = 2$ の場合：

$$\frac{\gamma(q_a, q_d) + \gamma(q_b, q_d)}{2}$$

となる．

表 3 は，無作為に構成した Q の 3 つのサンプル Q_1，Q_2，Q_3 に対して，それぞれ上記手順により γ の平均を計算した結果である．Q_1，Q_2，Q_3 を構成する連結成分の大きさ (最小, 最大, 平均) はそれぞれ (3, 5, 3.4)，(3, 8, 3.8)，(3, 9, 4.3) であり，最短経路長の最大値はそれぞれ 4，4，3 であった．いずれの場合も，関連度 γ はおおむね h が大きくなるにつれて下がる傾向にあり，関連性の高いものほど互いに近くに存在していることが分かる．

ただし，例外として，Q_3 の非連結な検索語どうしの関連性が $h = 3$ の場合より若干高くなっている．これは，検索語や URL の完全一致という Query Network の単純な構成方法では抽出しきれない関係がまれに存在するためであると考えられる．この “見落とされた” 関係の一部は，3 種類の要素（ユーザ，検索語，Web ページ）の総合的関係が表現されているという Query Network の特徴を利用することで効率的に発見できる．その方法を次項で紹介する．

4.1.3 関連性の発見

一般にユーザの検索要求は複数のコンテキストにわたる（たとえば，趣味的なものと仕事上のもの）．このことを考慮して，Query Network では同一ユーザの行為もあえて分離している．

しかし，同一ユーザによる 2 つの行為でそれぞれ使用された検索語，あるいは閲覧された Web ページの間に高い関連性が認められる場合には，これらは同じ検索要求に基づく行為と見なせる．この考え方は次のように一般化できる．すなわち，（同一ユーザによるとは限らない）2 つの行為において，ユーザ，検索語，Web ページのいずれか 2 つに関連性が認められた場合に，これらの行為に関連性があると見なす．

表 4 発見された関係の例
Table 4 Discovered relationships.

発見された関連性	手がかり
w3m，lynx	U+Q
八ヶ岳，小淵沢	U+Q
東京ディズニーランド，tdl	U+Q
rfc2960，sctp	U+Q
東洋水産，明星食品	U+Q
linux，redhat	U+Q
ガステーブル，魚焼器	U+P
adsl，isdn	P+Q
クマガイソウ，アツモリソウ	U+P+Q

図 7 リンク数とグラフの構造
Fig. 7 Relationship between the number of links and graph structure.

この方法を $N(I_1)$ の異なる連結成分に属する行為間に適用して新たな関連性の発見を試みた．これは，前項で触れた “見落とされた” 関係の発見にほかならない．ユーザ，検索語，Web ページが満たすべき関連性の条件としては，それぞれ，同一であること，γ が 0.13 以上であること，URL のホスト名とパス名が第 3 階層まで一致していることとした．ここで，0.13 という値は，$N_q(I_1)$ 中 1 ホップで到達可能な検索語どうしの γ の平均（0.135）を基準に選んだ．なお，この基準値は文書集合 U や検索履歴に依存して決まるので，ネットワークごとに個別に計算する必要がある．

表 4 は発見された行為間の関連性の例である．表にはそれぞれの検索行為で使用された語と，どのような手がかりからその関係が発見されたかを示している．U，Q，P はそれぞれユーザ，検索語，Web ページであり，それらを “+” でつなげたものは複数の手がかりが用いられたことを示している．

4.2 構造的特徴

グラフ構造のおおまかな特徴はノード数に対するリンク数の比で把握できる．このことを説明するために，**図 7** のように，連結グラフにノードを 1 つずつ順次つなげていくことを考える．明らかに，1 つのノードをつなげるためには最低 1 本のリンクが必要である．

表 5 $N_{q+p}(I_n)$ のリンク数とノード数
Table 5 Number of links and nodes of $N_{q+p}(I_n)$.

n	連結成分数	リンク数 < ノード数
1	8,362	8,320 (99.5%)
2	15,181	15,097 (99.4%)

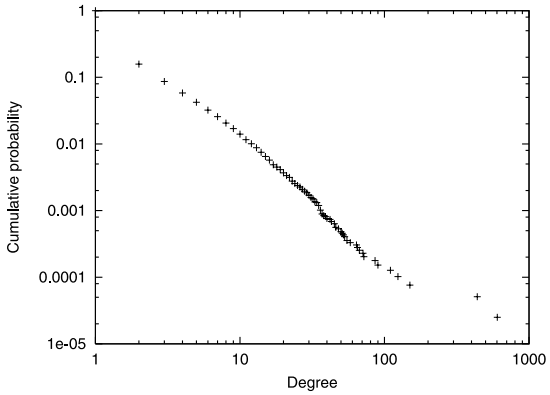

図 8 $N_{q+p}(I_1)$ の次数分布
Fig. 8 Degree distribution of $N_{q+p}(I_1)$.

