7 大規模Webアーカイブからのデータマイニング

豊田 正史（東京大学生産技術研究所・戦略情報融合国際研究センター）
mtoyoda@acm.org

喜連川 優（東京大学生産技術研究所・戦略情報融合国際研究センター）
kitsure@tkl.iis.u-tokyo.ac.jp

Webの発展過程マイニング

Webは膨大な文書の集合であると同時に，ハイパーリンクで結合された文書の巨大なネットワークでもある．少なくとも80億以上存在する☆1Webページは，生成，更新，および消滅の過程を経て日々変化しており，それに応じてWebのネットワーク構造も動的に変化を続けている．こうしたWebの発展過程は実世界の動向と密接な関係を持つ傾向を強めつつある．たとえば，テロのような重大な事件が発生すると，それに関したWebページが次々と作成され，重要な情報には多くのページからハイパーリンクが張られていき，多数のユーザが訪れるようになる．インターネット人口の増大，およびblog等の誰でも使える簡便な出版ツールの普及が，この傾向にますます拍車をかけている．今後，Webは実世界の事象をより広範囲かつ迅速に反映するようになると予想される．

Webの発展過程を把握することは，実社会で起きる事象の背景や予兆を探る上で重要な課題であり，以下のような状況で有用である．

1．Webにおけるトピックの履歴に関する質問に答える．たとえば2001年9月11日のアメリカでのテロ事件について，関連するページがどの程度作られてきたか，など．
2．新たな情報の発生を監視または観察し，トレンドを分析することで，市場調査などに応用する．また，新しい情報が頻繁に生まれる場所を特定することは，検索エンジンを新鮮に保つためにも有用である．
3．Webにおける社会学的な現象とその推移を調査する．また，実社会における社会学的現象がWebにどのように反映されるかを調査する．

しかし，現状の主要な検索エンジンでは，最新のページをいかに検索するかに焦点が当てられており，こうした過去を紐解くような調査方法はほとんど提供されていない．

動的に変化するWebの発展過程をマイニングするためには，定期的にWebページを大規模に収集して蓄積するWebアーカイブが必要である．収集したページの履歴をすべて蓄積することで，ページの発生，更新および消滅を観測することが可能となる（ただし現状では変化の速度が，収集の速度を遥かに上回るため，完全な観測はアーカイブをもってしても不可能である）．アーカイブは通常，ハイパーリンクを辿りながらWebページを収集するクローラ☆2と呼ばれるソフトウェアを定

☆1 2004年11月時点においてGoogle (www.google.com) で検索できる文書数．
☆2 スパイダー，ロボットと呼ばれる場合もある．

期的に運用することで構築される．現在公開されている中で最大のWebアーカイブはInternet Archive (www.archive.org)である．1996年からWebページの収集を始め，延べ300億ページ以上を蓄積している．

Webアーカイブを用いた発展過程のマイニングは，図-1に示すように，定期的に収集したWebのスナップショットを比較しどのような変化がどの程度起きたかを調査することが基本となる．本稿では，以下に示す2種類のマイニングの試みを紹介する．

さまざまな変化量の測定　検索エンジンの索引を新鮮に保つためには，更新の量や頻度の多い部分を集中的に収集しなくてはならない．クローラを効率化することを目的に，Webのさまざまな変化量の測定に基づいて，将来の変化を予測可能かどうかが調査されている．

トピックの発展過程　Webにおけるトピックの出現を検出し，その発展過程を追跡する．テキスト解析に基づく方法と，リンク解析に基づく方法を紹介する．

■さまざまな変化量の測定

Webの検索エンジンは，日々変化するWebページを検索可能にするために，索引を常に最新に保つ必要がある．しかしすべてのページを高頻度で(たとえば毎日)収集し続けることは不可能である．このため，検索結果に影響を与えるような顕著な変更が，高い頻度で行われるページの傾向を把握し，そのようなページを集中的に収集できるようにすることが重要となる．効率的なページの収集を目指して，さまざまな変化量の測定が行われている．

各ページが変更される度合

Fetterly，Manasse，Najork，Wienerは，大規模なクローリングを行って15億ページを収集し，以降1週間ごとに11週間にわたってこれら15億ページを繰り返し収集することで，各ページの変化の度合を解析した[1]．以下に示す結果から分かるように，この解析からページの未来における変化度がある程度予測できることが明らかになった．

