WWWのユーザ操作履歴によるHTML文書の相関関係の解析

風間一洋[†] 佐藤進也[†]
清水奨[†] 神林隆[††]

本論文では，プロキシサーバのログファイルから，ユーザのHTML文書単位のアクセスシーケンスを分析し，そのURLの2-gramから，あるコミュニティ中のHTML文書の相関関係を示すHTML文書関連グラフを構成する手法を提案し，HTML文書関連グラフの構成とパスの頻度から関連度を定義した．さらに，実際に小規模なコミュニティにこの手法を適用し，得られた結果を評価する．

HTML Document Correlation Analysis by User's Behavior in World Wide Web Navigation

KAZUHIRO KAZAMA,[†] SHIN-YA SATO,[†] SUSUMU SHIMIZU[†]
and TAKASHI KAMBAYASHI[††]

This paper describes the method to construct HTML document correlation graph that indicates mutual relations of them by analysis of user's access sequences to them and their URL 2-grams, and defines the correlation by the graph structure and path's frequency. And we evaluate the results that were acquired by this method in the small community.

1. はじめに

Webの普及にともなって，インターネットでさまざまな情報が得られるようになったが，情報源が分散と膨大な情報量のために，目的の情報を探し出すのは困難になった．

たとえば，情報の探索手段として，サーチエンジンやディレクトリサービスがあるが，収録された情報が多いほど，前者は情報の絞り込みが，後者は目的の情報の探索が難しくなる．プッシュ型情報サービスやダイジェストメールサービスは，情報がよく整理，選択されている半面，得られる情報の性質や範囲が制限され，いったん蓄積された情報の検索に対しては同じ問題が存在する．

そこで，あるコミュニティ内の情報閲覧者側の視点に基づく相関関係情報を導出して，各個人の広域情報探索時の支援に適用する方法について考察する．

まずユーザの情報閲覧履歴からHTML文書単位で抽出した2-gramからHTML文書関連グラフを再構成し，プログラムからそのグラフに関する任意の情報を獲得できるURL相関関係データベースを作成する方法について述べる．次に，情報閲覧履歴を収集する方法について分析し，プロキシサーバのログ分析を採用した理由と，ログからHTML文書のURLを抽出する方法について述べる．最後に，実際に小規模なコミュニティに適用し，得られたHTML文書関連グラフの性質や，自動閲覧プログラムの影響について述べ，最後に高相関度と判断された2-gramを分析し，有効性を示す．

† NTT未来ねっと研究所
NTT Network Innovation Laboratories
†† 日本テレマティーク株式会社
Nippon Telematique Inc.

2. HTML文書の相関関係と情報閲覧

2.1 HTML文書の相関関係

HTMLはハイパーテキストを記述するためのマークアップ言語である[1)]．単一の文書をHTMLで記述すると，1つ以上のリンクで結合したHTMLファイルや画像データなど複数のファイルで構成されることが多い．これらの統一した意図に基づいて作成されたファイルの集合を，まとめて**HTML文書**と呼ぶことにする．

これらのHTML文書間の相関関係として，次のような種類が考えられる．

2.1.1 情報提供者の意図に基づく相関関係

HTML文書では，情報作成者が意図した相関関係が明示的にリンクとして文書中に埋め込まれるために，

情報提供者の意図が直接反映されやすいが，リンクは比較的狭い範囲に限定される．また，リンクの質や重要度の差は表現できないだけでなく，宣伝などの内容に関連のない恣意的なリンクが行われることがある．この指定は静的であり，リンクされているHTML文書のURLの変更や内容変更に必ずしも追従しない．

2.1.2 内容の類似性に基づく相関関係

HTML文書の相関関係を内容の類似性で調べるために，サーチエンジンやディレクトリサービスが使用されてきた．たとえば，サーチエンジンは文書のキーワードに，ディレクトリサービスは人間によるカテゴリ分類に基づいている．

ただし，類似性を計算機で処理する場合は，自然言語の冗長性のためにノイズが多く含まれ，情報閲覧者側の認識とはかなり異なるスコアリングが頻繁に行われる．逆に，人間が処理する場合は，大量の情報を処理するのは難しい．

2.1.3 情報閲覧者の認識に基づく相関関係

以上の問題点のために，情報提供者が作成するリンクや内容の類似性に基づく場合は情報閲覧者が期待する相関関係とは頻繁にくい違いが生じるので，次のような情報閲覧者側の認識が反映された相関関係が導出されることが望ましい．

- 新しい情報によって古い情報の意味が失われるような情報の相対的な新旧
- ほぼ同じ内容を表す複数の情報の相対的な価値
- 情報の真偽や信頼度
- 情報作成者の意見の支持
- 情報の更新頻度

たとえば，情報閲覧者が自分自身の判断に基づいて，Webブラウザの Bookmark を使って HTML 文書を階層構造に整理することもあるが，この時点で情報が静的になるだけでなく，多くの関連する情報が失われてしまうので，情報閲覧者にとって必要な相関関係を動的に導出する．

なお，評価結果は情報閲覧者の知識レベルと興味分野によって異なる．たとえば，あるコミュニティ中にはさまざまな分野の専門家と初心者が混在する．しかし，各専門分野ごとの全体の活動を考えれば，その分野の専門家の活動が支配的であると考えられる．そこで，情報閲覧者単位でなくコミュニティ全体として評価することで，各分野の専門家の知識や経験を累積し，他の人から再利用できるようにすることを目指す．