そして，1 本でつなげている限り，つねにリンク数はノード数より少なくなっており，ループは生成されない．しかし，あるノードが 2 本以上でつながるとリンク数はノード数以上となり，ループが発生する．つまり，ノード数とリンク数の比はループ構造の有無を示している．

この関係を利用して，$N_{q+p}(I_n)$ におけるループの有無を調べた．その結果を**表 5** に示す．第 3 カラムの数値は，リンクがノードより少ない連結成分の数と割合である．$n = 1, 2$ のいずれの場合でも，ほとんどの連結成分ではリンク数がノード数を下回っており，ループが存在していない（すなわち木構造をなしている）ことが分かる．

さらに N_{q+p} の構造上の特徴を把握するため，次数（ノードごとのリンク数）の分布を調べた．その結果が**図 8** である．グラフは $N_{q+p}(I_1)$ における次数 x の累積存在確率，すなわち x より大きな次数を持つノードが存在する確率を両対数スケールで示したものであり，冪法則が認められる．

以上から，N_{q+p} はリンク密度が不均一な木構造の集合体で，全体としてスケールフリーネットワーク[6]になっていることが分かる．

4.3 時間発展

時間経過にともなう Query Network の成長を大きさ（ノード数）の変化で見ると，履歴の蓄積期間に比例した増加が認められる（**表 6**，$|N|$ はネットワーク N のノード数）．一方，構造に関しては，高頻度で使用（閲覧）されてきた検索語（Web ページ）は引き続き高い頻度で使用（閲覧）されるという，いわゆる “preferential attachiment” [7] の作用によってスケールフリーという特徴は維持される．

表 6 派生ネットワークの成長
Table 6 Growth of the derived networks.

n	$\lvert N_{q+p}(I_n)\rvert$	$\lvert N_q(I_n)\rvert$
1	39,395	9,706
2	79,045	18,255
3	123,001	26,922

表 7 成長にともなう $N_q(I_n)$ の近傍の変化
Table 7 Changes in neighborhoods with the growth of $N_q(I_n)$.

n	平均リンク数	$\overline{\gamma}$
1	2.823	0.135
2	5.274	0.130
3	7.106	0.127

成長にともなうネットワークの局所的な変化を見るため，$N_q(I_n)$ のノードあたりのリンク数の平均と，リンクでつながれた 2 つのノード（検索語）の関連性 γ の平均 $\overline{\gamma}$ を調べた（**表 7**）．リンク数は 1 ホップで到達可能なノード数，すなわち，あるノードの近傍にどれだけ他のノードが存在するか（量）を示す値である．また，$\overline{\gamma}$ は，近傍にどれだけ関連性のあるものが集まっているか（質）を示している．

期間の延長にともなってリンク数が有意に増加している一方で，$\overline{\gamma}$ の減少は低く抑えられており，Query Network の成長が局所的な関連情報の質を維持しつつ量の増加をもたらしていることが分かる．これは，関連情報を発見する方法として，ネットワークの特定箇所に視点を固定しその近傍を観察するという方法が有効であることを示している．

5. Query Network Navigator

以上の議論をふまえて，情報発見・収集支援システム Query Network Navigator（QNN）を試作した．

QNN はサーバ/クライアント型のシステムであり，クライアントには Web ブラウザを利用している（**図 9**）．図中，左側のフレームが QNN のユーザインタフェースである．ユーザがキーワードを入力すると，QNN は派生ネットワーク N_{q+p} から該当する検索語の近傍を探し出し，ユーザに提示する．ユーザは表示されたものの中から興味のある部分を選択し，展開することで，少しずつ近傍を広げていくことができる．

キーワードは複数指定可能で，対応する近傍をかわるがわる表示できる．図の例では「時刻表」のほかに

図 9 Query Network Navigator の利用例
Fig. 9 Screenshot of Query Network Navigator.