- クロールごとに取得できなくなるページの数は増加する．11週間後には10数％のページが取得できなくなっていた．このうち，サーバ管理者の設定するロボットルールによって排除されたケースも3％ほど存在する．これは，Webの観測行為が観測される側の行動を変化させ得ることを示している．

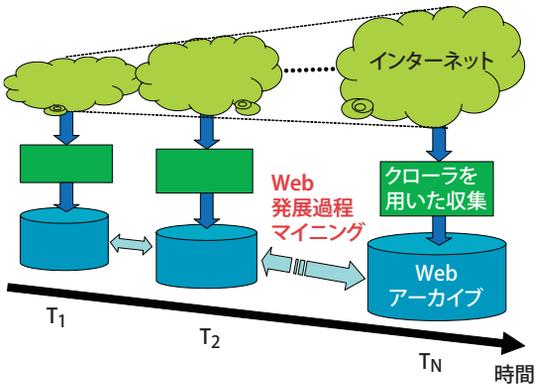

図-1　Webアーカイブを用いた発展過程マイニング

- ページの内容は1週間ではほとんど変更されない．任意の時点で2週間続けて取得できたすべてのページ対について変化を調べたところ，65%の対で更新がなかった(全体のチェックサムが一致した)．変更があったとしても，HTMLマークアップのみの変更か，些細な変更がほとんどであった．この結果から，顕著な変化の起こるページを予測することの重要性が確認できる．
- サイズの小さいページより大きなページの方が変化の度合が大きい．これはページのサイズが，未来の変化度を予測する手がかりとなることを示している．
- 各ページの過去における変化の度合は，未来における変化の度合と相関がある．すなわち，前の週に変化の小さいページは次の週も変化が小さく，大きいページは，次の週も変化が大きい．したがって，過去における変化度も未来の変化度を予測する手がかりとなる．

解析に用いたデータセットの大きさは最初のクロールだけでも6.4TBにのぼるため，テキストのままで各ページの変化度を測定するのは困難である．ここでは，大規模なデータセットからの変化度の測定手法を簡単に紹介する．

ページの比較を容易にするため，各ページは，ページ内容のチェックサム，および84の整数からなる特徴ベクトルに要約される．チェックサムを比較することでページに変化があったかどうかが分かる．特徴ベクトルは，2つのページの特徴ベクトルが共有する要素の数が多いほど内容が一致していると見なせるように定義しており，以下の手順で抽出する．

- ページ内のHTMLタグを除去して残った内容を単語の列と見なし，固定長の連続する部分列をすべて抽出する．実験では5語からなる部分列を用いている．この部分列はシングル(shingle)と呼ばれる．
- 各シングルのチェックサムを算出しその中から，84

のチェックサムを選択し特徴ベクトルとする．ただし，同じチェックサムの集合からは，必ず同じチェックサムが選ばれるような選択手法を用いている（詳細は文献1）を参照されたい）．

この手法を用いると，数値の比較のみでページの変化度を高速に測定できる．しかし，URL，チェックサム，特徴ベクトルからなるログのみでも規模は大きい．11クロール分全体では1.2TBとなり，各URLの時間ごとの差分データのみを抽出しても222GBとなる．222GBのデータは読むだけでも10時間☆3程度かかるため，データを読みながらさまざまな解析を同時並列に行えるようなシステムを作成して解析を行っている．

新しい情報の出現する度合

Fetterlyらの手法は，URLの集合を固定して繰り返し収集するため，新しいページの出現は検知できない．Ntoulas，Cho，Olstonは，Google Directoryから著名な150サイトを選択して固定し1週間ごとにそれらのサイトの中身を取り尽くすことで新規ページや新規ハイパーリンクの発生する度合を調査した[2]．Ntoulasらが1年分（52週間）のアーカイブから得た結果を以下に示す．