2.2 広域情報探索支援

私たちの Ingrid（INformation GRID）と呼ぶ広域情報探索システムでは，ネットワーク上に分散した Ingrid サーバに登録された情報リソース中に含まれるキーワードの組合せを元に，情報リソースを互いにリンクで結合して，Ingrid トポロジを作成する[2]．Ingrid ナビゲータに閲覧したい情報を示すキーワードの組を入力すると，Ingrid トポロジ中のリンクをたどって分散した Ingrid サーバ中から目的の情報を探し出すことができる．

この Ingrid ナビゲータでは，Web ブラウザ本来の操作である HTML リンクに基づく情報探索に，情報の類似性に基づく情報探索を融合し，リンクと内容の類似性の両方を手がかりにして Web 情報空間を探索できるユーザインタフェースを実現した．

しかし，実験を通じて，現在の自然言語処理の限界から計算機の処理結果は多くのノイズを含み，特にインターネットを対象とした膨大な情報を扱う場合には，必ずしも効率的に情報を絞り込むことができない場合があることが明らかになった．そこで，対象とする情報のメタ情報を抽出し，それを使用して情報を絞り込ませることで，ある程度良好な結果を得た[3]．

ただし，一般にメタ情報の自動抽出は困難であり，必ずしも適切なメタ情報の抽出ができるわけではない．また，抽出が容易なメタ情報は，情報閲覧者にとって必ずしも直感的ではないことが多い．そこで，情報閲覧者側から HTML 文書の相関関係情報を抽出し，広域情報探索時の，より効率的な情報の絞り込みや，関連情報の発見や移行などの作業を支援に利用する．

3. Web 情報閲覧

3.1 Web 情報閲覧

インターネットの普及にともなって情報の電子化と公開が進んだおかげで，かなりの情報が Web ブラウザを使ってアクセスできるようになった．分野によって情報の電子化の度合いは異なるが，たとえば計算機の分野では規格書やマニュアルやバグレポート，処理系の最新版の入手に至るまで，ほとんどのすべての情報を Web から入手して仕事することができる．

このように情報獲得手段としての Web ブラウザへの依存度が高まるにつれて，閲覧中に別のトピックに興味を持ったり，ときどき最新情報をチェックしたり，別の仕事を依頼されて情報閲覧作業をいったん中断し別のトピックの情報閲覧作業を開始するなど，複数のトピックについての情報閲覧作業を並行に行うことも多くなる．新しいトピックのための別ウィンドウの作成や，Web ブラウザのヒストリ機能とブックマーク機能は，これらの並行性を支援している．

図 1 に，Web ブラウザを使って並行に情報を閲覧

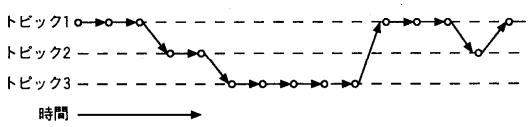


図1 情報閲覧の並行性
Fig. 1 Concurrency in Web information browsing.

する例を示す．図中の丸はHTML文書を示し，横点線は同一トピックを示す．

3.2 HTML文書の相関関係の導出

図1のようなHTML文書へのアクセスシーケンスから，情報閲覧者の認識に基づいたHTML文書間の相関関係を導き出すために，次の2種類の統計情報の利用法を検討した．

(1) アクセスシーケンス中の近接するHTML文書の出現頻度情報
(2) アクセスシーケンス中の隣接するHTML文書の出現頻度情報

前者は共起統計情報に相当する．自然言語処理の分野では，単語が近接して使われる度合を統計的に数値化した単語の共起確率は，実際の文書で単語が使用されている状況を反映するデータとしてよく使用される[4]．HTML文書の共起統計情報も，文章中に現れる単語のように，あるトピックに基づいたユーザシーケンス中に現れるHTML文書は密接な関係を持つことから，同様に有効であると考えられる．

しかし，自然言語処理においては，文書中の文脈が一貫して流れているので，単語の近接性を2単語間の単語の個数や1つの文内で判断することができるが，並列に実行されることが多いWeb情報閲覧ではアクセスシーケンスの分離が必要になり，この処理を自動化するのは非常に困難である．

自然言語処理では，ある n 個の連続した文字の組合せの出現頻度を調べてn-gram統計として利用することが多いが，同様に隣接するHTML文書のURLをn-gramとして抽出し，出現頻度の統計を収集して，Webブラウザによる閲覧行動の解析に関する基礎的データとして利用する．この方法では，ユーザのアクセスシーケンス中に図1に存在するような不連続部分も抽出されることになるので，$n=2$ とすることで不連続部分を明確化することで不連続部分の2-gramの正規化頻度は低くなり，同時に後述する相関度は低くなるので無視できるようになる．

3.3 HTML文書関連グラフ

アクセスシーケンス中のHTML文書のURLの2-gramの出現頻度の統計情報を元にして，各HTML文書をノードとして再構成したグラフを**HTML文書関連グラフ**と呼び，次のような特徴を持つ．