「検索サイト」と「ユビキタス」がキーワードとして指定されている．近傍の広がりは，N_{q+p} の木構造という特徴を利用して階層的に表示される．Web ページは URL の代わりにタイトルを表示し，検索語はタイトルと区別がつくように文字の背景色を変えてある．また，Web ページの内容はユーザの要求に応じて右フレームに表示できるようになっている．

図 9 の例は，図 4 のネットワーク中にある情報の「時刻表」を出発点とした探索に対応している．いま，ユーザの興味が交通機関の情報だけにあり宿泊情報を必要としていないと仮定する．このとき，ユーザは図 9 中の「ホテル」の部分には興味を示さず，そこからのつながりを展開しないはずである．その結果，図 4 のネットワークの左方向の探索は抑止され，右側だけが得られることになる．このように，QNN を用いることで効率的な情報収集が可能になる．

さらに QNN は，ブラウザ上の表示と Query Network の最新状況との差分を示すことができる．図 9 は期間 I_1 の検索履歴から得られた「時刻表」の近傍であるが，その 1 週間後（I_2 経過後）にこの近傍を更新すると，**図 10** のように，1 週間のうちに新しく加わった情報がボルドイタリック体で示される．この機能により，ネットワークの成長を利用した効率的な情報の発見・収集が可能になる．

以上をユーザの視点でまとめると，情報の発見・収集の手順は以下のようになる．

図 10 ネットワークの成長を利用した情報の発見・収集
Fig. 10 Information discovery by making use of network growth.

(1) QNN へのアクセス
Web ブラウザにより QNN にアクセスする．

(2) キーワードの入力
興味のあるトピックを表すキーワードを QNN のフォームに入力する．その結果，キーワードと関連性の高い Web ページ（のタイトル）や検索語が提示される．さらに，それらに関連する他の情報（Web ページや検索語）の有無が

マークで示される．さらなる関連情報が存在する場合には，それぞれ個別にを QNN に要求する（存在を示すマークをマウスでクリックする）ことができる．

これら Web ページと検索語のつながりは，キーワードをルートとし，関連性の高いものどうしが枝でつながれた木構造を形成する．この構造は，ブラウザ上ではインデント付けによって表現される．

(3) 関連情報の収集

興味に合致した Web ページや検索語の関連情報を順次要求することにより情報の収集を行う．木構造により，関連性の高い情報は互いに近く配置される一方で関連性の低い情報は枝分かれによって分離されているため，ブラウザ上に提示された情報中，興味に合致する部分の特定が容易になり，効率的な関連情報の収集が支援される．また，入力キーワードとの関連性が高い情報だけを提示するだけでなく，関連性の連鎖でつながっている情報が示されることで多様な話題が提供され，情報の発見が促される．

(4) 表示内容の更新

興味に合致した Web ページや検索語を十分に収集した時点で，QNN による情報収集はいったん完了するが，時間経過にともなう検索履歴の増加により，新たな関連情報を発見できる可能性がでてくる．

この新しい情報は，QNN の更新ボタンを押すことで取得できる．ユーザのリクエストに応じ，ブラウザ上の表示と最新状況との差分が抽出される．その新着情報のうち，ブラウザに表示されている情報（すなわち，ユーザの興味に合致した Web ページや検索語）に関連するものだけが既存情報に追加表示される．更新を定期的に行えば，そのトピックの状況の変化を追いかけることができる．

QNN はこのほかにも関連性のある連結成分（4.1.3 項）を示す機能などを持ち，多面的に情報の発見・収集を支援する．

6. 関連研究

サーチエンジンやディレクトリサービスが主に情報そのものの特徴に基づいてユーザの情報検索要求に応えているのに対し，情報利用の状況（たとえば，その時時で何が注目されているか）を情報の取捨選択に積極的に用いるものとして推薦システム[8] というアプローチがある．

表 8 QNN と CSA の関連語の比較
Table 8 Comparison of relevant terms selected by QNN and CSA.

	QNN	CSA
エルメス	carhartt，アルマーニ，カルティエ，セオリー，ジャーナルスタンダード，ナラカミーチェ	バッグ
狂牛病	bse，プリオン，肉骨粉	bse
求人	ハローワーク	読売
学会	情報処理	civil, japanese, 情報処理，電気，電気学会，土木学会，日本建築学会
java	jre	japanese，java 入門，sun，プログラミング

検索という Web 上の情報利用活動を利用して，複数のユーザが互いに関連情報を提示し合うことを支援する QNN は，Web 上の行動解析（Web usage mining）[1] に基づく推薦システムとして位置付けられる．この範疇に属するシステムには，閲覧履歴，ブックマーク，検索履歴を解析対象とするものがあり，それぞれの例として Recer [9]，Siteseer [10]，Community Search Assistant [11] があげられる．