- 選択したサイト内のページは，速いペースで新しいページと入れ替わっている．1週間ごとのアーカイブ中，平均で8%が新しく作られたページである．また初回に取得できたページのうち半数は，半年後には取得できなくなり，1年後には60%が取得できなくなっている．
- ページのコンテンツの入れ替わりは，ページ自体の入れ替わりよりも緩やかである．全ページから全シングル（本実験では50語）を抽出し，ユニークなシングル数の推移を調べたところ，1週間ごとに5%の新しいシングルが加わっており，1年後には50%のシングルが生き残っていた．これは新規ページに必ずしも新しいコンテンツが含まれているとは限らないことを示している．
- ハイパーリンク数の入れ替わりは，ページの入れ替わりよりさらに急で，1週間ごとに25%のリンクが新規に作成されており，1年後には80%のリンクが消滅している．
- ページの内容の変化については，概ねFetterlyらと同じ結果が得られている．すなわち，ページの内容はほとんど変化しておらず，過去の変化の度合と未来の変化の度合に相関があることを確認している．

これらの結果は，あくまで著名な150サイトに関する結果である．論文中には，これらの統計をそのままWeb全体に敷衍するような記述が見受けられるが，正しい推計にはなっていないと思われる．たとえばページの消滅する割合は，Fetterlyらの結果よりも急である．

トピックの発展過程

単語の出現頻度からの発展過程抽出

Internet Archive（www.archive.org）は，1996年から継続的にWebページを大規模に収集し続けている．2003年9月の時点で延べページ数は約300億に達している☆4．このアーカイブは，ページが更新されても古いバージョンのページを残しておくため，ページ更新の履歴を見ることができる．2001年には，URLを入力するとそのURLの変更の履歴を閲覧できる，Wayback Machine（web.archive.org）インタフェースを公開している．

Recall（recall.archive.org）は，Internet Archiveの上に構築された全文検索エンジンである[3]．1996年から2003年5月までに収集された約110億ページ（0.5ペタバイト）を索引化し，2003年9月にβ版として公開された．

Recallは単なる全文検索エンジンではなく，アーカイブから自動的にカテゴリーおよびトピックを抽出してトピックの発展過程を調査できる機能を持っている．キーワードを指定して検索を行うと，そのキーワードが出現する頻度の時系列的な変化をグラフで見せると同時に，関連するキーワードの出現頻度もグラフ化して提示する．これらのキーワードの動きからトピックの変遷を調査することができる．

図-2に，“terrorism”というキーワードで検索を行った結果を示す．中央右寄りの赤い線のグラフが，月ごとのキーワードの出現頻度を表している．2001年の後半付近に最大のピークが見られ，9月11日の同時多発テロに対応していることが分かる．左側には，関連したトピックを表す語句の出現頻度を表すグラフが表示される．ここでは，“Counter Terrorism”や“Prevention of Terrorism”などが関連トピックとして自動抽出され