- 有向グラフである．
- 自分自身をさすパスを持たない．
- 各パスは2-gramの出現頻度情報を持つ．

3.4 相関度

HTML文書関連グラフにおいて，ノード u_x からノード u_y へのパスについて考える．このパスの頻度を $F(u_x,u_y)$，そしてノード u_x から出ているすべてのノードの集合を G_{u_x} とすると，u_x から出ているすべてのパスの頻度の和は $\sum_{u_z\in G_{u_x}} F(u_x,u_z)$ になる．さらに，u_x から u_y へ移行する確率 $P_{u_x}(u_x,u_y)$ は，次の式で表される．

$$P_{u_x}(u_x,u_y)=F(u_x,u_y)\Big/\sum_{u_z\in G_{u_x}}F(u_x,u_z)$$

$P_{u_x}(u_x,u_y)$ は，$F(u_x,u_y)$ が増えると増加し，u_x から u_y 以外に出ているパスの総数が増えると減少する．言い換えれば，u_x から u_y へのパスが特殊なほど値が高くなる性質を持つ．

同様に，u_y に移行する前に u_x を見ていた確率 $P_{u_y}(u_x,u_y)$ が求められ，これらの確率の相乗平均は $\sqrt{P_{u_x}(u_x,u_y)P_{u_y}(u_x,u_y)}$ になり，0と1の間の値をとる．この相乗平均値は u_x から u_y へのパスの特殊性を示す指標となるが，パスの絶対的な頻度には依存しない．

実際には，パスの頻度が低いほどノイズである可能性が低く，逆に高いほど多くの人に関連が深いと認識されていることになるので，u_x と u_y の間の相関関係の程度を示す**相関度** $E(u_x,u_y)$ を次の式のように相乗平均値と頻度の積で定義する．

$$E(u_x,u_y)=F(u_x,u_y)\sqrt{P_{u_x}(u_x,u_y)P_{u_y}(u_x,u_y)}$$

3.5 URL相関関係データベース

HTML文書関連グラフの解析や，プログラムによる相関度の利用を容易にするために，2-gramの集合をハッシュ表で管理して，あるURLを2-gramの最初または最後に持つ2-gramの集合を検索したり，その2-gramの出現頻度や相関度のデータを得ることができるURL相関関係データベースを作成した．

現在，Webブラウザの HTTP リクエストをモニタして，現在閲覧しているURLの前後のURLを相関度を使ってソートして表示することで，あるURLに関連するURLの集合を相対評価するのに用いている．

このようなデータベース化により，広域情報ナビゲーション支援のための予測/例示機能の付加や，情報利用者側から見たWeb情報空間の構造の視覚化，

Bookmark やリンク集の自動生成にも利用できると思われる．

4. モニタと縮退処理

4.1 Web 情報閲覧のモニタ

HTML 文書の 2-gram を得るためには，ユーザの Web 情報閲覧行動をモニタしなければならない．

Abrams ら[5] の研究では，Web をモニタする手法を，情報を収集するポイントに基づいて client logs, proxy logs, network logs, server logs の 4 種類に分類した．現在，行われているモニタ手法は，この 4 種類のいずれかに分類できる．

(1) ユーザの Web ブラウザ操作の監視（client logs）
Cunha らの研究[6]，Catledge らの研究[7]，Tauscher らの研究[8] では，ユーザの Web ブラウザの操作を記録する機能を Web ブラウザを改造して付加し，ファイルに記録している．

(2) プロキシサーバのログの利用（proxy logs）
私たちの研究では，プロキシサーバのログファイルを分析し，Web サーバの状態の推定[9] や，グループ指向の検索アシスタント[10] に利用している．

(3) 通信パケット内容の解析（network logs）
Abrams らの研究では，内部で tcpdump を使用し HTTP のポートの通信内容をモニタする httpfilt や，任意のポート番号の HTTP の通信内容をモニタする httpdump を作成して，通信内容をモニターしている．

(4) Web サーバのログファイルの利用（server logs）
Web サーバの分析ツール開発を容易にするために，ログファイルの共通形式として common logfile format（CLF）が定義されている[11]．CLF は apache をはじめとする多くのサーバでサポートされ，wwwstat など多くのツールが開発されている[12]．

特に次の点に注目し，プロキシサーバのログファイルを解析する手法を用いる．

- 広範囲なユーザの獲得
Web ブラウザの動作プラットフォームやバージョンに依存しないモニタ手法が望ましい．Web ブラウザを改造する手法では，改造したプログラムをユーザに常用してもらうのは難しいだけでなく，改造作業の負荷が無視できないほど大きい．
- 目的と情報収集範囲の妥当性
たとえば，Web サーバのログファイルは自分自身へのアクセスしか記録できないので，操作シーケンスに深刻な欠落を起さないようにログファイルを集めるのは難しい．また，通信パケットの収集は物理的に同一のネットワークセグメント内に限定される．これに対して，プロキシサーバは複数のネットワークセグメント間で共有され，広範囲の情報収集が容易である．

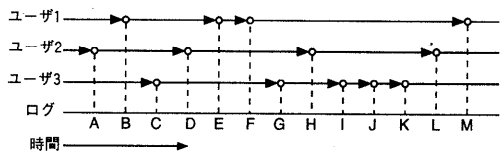


図 2 ユーザのアクセスとログ
Fig. 2 User's access sequences and logging.