Community Search Assistant（CSA）は，QNN 同様，関連性のある検索語をユーザ間で共有するシステムである．CSA では，検索語 q_i にその検索結果の上位 10 件 R_{q_i} を対応付け，$R_{q_i} \cap R_{q_j} \neq \phi$ であるときに q_i と q_j に関連性があるとする．**表 8** は，$N_q(I_1)$ における検索語間の関連性と，同じ語の集合に対して CSA のアルゴリズムを適用して得られる関連性を比較したものである．「エルメス」の例をとると，QNN では関連語として種々のブランド名が抽出されているのに対し，CSA では「バッグ」という一般的な語が選び出されている．その他の例でも，QNN の方が，情報を利用する側の興味を強く反映した，より特化された語を抽出する傾向にあり，新たな情報の発見を支援するという点においては，CSA より優れていると考えられる．

7. むすび

本論文では，Web 検索履歴の 1 つの表現である Query Network を利用した情報発見・収集支援法を提案し，その妥当性を確認した．

本手法では，Query Network を構成する要素のうち主に検索語と Web ページの相互関係に注目した．もう 1 つの要素である検索の行為者は，自律的に情報を扱うものとして，他の 2 要素以上に重要な役割が期待

される．実際，従来の推薦システムでは情報流通のために人と人とのマッチメイキングが利用されている．Query Network においても，行為者同士の関係を解析し情報発見に役立てることが期待される．

参 考 文 献

1) Kosala, R. and Blockeel, H.: Web Mining Research: A Survey, *SIGKDD Explorations*, Vol.2, No.1, pp.1–15 (2000).
2) 佐藤進也，原田昌紀，風間一洋：検索履歴可視化の一手法，情報処理学会研究会報告，2003-FI-71, pp.119–125 (2003).
3) 風間一洋，原田昌紀，佐藤進也：サーチエンジンの検索結果のマルチレベルグルーピングの評価，コンピュータソフトウェア，Vol.17, No.4, pp.58–69 (2000).
4) 原田昌紀，佐藤進也，風間一洋：索引篩法 — 大規模サーチエンジンのための高速なランキング検索法，第 14 回データ工学ワークショップ（DEWS2003），5-A-3 (2003).
5) Ng, H.T., Goh, W.B. and Low, K.L.: Feature selection, perceptron learning, and a usability case study for text categorization, *Proc. SIGIR '97*, pp.67–73 (1997).
6) Albert, R., Jeong, H. and Barabási, A.-L.: Diameter of the World Wide Web, *Nature*, No.401, pp.130–131 (1999).
7) Barabási, A.-L. and Albert, R.: Emergence of scaling in random networks, *Science*, No.286, pp.509–512 (1999).
8) Resnick, P. and Varian, H.R.: Recommender Systems, *Comm. ACM*, Vol.40, No.3, pp.56–58 (1997).
9) Chalmers, M., Rodden, K. and Brodbeck, D.: The order of things: activity-centered information access, *Computer Network and ISDN Systems*, Vol.30, pp.359–367 (1998).
10) Rucker, J. and Polanco, M.J.: Siteseer: personalized navigation for the Web, *Comm.ACM*, Vol.40, No.3, pp.73–76 (1997).
11) Glance, N.S.: Community Search Assistant, *Proc. Intl. Conf. on Intelligent User Interfaces*, pp.91–96 (2001).

(平成 15 年 9 月 25 日受付)
(平成 16 年 1 月 19 日採録)

(担当編集委員　石川 博，市川 哲彦，原 隆浩，
佐藤 聡，土田 正士)

佐藤　進也（正会員）

昭和 63 年東北大学大学院理学研究科数学専攻修士課程修了．同年日本電信電話（株）入社．協調作業における情報活用支援の研究に従事．現在 NTT 未来ねっと研究所主任研究員．ACM，Internet Society，電子情報通信学会各会員．

原田　昌紀（正会員）

平成 10 年東京大学大学院総合文化研究科広域科学専攻修士課程修了．同年日本電信電話（株）入社．情報検索の研究に従事．現在 NTT 未来ねっと研究所所属．

風間　一洋（正会員）

昭和 63 年京都大学大学院工学研究科精密工学専攻修士課程修了．同年日本電信電話（株）入社．現在 NTT 未来ねっと研究所主任研究員．分散協調処理，情報検索の研究に従事．ソフトウェア科学会，ACM 各会員．