Query Network の構造と時間発展を利用した情報発見・収集支援

Information Discovery based on Structural Characteristics and Growth of the Query Network

佐藤 進也† 原田 昌紀‡ 風間 一洋‡

Shin-ya SATO Masanori HARADA Kazuhiro KAZAMA

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みなせば，Query Network は複数ユーザによる協調的情報収集の結果として捉えることができる．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，その構造的特徴と時間発展を情報発見・収集支援に利用する方法を提案する．

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 especially put our focus onto subnetworks consisting of queries and Web pages, and propose techniques for making use of their structural characteristics and growth for information discovery.

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]はそのための一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挙げられる．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，言い換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．その解析手法の一つに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）推薦する作業とみなせば，Query Networkは複数ユーザによる協調的情報収集の結果として捉えることができる．

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，その構造的特徴と時間発展を情報発見・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．

2. Query Network の構成

2.1 基本ネットワークと派生ネットワーク

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，例えば，「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みなしグラフによって表現したものである．

図１ Query Network の最小単位グラフ

Fig.1 Atomic graph of the Query Network

ユーザ（Bobbi）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．例えば，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，この二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．

† 正会員 NTT未来ねっと研究所
sato@ingrid.core.ntt.co.jp
‡ NTT未来ねっと研究所 {harada, kazama}@ingrid.core.ntt.co.jp

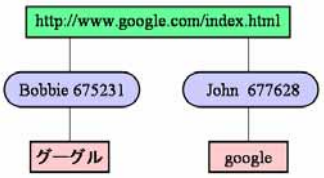

図 2 連結された最小単位グラフ

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[1]．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]．

ここで定義した Query Network から派生するものとして，特定の要素（検索語，Web ページなど）間の関係に注目することで得られるネットワークが考えられる．これを派生ネットワークと呼ぶ．派生ネットワークに対比する意味で，もともとのネットワークを基本ネットワークと呼ぶことにする．

2.2 Query Network の実例

ここで，Query Networkの実例をいくつか示しておく．

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[2]に寄せられた検索リクエストのうち，検索語に一語のみを用いているレコードだけを抽出し，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基本ネットワークを$N(I_n)$とする．図３は$N(I_1)$の一部分を示したものである（以下これをN_0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．

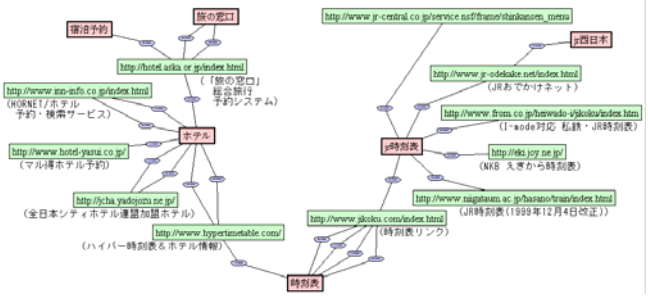

図 3 Query Network の例

Fig.3 Example of the Query Network

$N(I_n)$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を$N_{q+p}(I_n)$とする．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つ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．

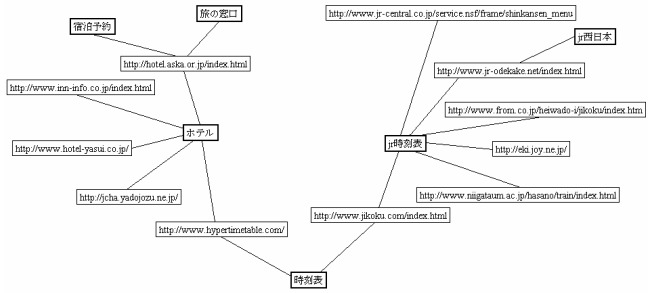

図 4 検索語と Web ページからなる派生ネットワーク

Fig.4 Derived network of queries and Web pages

派生ネットワークのもう一つの例として，検索語をノードとし，$N_{q+p}(I_n)$中で少なくとも1つのWebページによって結びつけられているものの間にリンクを張って得られるネットワーク$N_q(I_n)$が考えられる．図３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．

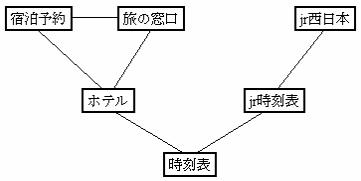

図 5 検索語のみからなる派生ネットワーク

Fig.5 Derived network of queries

以降，本論文ではこれら$N(I_n)$， $N_{q+p}(I_n)$， $N_q(I_n)$を用いて議論をすすめる．

3. 情報発見・収集の支援

Query Network を利用した情報発見・収集支援の方法として，本論文では次の２つを提案する．その１つはネットワークのつながり（トポロジー）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 トポロジーの利用

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ー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．一つは，検索語どうしのネットワーク$N_q(I_n)$上での近接性（何ホップで到達可能か）と意味的な関連

†1本論文では，ノードとリンクを構成要素にもつ（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 http://odin.ingrid.org/.