ただし，大部分の Web ブラウザが使用しているローカルキャッシュのために，プロキシサーバでは操作シーケンスの一部が観測できずにバイアスがかかる．ローカルキャッシュによって欠落する主要な操作はバック操作と考えられるが，Taucsher らの Web ブラウザの操作履歴を使った revisitation pattern の研究では，バック操作は履歴の約 30%を占めており，このように多用される主な要因は，ここで分類されている Hub-and-spoke, Guided tour, Depth-first search の 3 種類の情報閲覧パターンのうち，特に Hub-and-spoke と Depth-first search 閲覧パターンで新しいページに移動するために一度元のページに戻るヒストリ機構として使用されているためと解析されている[8]．このような場合に新しいページに直接関係づけることは妥当であり，またそのように操作シーケンスがフィルタリングされる方が人間の情報閲覧の意図がより的確に観測できることになる．

4.2 アクセスシーケンスとログファイル

図 2 に，ユーザのアクセスシーケンスと，実際にログファイルに記録されるシーケンスの関連を示す．

ログファイル中にはユーザ 1（B → E → F → M），ユーザ 2（A → D → H → L），ユーザ 3（C → G → I → J → K）の独立したアクセスが，直列化されて記録される．

この直列化されたアクセスシーケンスをユーザ単位で追跡するために，ログファイル中に同時に記録される IP アドレスを元にアクセスシーケンスを分離する．

4.3 HTML 文書の 1 ページに対する縮退処理

HTML 文書は，1 つの文書であっても複数の HTML ファイルや画像ファイルで構成されることが多いので，ログファイル中では複数のファイル獲得シーケンスに展開される．

このようにプロキシに記録されるログの例を**図 3** に示す．この例では，`http://www.apple.com/` へのアクセスが，GIF ファイルなどを含む複数のファイルへ

```
890376962.453   1865 163.138.96.44 TCP_MISS/200 6097 GET http://www.apple.com/ - DIRECT/www.apple.com text/html
890376962.508    708 163.138.96.44 TCP_REFRESH_HIT/200 290 GET http://www.apple.com/main/elements/spacer.gif - DIRECT/www.apple.com image/gif
890376962.531    734 163.138.96.44 TCP_REFRESH_HIT/200 290 GET http://www.apple.com/main/elements/spacer.gif - DIRECT/www.apple.com image/gif
890376963.183   1384 163.138.96.44 TCP_MISS/200 3373 GET http://www.apple.com/main/elements/apple.gif - DIRECT/www.apple.com image/gif
890376963.323     93 163.138.96.44 TCP_HIT/200 3549 GET http://www.apple.com/main/elements/navbar.gif - NONE/- image/gif
890376964.195   1607 163.138.96.44 TCP_MISS/200 4330 GET http://www.apple.com/home/images/ticker.gif - DIRECT/www.apple.com image/gif
890376964.889   2222 163.138.96.44 TCP_MISS/200 12111 GET http://www.apple.com/home/images/promos/toasted.gif - DIRECT/www.apple.com image/gif
890376965.503   2086 163.138.96.44 TCP_MISS/200 9922 GET http://www.apple.com/home/images/promos/thinknow2.gif - DIRECT/www.apple.com image/gif
890376965.684    724 163.138.96.44 TCP_MISS/200 273 GET http://www.apple.com/home/images/blackline.gif - DIRECT/www.apple.com image/gif
890376966.185   1915 163.138.96.44 TCP_MISS/200 6726 GET http://www.apple.com/home/images/promos/thinkfirst2.gif - DIRECT/www.apple.com image/gif
890376970.471   7950 163.138.96.44 TCP_MISS/200 65345 GET http://www.apple.com/home/images/toasthome.gif - DIRECT/www.apple.com image/gif
```

図 3 ログファイルの記録例
Fig. 3 Example of a log file.

の HTTP リクエストとしてログファイルに記録されている．

これは非本質的な展開であり，本来ひとまとめとして扱うべきものであるので，以下の方針でアクセスシーケンスの縮退処理を行い，本質的な情報を含む URL に集約する．

4.3.1 プロトコルの制限

ログファイル中には，HTTP の履歴以外に，プロキシサーバのキャッシュの通信の履歴や FTP などの他のプロトコルの履歴も含まれている．HTML 文書の相関関係だけに注目するために，HTTP リクエストだけを抽出する．

4.3.2 MIME タイプの制限

ログファイルには HTTP リクエスト単位で記録されているので，HTML ファイルで使用されている画像データやアプレットは別々に記録される．そこで MIME タイプを元に，テキストファイルと HTML ファイルだけを抽出する．

ただし，ステータスコードが 200（OK）の場合にはファイルの MIME タイプが正しく付加されてくるが，304（Not Modified）の場合にはファイルの MIME タイプではない “text/html” を返してくるので，ステータスコードが 200 のときの MIME タイプを記録し，304 のときにはそれに基づいて判定している．

4.3.3 ファイル拡張子の制限

たとえば，HTTP サーバが “.class” 拡張子を持つ Java クラスファイルに対応していない場合は，誤った “text/plain” を MIME タイプとして返すことがある．そのような例外に対応するために，拡張子をチェックしている．

