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

2003－DBS－130 (14)
2003－ＦＩ－ 71 (14)
2003／5／23

検索履歴可視化の一手法

佐藤 進也　原田 昌紀　風間 一洋
NTT 未来ねっと研究所
東京都武蔵野市緑町 3-9-11

Web サーチエンジンのログに記録されている検索行為の相互関係を，行為者，検索語，検索結果中から選択・閲覧した Web ページという属性データの相互関係を使い，Query Network というグラフ構造によって表現する方法を提案する．サーチエンジン ODIN の検索ログから得られた Query Network の実例を示しながら，その構造的特徴について述べるとともに，このネットワークを視覚化するツールを紹介する．さらに，Query Network は複数ユーザのインプリシットな協調作業の結果として捉えられることを述べ，その情報探索手段としての可能性について議論する．

A Method for Visualizing Web Search Histories

Shin-ya SATO, Masanori HARADA and Kazuhiro KAZAMA
NTT Network Innovation Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo

We propose a method for interpreting Web search histories into the graph structure called *Query Network*, which can be obtained by compiling relationships between users, queries and web pages in the histories. We describe structural characteristics of Query Networks based on some examples derived from logs of the ODIN search engine. We also introduce a tool for visualizing these networks. Furthermore, we claim that the Query Network can be thought of an outcome of implicit collaborations among users. Its potential for collaborative information discoveries is explored from this point of view.

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 マイニングは，解析の対象ごとにコンテンツマイニング，構造マイニング，利用形

態マイニングの 3 タイプに分類される [1].

利用形態マイニングでは，Web サーバのログやブラウザの閲覧履歴などを解析しユーザの情報利用行動の特徴を抽出する．Web サーバのチューニングやコンテンツデザインの改善，マーケティングなどがその主たる目的とされるが [1]，情報獲得支援のためにも有効な手段である (例えば，[2]). むしろ，“言葉の意味はその使われ方の有り様である”というウィトゲンシュタイン [3] やハヤカワ [4] の考え方に基づけば，情報利用履歴にこそ，言葉 (検索語) と情報 (Web ページ) の結び付きに関する知識を求められるはずである．

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．本論文では，ログに記録されている検索行為の相互関係を，行為者，検索語，検索結果中から選択した Web ページという属性データの相互関係を使いグラフ構造によって表現する方法を提案する．以下，このグラフを **Query Network**(あるいは略して **QN**) と呼ぶ．QN からは，行為の相互関係に加えて，検索語，Web ページ，行為者の総合的関係を読みとることができる．また，時間経過にともなって検索履歴の蓄積量が増えその結果 QN も成長するが，これは情報を探索している人々が協調して QN を育てているとみなすことができる．この QN の協調情報探索手段としての可能性についても本論文で議論したい．

以下，2 章で検索ログから QN を構成する方法を示し，3 章でサーチエンジン ODIN のログから得られた QN の実例を示しながらその構造的特徴について述べる．また，QN を表示するツール QN Grapher を 4 章で紹介し，その協調情報探索手段としての可能性について 5 章で議論する．

2 QN の構成方法

本章では，Web サーチエンジン ODIN のログから QN を構成する方法を示す．

ODIN では，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ログに記録している [6]. 具体例を挙げると，「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」という事実が，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記録される．

この各レコードを，QN 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．

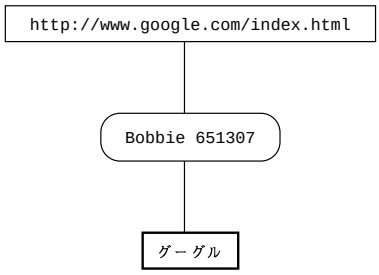

図 1: QN の最小単位グラフ

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．QN 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．

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 QN である．

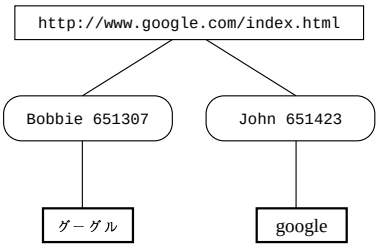

図 2: 連結された最小単位グラフ

なお，QN を構成するにあたって，今回は検索に一語のみを用いているレコードだけを処理対象とした．これは，検索の意図 (趣旨) を簡単なかたち (一つの語) で把握できるようにするためである．Web 検索の大半は一語のみによる [5] 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．検索語の大文字/小文字は同一視したが，いわゆる全角/半角の正規化は行なっていない．また，Web ページに関しては正規化[1] などの処理は一切行なっていない．