性に相関があることであり（4.1節），もう一つは，派生ネットワーク$N_{q+p}(I_n)$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別れていたように，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 時間経過に伴う成長の利用

時間の経過に伴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，時間経過に伴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

語間の意味的な関連度を測るためには，CLASSIシステム[4]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（たとえばxはUに属する文書で語q_1, q_2ともに出現するものの数）を考えたとき，

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

で与えられる．

	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

表1 2つの語の出現によるUの分割表

Table 1 Crosstable for testing occurrence dependency of two terms in documents in U

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を計算し，ネットワーク$N_q(I_1)$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，全ての関係を調べ上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γ(q_1, q_2)を計算する．ここで得られたγを hごとに平均をとり，hとの関係を調べる．

表2は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．

サンプル ＼ 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

表2 最短経路長hとγの平均

Table 2 Relationships between shortest path length h and average of γ

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QueryNetwork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．

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．このことは，初期状態として1ノードのみからなるネットワークにノードを1つずつ順次つなげていくことを考えると分かりやすい．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．そして，1本でつなげている限り常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3に示す．

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%)

表3 $N_{q+p}(I_n)$のノード数とリンク数

Table 3 Number of links and nodes of $N_{q+p}(I_n)$

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．

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）の変化でみると，履歴の蓄積期間に比例した増加が認められる．この成長にともなうネットワークの局所的な変化をみるため，$N_q(I_n)$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 <γ> を調べた（表５）．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，<γ> 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．

n	平均リンク数	<γ>
1	2.823	0.135
2	5.274	0.130
3	7.106	0.127

表５ 成長に伴う$N_q(I_n)$の近傍の変化

Table 5 Changes in neighborhoods with the growth of $N_q(I_n)$

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，<γ> 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. 関連研究

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本手法は，Web上の行動解析（Web usage mining）[1]にもとづく推薦システム[5]として位置づけられる．この範疇に属するアプローチには，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[6]，Siteseer[7]，Community Search Assistant[8]が挙げられる．Community Search Assistant（CSA）は，Query Network同様，複数ユーザによる関連検索語の共有を支援する．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，$R(q_i) \cap R(q_j)$が空でないときにq_iとq_jに関連性があるとする．CSAとQuery Networkを期間I_1のODINの検索履歴から抽出される関連語で比較すると，たとえば「エルメス」の関連語としてQuery Networkでは「アルマーニ」や「カルティエ」など種々のブランド名が抽出されるのに対し，CSAでは「バッグ」という一般的な語が選び出される．ここでQuery Networkでは$N_q(I_1)$上でリンクで直接つながっている語を互いに関連あるものとみなした．この他の例においても，Query Network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．

6. むすび

本論文では，Web 検索履歴の一つの表現である Query Network の構造的特長と時間発展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．

今回の提案では，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．もう一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，行為者どうしの関係を解析し情報発見に役立てることが期待される．

なお，我々は本手法に基づく情報発見・収集支援システム Query Network Navigator を Web ブラウザの機能を利用して実装している。その詳細については文献[9]を参照されたい．

[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] Ng, H. T., Goh, W.B. and Low, K. L.: Feature selection, perceptron learning, and a usability case study for text categorization, SIGIR'97 Proceedings, pp. 67-73 (1997).

[5] Resnick, P. and Varian, H. R.: Recommender Systems, Communications of the ACM, Vol. 40, No. 3, pp. 56-58 (1997).

[6] 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).

[7] Rucker, J. and Polanco, M. J.: Siteseer: personalized navigation for the Web, Communications of the ACM, Vol. 40, No. 3, pp. 73-76 (1997).

[8] Glance, N. S.: Community Search Assistant, Proc. Intl. Conf. on Intelligent User Interfaces, pp. 91-96 (2001).

[9] 佐藤進也，原田昌紀，風間一洋: Query Network による情報発見・収集支援，DBWeb2003, pp. 77-84, (2003).

佐藤 進也 Shin-ya SATO

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，情報処理学会，電子情報通信学会，日本データベース学会各会員．

原田 昌紀 Masanori HARADA

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．情報処理学会会員．

風間 一洋 Kazuhiro KAZAMA

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．情報処理学会，ソフトウェア科学会，ACM 各会員．