4.4 文書全体に対する縮退処理

HTML 文書という意味的にまとまりのある単位で扱う場合に，HTML ファイルの集合の表現方法，HTML 文書の境界のあいまいさ，そして 1 つの HTML 文書が複数のサブ HTML 文書を含む構成の扱い方などが問題になる．

ここでは，HTML 文書をある URL で代表させる．たとえば，他の HTML 文書からのリンクや，ブックマーク，リンク集，ディレクトリサービスには，情報提供者がある HTML 文書を代表するのにふさわしいと認識している URL をリンクとして登録しているはずであり，たいてい情報閲覧者はその URL から HTML 文書の閲覧を開始する．

このように文書全体を代表する URL として，ユーザは HTML 文書を代表する URL に一番最初にアクセスするという経験則から，アクセスシーケンス中で Web サーバが切り替わった最初の URL を使用する．

たとえば，今回使用したプロキシサーバのログで，比較的頻繁にアクセスされている `java.sun.com` を調べたところ，129 個ある HTML ファイルやテキストファイルの URL が 34 個にまとめられた．この中には，最後が “/” で終了している URL が 9 個，“index.html” で終了している URL が 8 個あり，HTML 文書を代表する URL としてふさわしくないと考えられる URL は 5 個であった．これらは，複数の並行に進行しているアクセスシーケンスが交錯したか，サーチエンジンの出力結果を閲覧したのが原因と考えられるが，この経験則は比較的妥当だと考えられる．

5. 評　　価

5.1 評価について

HTML 文書関連グラフや HTML 文書間の相関度の性質を分析するために，あるグループが使用している Squid の 1997 年 9 月 1 日から 1998 年 2 月 28 日まで半年分のログデータを評価に使用した．

評価対象のグループは分散システムの基礎研究グループであり，分散アルゴリズムや通信プロトコルの設計から，プロトタイプシステムの実装まで行っている．仕事は分業化されており，各自の専門分野においては高度な知識を有するが，必ずしも他人の分野に詳しいわけではない．総ユーザ数は 4 人であり，うち 3 人は同じプロトタイプシステムの実装を行っている．

プロキシサーバにアクセスしたクライアントの総数は 14 台である．この中には，キャッシュ通信を行っ

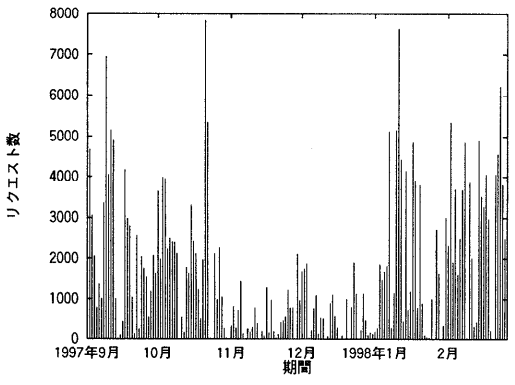


図 4 リクエスト数の推移
Fig. 4 Transition of requests.

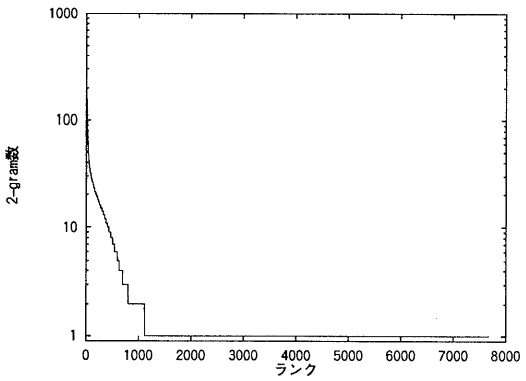


図 5 2-gram の分布
Fig. 5 Distribution of 2-grams.

ている Squid も 1 台含まれる．評価対象の環境では，ユーザは計算機を 1 人 1 台以上使用しているが，他人と共有使用することはほとんどない．

この期間中に発行された HTTP リクエストを分析すると，半年間の総リクエスト数は 294858 件であり，1 日の平均リクエスト数は 1647 件，最高リクエスト数は 7831 件，最小リクエスト数は 0 件であった．この期間の 1 日のリクエスト数の推移を**図 4** に示す．休日があるにもかかわらず，1 人あたり 1 日平均 400 リクエスト以上であり，評価対象のグループは，Web 情報閲覧の依存性がかなり高いと判断できる．

5.2 2-gram 分布

このログファイルを処理して得られた 2-gram 数の合計は 22356 個，7678 種類である．この結果を，出現数が多い順番に順序付けした分布を，**図 5** に示す．出現数が 1 回の 2-gram は 6560 種類であり，これは全体の 85.4%に相当する．このような低頻度の 2-gram は，それがアクセスシーケンス中の切れ目を表すのか，頻度が低いだけで重要なシーケンスを表すか正しく判断するのは困難である．

次に，各クライアントの特徴を調べるために，最初に各クライアントの 2-gram の総数を計算し，クライアントを順序付けする．次に，その順番に基づいて 1 番目のクライアントの 2-gram 数で 2-gram をソートし，同一数の場合に限り 2 番目，3 番目と繰り返しソートしていくマルチレベルのソートを行って各 2-gram のランクを決定する．使用頻度の高い上位 4 クライアントに対して，この手法によって得られた各クライアントごとの 2-gram 数の分布を**図 6** に示す．なお，この上位 4 クライアントは，4 人のユーザが主に使用している計算機である．