3 QN の特徴

本章では，2001 年 10 月 1 日から同年 10 月 7 日までに ODIN に寄せられた検索リクエストの履歴 53,585 レコードのうち，検索語が一つのみである 36,862 レコードをもとに作成した Query Network N_0 の特徴を調べる．

3.1 連結成分

定義から明らかなように，QN は全体として連結であるとは限らず，一般には複数の連結成分からなる．N_0 の場合，76,118 個のノードが 8,388 個の連結成分を形成している．連結成分の大きさ (成分に含まれるノード数) の分布を図 3 に示す．これは，大きさについて降順に整列させた連結成分の順位 (横軸) と大きさ (縦軸) の関係を示す両対数グラフである．最大成分を例外としてべき分布を呈しているが，その理由は次のように考えられる．まず，QN において，使用頻度の高い検索語は連結性の向上に寄与する．一方，検索語の使用頻度はべき分布に従うことが知られており [7]，その影響を受けるかたちで連結成分の大きさもべき分布に従うと考えられる．なお，閲覧 Web ページ (URL) の頻度分布にもやはりべき法則が認められるが，重複して使用 (閲覧) される頻度は検索語の場合に比べて低い[2] ので連結性向上への寄与も比較的弱いと考えられる．

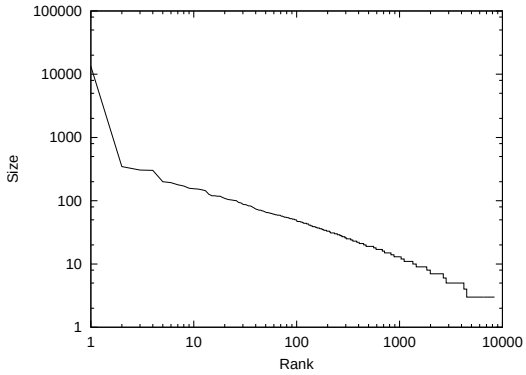

図 3: 連結成分の大きさの分布

図 4 に連結成分の一例を示す．“郵政省”や “郵便番号”，一般小包郵便の名称である “ゆうパック” とその異表記 “ゆうぱっく” などの互いに関連性のある語とそれらに関係する Web ページで構成されており，連結成分全体として意味的な一貫性が認められる．

連結成分が全体として意味的なまとまりをもつうえで，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得る．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

[1] `http://foo.jp/` と `http://foo.jp/index.html` を同一視する，など．

[2] 履歴中の平均出現頻度は，検索語の場合 3.73 回/語であるのに対し Web ページは 1.24 回/URL であった．

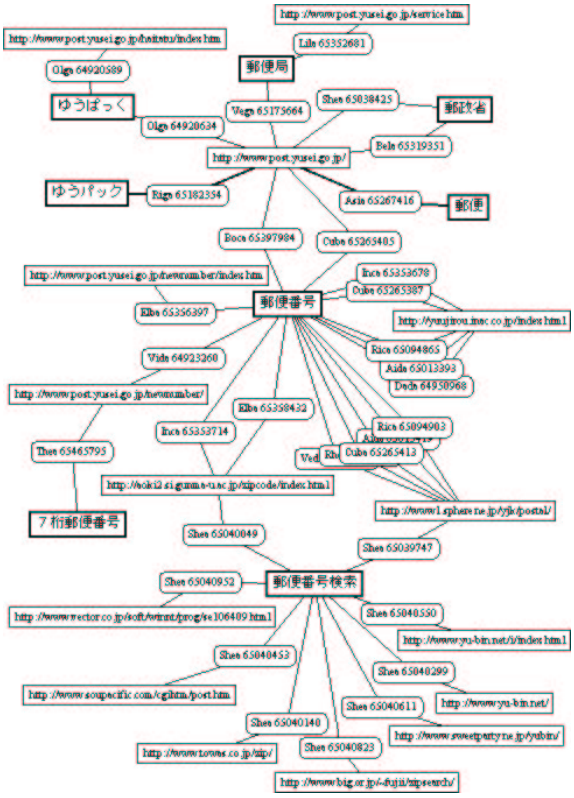

図 4: 連結成分の例

与えていると考えられる．

全体の意味的なまとまり具合，あるいは逆に話題の多様性は，各連結成分によってまちまちである．たとえば，検索語 “JR” を含む連結成分は 119 ノードから構成され，JR 関連情報に加えて時刻表や宿泊情報などの話題を含む．一方, “狂牛病” を含む連結成分のサイズは 154 ノードと “JR” の場合よりも大きいが，話題は人獣共通感染症という狭いトピックに閉じている．