☆3　解析にはCompaq DS20サーバ4台からなるクラスタを使用している．
☆4　同じページを複数回収集していることに注意されたい．

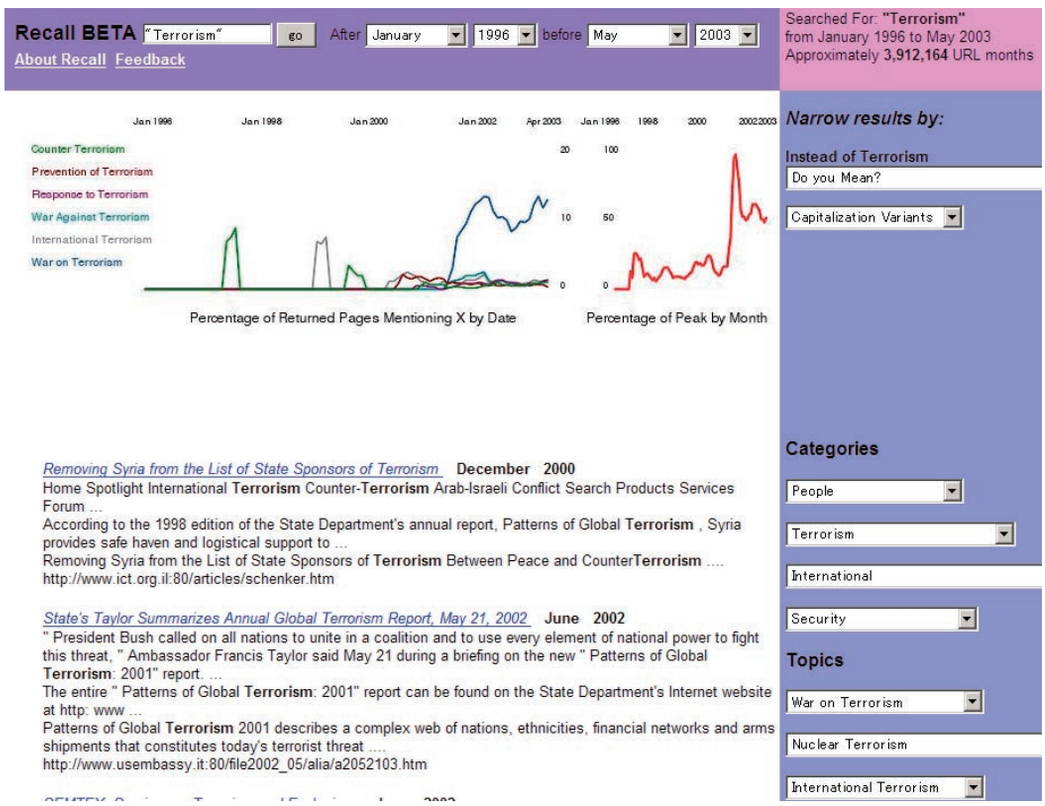

図-2　Recallを用いた検索例

その頻度が表示されている．2001年の後半に“War on Terrorism”というトピックが急激に成長しており，テロ直後のブッシュ大統領の演説に対する反応の激しさをうかがい知ることができる．

また，図-2の右側には，人名などのカテゴリーや，テロに関連するトピックを選択するドロップダウンボックスが用意されており，これらを選択することで，検索の範囲を絞ることができる．これらのカテゴリーやトピックは，アーカイブを事前に解析することで自動抽出されている．1,400,000の語句が50,000のカテゴリーに自動分類されており．カテゴリーに含まれる語句数の中央値は12となっている．カテゴリーおよびトピックの抽出手法については，現在のところ明らかにされていない．

リンク構造からの発展過程抽出

Web上では，互いに関連するページは比較的多くのリンクで結合されており，リンクによるネットワークにおいて近くに存在している傾向がある．理由として，同種類のページへのリンクを集めたリンク集が数多く作成されていること，Webページの作者は自分のページに関連する情報を持つページにリンクを張る傾向があること，が挙げられる．

関連ページ同士はしばしばハブおよびオーソリティからなるリンク構造のパターンを持ち，Webコミュニティ

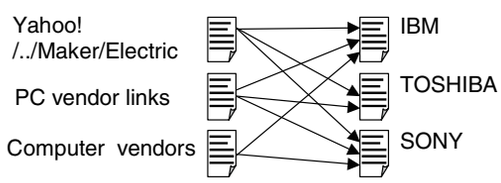

図-3　典型的なWebコミュニティのリンク構造

と呼ばれる．コミュニティの典型的な例を図-3に示す．右側にある大手コンピュータメーカのようによくリンクされているページがオーソリティで，左側にあるリンク集のようにオーソリティに多くリンクを張っているのがハブである．ハブからオーソリティへの密なリンク構造がコミュニティを形成する．このリンクのパターンはさまざまな分野のページ間で共通して見ることができる．1997年にKleinbergが単純な繰り返し計算によるコミュニティ抽出手法を提案し，その後さまざまな発展形が提案されている．

Webコミュニティは何らかのトピックを表すため，新しいトピックがいつ発生して，どのように発展したかを，コミュニティを単位として把握することができる．著者は，定期的に収集した各スナップショットから主要なコミュニティをすべて抽出してそれらの相関図を構築し，アーカイブ間でコミュニティの比較を行うことで時系列

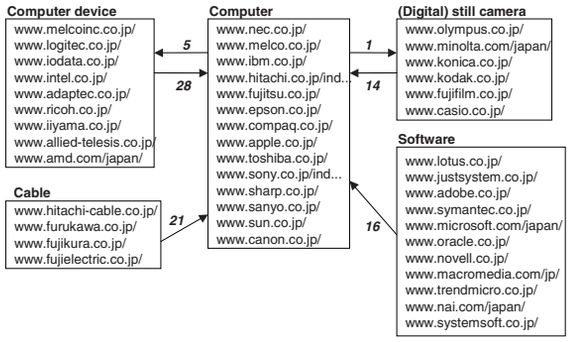

図-4　コンピュータメーカを中心としたコミュニティチャートの一部

図-5　発展過程抽出システム

的変化を抽出する手法を提案した[4]．この手法に基づき，1999年から継続的に収集を続けている日本のWebアーカイブを用いて，コミュニティの発展過程抽出システムを構築している．