図 6 から，類似の仕事内容でありながら，各ユーザのアクセスパターンはかなり異なることが観測できる．各ユーザのアクセスが分離している部分は興味領域が異なることを示し，2-gram 数が多ければ，その興味領域に対して専門的な知識を有すると考えられる．逆にアクセスが重なっている部分は，興味領域が同じことを示し，2-gram 数の大きいパターンを持つユーザの方が，より専門的な知識を有すると考えられる．

5.3 相関度分布

図 7 に，最終的に得られた相関度と 2-gram の出現頻度の関係を示す．グラフ中の線は，$P(u_x, u_y)$ の値域が 0 から 1 の間であることから決定される，出現頻度と論理的な相関度の上限値の関係を示す．

5.3.1 自動巡回プログラムの影響

今回テストしたネットワーク環境では，AutoNews[13] と NewsCast[14] という 2 種類の Web サーバを自動巡回するプログラムを使用している．図 6 で類似形のパターンを持つ部分は，これらの自動巡回プログラムの影響も大きい．これらの自動巡回プログラムは，それぞれ巡回先 URL のリストを持っているので，巡回先リストに含まれている 2-gram を抽出した．

図 8 に，AutoNews の相関度分布を示す．AutoNews は，JavaScript で Web ブラウザに巡回表示をさせる簡単なプログラムであり，巡回順序は変化しないが，巡回の停止，直前の URL または次の URL への移動ができる．この巡回戦略は単純なので，全体的に非常に高い相関度を示す．

図 9 に，NewsCast の相関度分布を示す．AutoNews は，巡回先が分散管理され，動的に変化可能なシステムであり，プログラム内部で保持した巡回先リストに対して，ユーザ操作情報を利用したプライオリティ付け，巡回経路のランダム化を行うプログラムである．巡回経路が多様化するために，AutoNews に比べて相関度はかなり低くなる．

図 8 と図 9 から，このような自動プログラムはどう

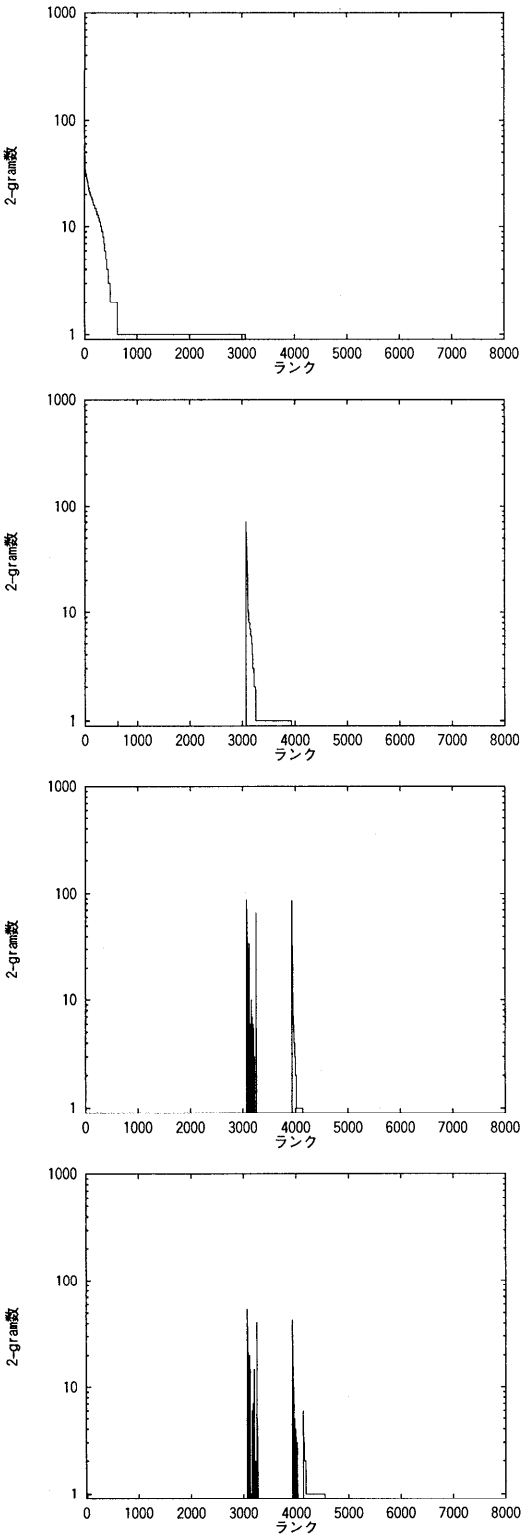


図 6 各クライアントの 2-gram の分布
Fig. 6 Distribution of 2-grams for each client.

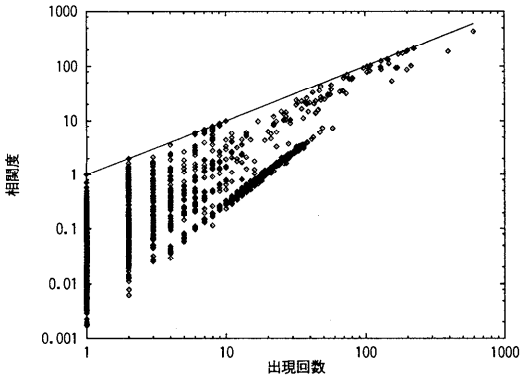


図 7 相関度分布
Fig. 7 Distribution of correlation.