3.2 構造的特徴

QN の連結成分中には，構造的に特徴をもつ部分グラフが存在する．以下，図 4 の連結成分に含まれる部分グラフの例を示し，そこから読みとれる検索語と Web ページの関係について述べる．

3.2.1 スター状グラフ

図 5 は，検索語 “郵便番号検索” から 3 ホップ以内にあるノードからなる部分グラフで，スター状になっている．この部分グラフ中の Web ページは，Web 上の郵便番号検索サービスあるいは郵便番号検索用ソフトウェアのページである．これは，スターの中心にあることば (概念) に複数の実体 (Web ページ) が対応し，それぞれが等しく (偏りなく) ユーザに利用されている状況を示している．

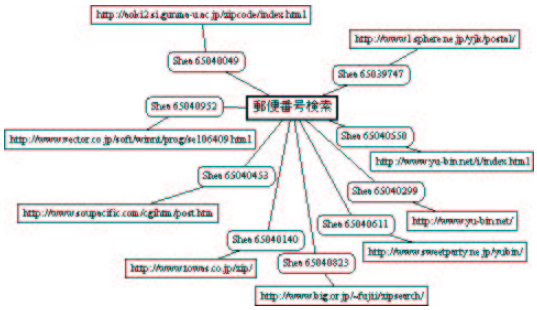

図 5: スター状部分グラフ

スター状グラフの中心は，Web ページになることもある．実際，図 4 の連結成分には，Web ページ `http://www.post.yusei.go.jp/` を “郵便”，“郵便局”，“郵政省”，“ゆうパック” という検索語が取り巻いている部分グラフがある．これは，当該ページ (あるいは当該ページを入口とする Web サイト) が提供する話題の多様さを示していると考えられる．

3.2.2 完全二部グラフ

スター状グラフの場合と異なり，検索要求に対応するページが少数に集中している場合もある．このとき，図 6 に示すような，検索語と Web ページの和集合と，ユーザ集合との間の完全二部グラフが現れる．

これは，複数のユーザが共通して同一の検索語と Web ページを結び付けているということであり，検索語と Web ページの強い関連性を示唆している．

図 6: 完全二部部分グラフ

4 QN Grapher

検索履歴の利用は，適切な検索語や Web ページを選択といった人間の知的判断を利用することである．とくに，複数のユーザによる検索行動の軌跡は情報の使われ方の趨勢を示しており，情報を取捨選択するうえでの判断材料を与えてくれる．**QN Grapher** は，このように示唆に富んだ検索履歴を眺め，情報利用の趨勢を把握し，情報探索に役立てることを支援するツールである．

4.1 QN の表示とナビゲーション

QN Grapher は，まず，指定された検索ログを読み込み QN を作成する．その後，ユーザが検索語を指定すると，QN 中から当該検索語に対応するノードを探し出し，その近傍を表示する．図 7 は，検索語として “検索” を指定したときの表示例である．

この近傍に含まれている検索語 “検索サイト” の枠が他のノードと異なる線で描かれているが，これは，このノードの先に (まだ表示していない) 連結成分のつながりが存在することを示している．このようなノードをマウスでクリックすることで，順次ネットワークを広げていくことができる．

また，各ノードごとにポップアップメニューが用意されており，関連情報などを得ることができる．図 7 では，ノード「Inca 64891726」から Inca というユーザの使用した検索語の一覧を得ている．この一覧から検索語を選択すれば，さらにその語の近傍を表示させることができる．Web ページのメニューには「Web ブラウザで開く」という項目があるので，ページの内容を確認しながら QN を拡張し関連情報を収集する，といった作業も可能である．