コミュニティの相関図は，Webコミュニティチャート[5]と呼ばれる有向グラフとして表現される．互いに素なページの集合であるコミュニティをノードとし，関連するコミュニティの間に重み付きの有向辺を持つ．チャート作成手法は，まずWeb全体から被リンク数の多いページ群を選んでシードセットとする．次に，各シードをオーソリティと見なし，ハブとオーソリティの関係で密に結ばれているシード同士をコミュニティとして分類する．最後に，2つのコミュニティのメンバ同士がハブとオーソリティの関係を持つ場合，それらのコミュニティ間に辺を作成する．分類のアルゴリズムの詳細については，文献5）を参照されたい．

図-4に，作成したコミュニティチャートの一部を示す．中央に大手コンピュータメーカのコミュニティがあり，その周りに関連するコミュニティとして，ソフトウェア，周辺機器，ディジタルカメラなど関連業種の会社のコミュニティが抽出されていることが見て取れる．

コミュニティの発展過程は，各スナップショットごとに作られたチャートを前後の時間のものと比較することで抽出する．たとえば，ある時間tのチャート(C_t)に含まれるコミュニティ($c_t \in C_t$)の変化を未来方向に追跡するには，次の時間において対応するコミュニティ(c_{t+1})を同定しなくてはならない．同定には，コミュニティ間で共有されているページの数を利用する．すなわち，時間$(t+1)$においてc_tと共有するページ数が最も多いコミュニティを，c_{t+1}とする．

対応するコミュニティが同定できれば，c_tとc_{t+1}を比較することで，c_tから消滅または分裂したページ数，c_{t+1}に合併または新規に現れたページ数を計算することができる．これらを基に，各時点での各コミュニティについて成長率，新規率，安定率など，興味ある発展過程の抽出に有用なメトリックスを算出する．これらのメトリックスを用いると，ある時点で最も成長したコミュニティや，最も新しいコミュニティなどを抽出することが可能になる．筆者は，コミュニティの発展過程についてさまざまな統計をとっているが，詳細については文献4）を参照されたい．

筆者は，こうした発展過程を柔軟に閲覧できる発展過程抽出システムを開発している(**図-5**)．表示する情報量が膨大になるため，高解像度の壁面ディスプレイを用いている．右側で最新のコミュニティチャートのグラフ表示を行い，左側で時系列的な発展過程を表示する．中央にはコミュニティに含まれる複数のページを同時に閲覧できるブラウザを配置してある．ここでは左側の発展過程のビューアを詳細に説明する．

発展過程ビューアは，与えられたキーワードやURLによるコミュニティの検索，指定したコミュニティの発展過程の表示，および発展のメトリックスを用いたコミュニティのソートといった機能を提供する．**図-6**は，「テロ」というキーワードでコミュニティを検索した結果を示している．コミュニティの変遷が左から右へ時間順に表示されている．各列はアーカイブを収集した時期に対応しており，上部にその時期が示されている．各矩形がコミュニティを表し，中にはページのリストが含まれている．コミュニティを結ぶ線は，その間でページが共有されていることを示し，太さが共有ページ数を表す．図-6からは，Recallの例(図-2)と同様に，同時多発テロの直後に爆発的にさまざまな種類のコミュニティが発生していることが分かる．主なところでは，テロに関するニュースの集まり，義援金を募集する団体の集まり，報復攻撃に反対する平和団体などの発生を見ることができる．このように，どのような種類のページが，どのくらい現れ，それがどう発展したかを概観することが可能である．