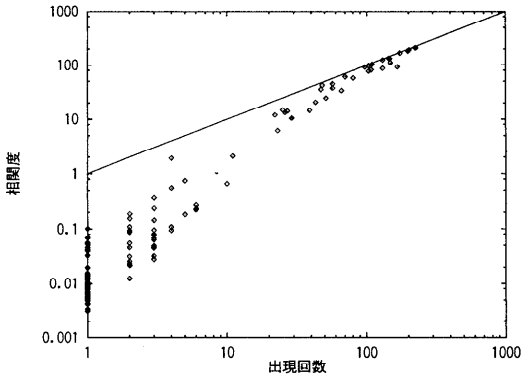


図 8 AutoNews の相関度分布
Fig. 8 Distribution of correlation that has relevance to AutoNews.

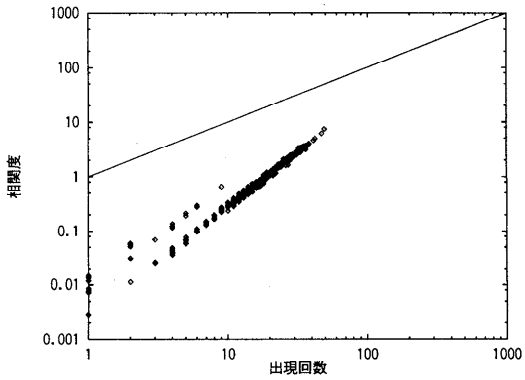


図 9 NewsCast の相関度分布
Fig. 9 Distribution of correlation that has relevance to NewsCast.

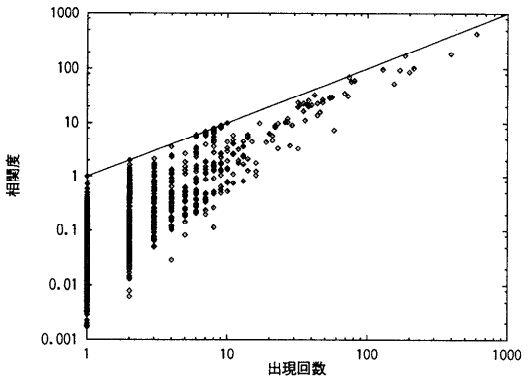


図 10 AutoNews と NewsCast の影響を除外した相関度分布
Fig. 10 Distribution of correlation that have no relevance to AutoNews and NewsCast.

しても人間よりもアクセス頻度が高くなるので，ユーザの相関度よりもかなり優位になる．

また，これらの自動プログラムは，人間が繰り返し行う作業を自動化しただけであるが，人間に比べると動作が規則的なので，人間の場合とは違うそれぞれの動作アルゴリズムを反映した特徴的な相関度分布パターンを示すことが分かる．

そこで，自動プログラムの影響を除外して，人間の操作部分だけを検討するために，2 つのプログラムに関連する 2-gram を除外すると，2-gram 数は 7125 種類に減少する．**図 10** に，この AutoNews と NewsCast の影響を除外した相関度分布を示す．

5.4 パスの双方向性

HTML 文書関連グラフ中には，片方向パスと両方向パスが混在し，2-gram の順不同の組合せ数は，全部で 7166 種類あり，片方向パスを持つのが 6166 種類，両方向パスを持つのが 512 種類である．さらに，512 種類から自動巡回プログラムに関連する組合せを除くと，299 種類になる．

片方向のパスが圧倒的に多いのは，ブラウザキャッシュの影響でミクロなリンクの往復などは観測されずマクロな情報参照方向だけが観測されること，HTML のリンクの指定が片方向であることが主な原因であると思われるが，情報閲覧者が情報間に何らかの格差や優先度を認識していることも考えられる．

5.5 高相関度の 2-gram の評価

残された 2-gram の組には，ノイズ情報も数多く混じっていると思われるので，次の式を満たす高相関度を示す 2-gram だけを抽出した．

$$E(u_x, u_y) > 1$$

高相関度を示す 2-gram の数は 232 個であり，これは全体の 3.3%にあたる．**表 1** に，これを分類した結果を示す．すでに削除されていたり，JavaScript プログラムが異常動作する，パスワードの入力を要求されアクセスできないなどの問題を持つ 2-gram が 36 個存在した，また，リダイレクトを表す 2-gram は 16 個あった．

表 1 2-gram の分類
Table 1 Classification of 2-grams.