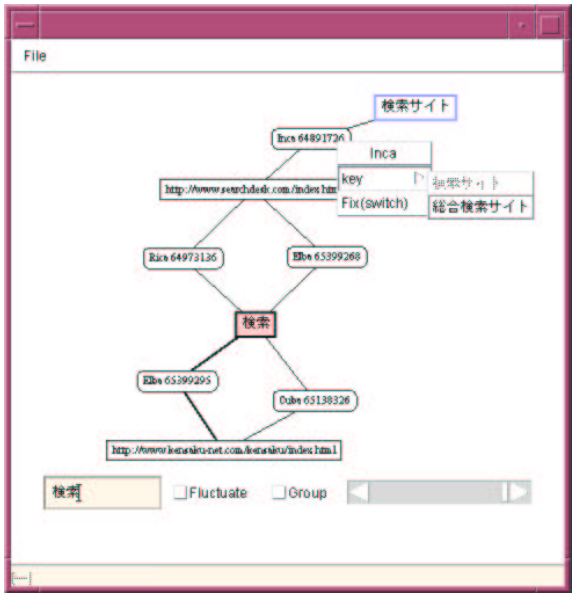

図 7: QN Grapher

4.2 時間的変化のトレース

QN Grapher のもう一つの特徴は，QN の時間的変化をトレースできる点である．右下のスクロールバーは，検索行為が発生した時刻 (各レコードに記録されている時刻) の範囲 $[t_s, t_e]$ を示している．スライダーによって時刻 $t \in [t_s, t_e]$ が選択されているとき，QN Grapher は表示の対象を $[t_s, t]$ に発生した検索行為 (に付随するノード) に制限する．t の値が大きくなるのに従い表示されるノードの数が増加するため，スクロールバーを少しずつ右にずらすことで，QN の成長する様子を観察できる．

図 8 は，この機能を使って “狂牛病” を含む連結成分の時間変化を調べたものである．この図では，“狂牛病” に対応するノードには色を付けて他のノードと区別できるようにしてある．10 月 1 日 11 時の時点では，17 ノードからなる単一のスター状グラフであったが (左)，12 時間後にはそれぞれ “プリオン” と “肉骨粉” を中心とするスター状部分グラフが連結成分に加わった (中央)．そし

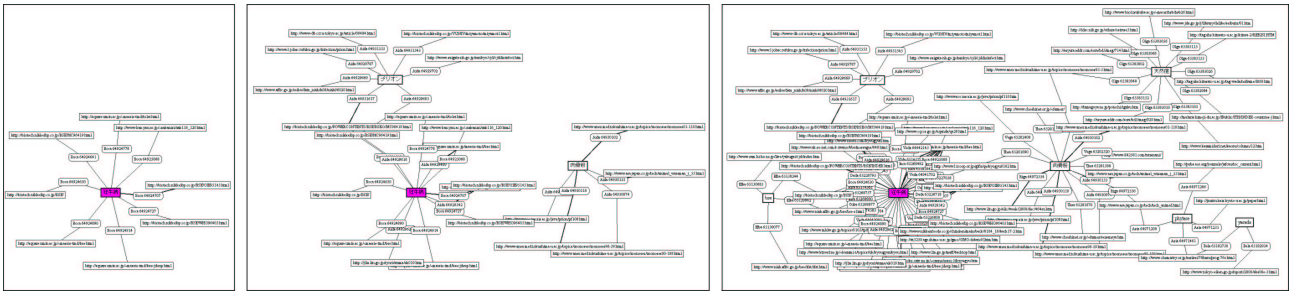

図 8: QN の成長

て，10 月 7 日 23 時には 154 ノードからなる大きな連結成分へと成長した (右)．特殊な話題であるにもかかわらず急速な成長を遂げており，狂牛病に対する社会的な関心の高さが見てとれる．

この機能は，いわば複数ユーザによる情報探索行動の再生であるが，サーチエンジンから検索の状況を直接取得すれば実況も観察できる．また，将来に向けて知識を育む場としても利用できる．たとえば，興味の対象 (検索語，Web ページ) を QN の種として蒔いておくことで，将来発生する他ユーザの検索行為を捕捉するという情報収集法が考えられる．このアプローチについては，次章で詳しく議論する．

5　QN による情報探索・獲得支援

ネットワーク上のコミュニティは明示的なものと暗示的なものとに分類される [9]．前者は，メイリングリストのように，コミュニティのメンバの活動する場が明示的に定義されているものである．後者の代表的な例としては推薦システム [8] が挙げられる．この場合，コミュニティ形成の要となるのは予め定義された場ではなくユーザの相互関係であり，より関係の強いものどうしが集まることで結果的にコミュニティが形成される．

QN において互いに結びつけられた検索行為者のグループは，この分類法に従えば暗示的なコミュニティであり，有用な情報の収集を支援する QN Grapher は推薦システムであると考えることができる．