内容	個数（パーセント）
相関関係がある 2-gram	111 (47.8%)
相関関係がない 2-gram	2 (0.8%)
Frame ごとに分解	67 (23.9%)
リダイレクト	16 (6.9%)
情報の削除	30 (12.9%)
プログラムの異常動作	5 (2.2%)
パスワード要求	1 (0.4%)
合計	232

なお，HTML ファイル中で Frame を使用していると，それを同一の URL に縮退することができない問題がある．たとえば，Frame を使って商用サーバで画面の一部に CM を入れる場合や，別のサーバのサービスや情報を参照する場合には，提供されるサーバが異なる場合には，情報閲覧者にとって無意味な高相関度を示す 2-gram が数多く生成されている．

これらを除外して，何らかの相関関係があると判断できた 2-gram は 111 個であった．ただし，この相関関係は多様であり，たとえば最新情報や検索システムなどの情報の種類による分類と，釣り情報のように情報の内容による分類が見られた．

ただし，相関関係があるとは判断できない 2-gram は 2 個発見された．1 つは，アクセスシーケンスの切替わりであった．これは出現頻度が 2 回であり，相関度は今回採用した閾値をわずかに上回る 1.3 であった．もう 1 つは，ある特定のメンバだけがよくアクセスするシーケンスであった．

以上のことから，次のように判断される．

- 相関度は，ある HTML 文書と相関関係を持つ HTML 文書群を順位付けする指標として使用できる．
- 相関度は，高頻度の 2-gram に対しては良い特性を示す．逆に，低頻度の 2-gram の場合には判断が難しいので，データ量が多い方が良い．
- 観測母集団が小さい場合には，特定のユーザの操作が顕著に現れることがある．
- Frame の処理など，HTTP リクエストを観測するだけでは処理できない問題がある．Web Robot で収集した情報を同様にデータベース化して併用するなどの対策が必要である．

- 情報が移動・削除される率は比較的高い．情報の動的な変化に対応するためには，観測周期をより短くする必要がある．

6. おわりに

本論文では，プロキシサーバのログファイルから，ユーザのHTML文書単位のアクセスシーケンスを分析し，そのURLの2-gramから，あるコミュニティ中のHTML文書の相関関係を示すHTML文書関連グラフを再構成する手法を提案した．このHTML文書関連グラフは頻度情報を持つ有向パスで構成され，そのグラフ構成と頻度から関連度を定義した．

さらに，実際に小規模なコミュニティにこの手法を適用し，得られた関連度を分析して，その有効性を示した．

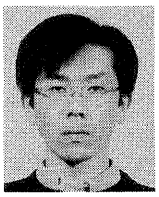
参考文献

1) Raggett, D.: *HTML 3.2 Reference Specification*, World Wide Web Consortium (1997).
2) Francis, P., Kambayashi, T., Sato, S. and Shimizu, S.: Ingrid: A Self-Configuring Information Navigation Infrastructure, *4th International World Wide Web Conference*, Boston, World Wide Web Consortium (1995).
3) Shimizu, S., Kambayashi T., Sato, S. and Francis, P.: A Framework for Multilingual Searching and Meta-information Extraction, *INET 97*, Kuala Lumpur (1997).
4) 丹羽芳樹，新田義彦：単語ベクトルを用いた多義語の意味推定—共起ベクトルと定義距離ベクトルの比較，自然言語処理，Vol.102, No.7, pp.49–56 (1994).
5) Abrams, M. and Williams S.: Complementing Surveying and Demographics with Automated Network Monitoring, *World Wide Web Journal*, Vol.1, No.3 (1996).
6) Cunha, C.R., Bestavros, A. and Crovella, M.E.: Characteristics of WWW Client-based Traces, Technical Report, BU-CS-95-010, Boston University (1995).
7) Catledge, L.D. and Pitknow, J.E.: Characterization Browsing Strategies in the World-Wide Web, *3rd International World Wide Web Conference*, Darmstadt, World Wide Web Consortium (1995).
8) Tauscher, L and Greenberg, S.: Revisitation Patterns in World Wide Web Navigation, *CHI '97*, Atlanta, ACM (1997).
9) 佐藤進也，風間一洋，清水　奨：アクセス履歴を利用したWebサーバの状態の推定，Japan World Wide Web Conference '97，横浜，日本インターネット協会 (1997).
10) 清水　奨，神林　隆，佐藤進也，風間一洋：グループ指向WWW検索アシスタントPA-searchの実現，*Japan World Wide Web Conference '97*，横浜，日本インターネット協会 (1997).
11) Nielsen, H.F.: *Logging Control In W3C httpd*, World Wide Web Consortium (1995). http://www.w3.org/Daemon/User/Config/Logging.html
12) Fielding, R.: *wwwstat – HTTPd Logfile Analysis Software* (1998). http://www.ics.uci.edu/pub/websoft/wwwstat/
13) 原田昌紀：*AutoNews 0.2* (1997). http://www.graco.c.u-tokyo.ac.jp/ harada/autonews/0.2/
14) 風間一洋，佐藤進也，清水　奨：Ingrid NewsCast—自律型ニュース配信システム，*SWoPP '97*，熊本 (1997).

(平成10年5月8日受付)
(平成11年2月8日採録)

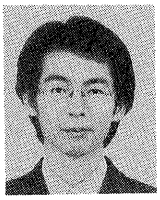
風間　一洋（正会員）

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．



佐藤　進也（正会員）

昭和38年生．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．電子情報通信学会，Internet Society，ACM各会員．



清水 奨（正会員）

平成4年東京大学工学部機械情報工学科卒業．同年日本電信電話（株）入社．現在NTT未来ねっと研究所．情報システムの研究に従事．Internet Society，ソフトウェア科学会，ACM 各会員．

神林 隆（正会員）

平成元年慶応義塾大学大学院理工学研究科修士課程終了．同年日本電信電話（株）入社．現在日本テレマティーク（株）に出向．