一般に，推薦システムにおいては推薦行為 (ユーザによる善し悪しの判断と，そのシステムへの入力) に労力がかかることが多く，その結果として推薦者数が確保できず適切な推薦が行なわれなくなってしまうという問題がある [10]．この問題を解決するため，推薦対象に付随する情報やユーザの振舞い，さらにユーザどうしの関係などを利用して推薦対象を選び出す方法が考えられている．たとえば，kMedia[11] では，Web ブラウザのブックマークからユーザの相互関係を導き出し，Web ページの推薦に利用している．QN Grapher もまた，検索というユーザの情報探索行動から導き出した検索語，Web ページ，ユーザの総合的関係を示すことにより，必要とする情報の発見を支援する．

QN において，ユーザは検索行為によってコミュニティにインプリシットな働きかけをしているが，QN Grapher による可視化によって，このコミュニティはより実在感をともなうものとなる．その結果として，ユーザのコミュニティへのより能動的な関与が促される．たとえば，前章で触れたように，QN Grapher の利用方法としては，各ユーザが自分の興味にあわせて QN を整形しながら育てたり，予め作成しておいた自分の興味の表す QN を「罠」として使い将来発生する他ユーザの検索行為を捕獲することが考えられる．複数ユーザの検索行為を収集してそこから QN という知識を得たように，これらの行為をさらにコミュニティにフィードバックさせることが，情報の取捨選択に関するさらに詳しい知識の獲得につながると考えられる．これは，梅木が提案する概念で，ユーザ

のコミュニティへの能動的関与をともなう情報の取捨選択機構である双方向協調フィルタリング [9] の具体化として位置付けることができる．

6 まとめ

本論文では，Webサーチエンジンのログに記録されている検索行為の相互関係を，行為者，検索語，検索結果中から選択したWebページという属性データの相互関係を使い，Query Network(QN)というグラフ構造によって表現する方法を提案した．一般に，QNは複数の連結成分からなり，各成分には意味的な一貫性が認められる．連結成分にはスター状と完全二部グラフという構造的特徴をもつ部分グラフが存在し，それぞれ検索語とWebページの関係を反映している．

また，本論文では，QNを可視化するツールであるQN Grapherを紹介した．QN Grapherによってユーザは QNの興味のある部分を抜き出して表示させることができる．表示されたネットワークからは，検索語，Webページ，ユーザ (行為者) の総合的関係に加え，各ユーザの使用した検索語一覧や Webページの内容などの関連情報が得られる．QNは，複数ユーザの検索行為を収集して得た知識であり，インプリシットな協調作業の結果形成されたコミュニティとして捉えることができる．そして，QN Grapher はその知識を利用したWebページ推薦システムとして位置付けられる．

QN Grapherによる可視化は，QNというコミュニティを顕在化し，そこへのユーザの能動的関与を促す．今後は，この双方向協調フィルタリングとしての機能を強化しその効果を評価する予定である．

参考文献

[1] R. Kosala, H. Blockeel, “Web Mining Research: A Survey,” SIGKDD Explorations, Vol. 2, No. 1, pp. 1–15, 2000.

[2] B. Mobasher, R. Cooley, J. Srivastava, “Creating adaptive web sites through usage-based clustering of urls,” Proc. of KDEX’99, 1999.

[3] L. Wittgenstein, “Philosophical investigations,” Macmillan, 1953.

[4] S. I. ハヤカワ, “思考と行動における言語,” 岩波書店, 1965.

[5] 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,” DEWS2003, 5-A-3, 2003.

[6] 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.

[7] 佐藤進也, 原田昌紀, 風間一洋, “サーチエンジンへの問い合わせの解析,” 情報処理学会研究報告, 2000-FI-57, pp. 135–142, 2000.

[8] P. Resnick, H. R. Varian, “Recommender Systems,” Communications of the ACM, Vol. 40 , No. 3, pp. 56–58, 1997.

[9] 梅木秀雄, “ネットワークコミュニティ形成支援技術,” 人工知能学会誌, Vol. 14, No. 6, pp. 943–950, 1999.

[10] N. Good, J. B. Schafer, J. A. Konstan, A. Borchers, B. M. Sarwar, J. L. Herlocker, J. Riedl, “Combining Collaborative Filtering with Personal Agents for Better Recommendations,” Proc of AAAI-99, pp. 439–446, 1999.

[11] 濱崎雅弘, 武田英明, 松塚健, 谷口雄一郎, 河野恭之, 木戸出正継, “Bookmarkからの共通話題ネットワークの発見手法の提案とその評価,” 人工知能学会論文誌, Vol. 17, No. 3, pp. 276–284, 2002.