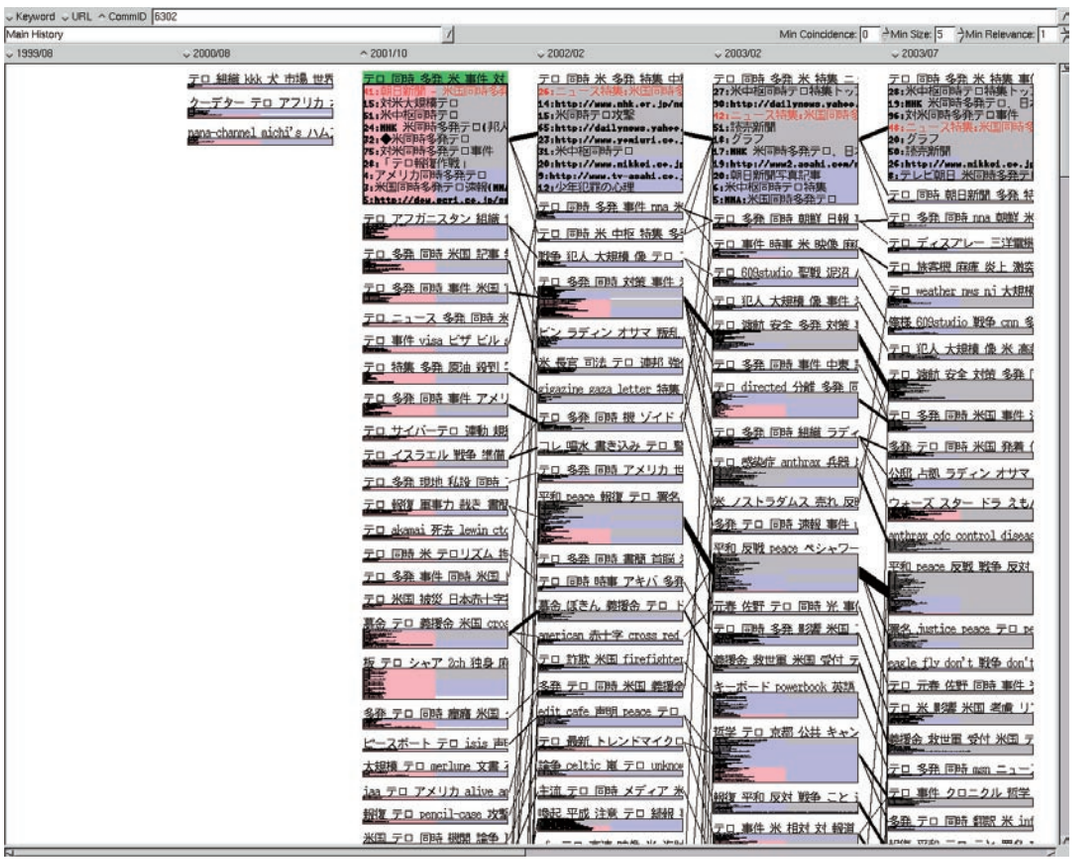

図-6　発展過程ビューア

各コミュニティの中では，ページの背景が左右に色分けされており，これがページの挙動を表現している．左側の色は前の時間からの変化を表現しており，灰色は前にも同じコミュニティに存在したページ，赤が新規に現れたページ，青が他のコミュニティから合併してきたページを表す．同様に右側では，灰色が次にも同じコミュニティに存在するページ，白が消滅したページ，青が他のコミュニティに分裂したページを表す．赤や青の多い部分では動きが激しく，灰色が多いところは安定しているなど，トピックの活発さを色から判断することができる．

発展過程ビューアは，お茶の水女子大学ジェンダー研究センターとの共同研究で，ジェンダーに関する社会学的な動向の調査に利用されており，専門のポータルサイト構築などへ応用する予定である[6]．

■今後の展望

大規模なWebアーカイブを用いてWebの発展過程をマイニングすることはデータ量の多さから困難であるため，現状ではあまり複雑な解析はなされていない．今後は，テキスト解析やリンク解析などさまざまな手法を組み合わせて精度を向上するとともに，大規模なアーカイブに耐えられるスケーラブルなアルゴリズムを開発することが望まれる．

また，近年blog等の新しいメディアに焦点を絞り，その中で話題が伝播する様子を把握する研究も盛んに行われており，今後の発展が期待される．blogでは，ページ単位より細かい更新時間情報が得られるため，より詳細な話題の構造が把握できる．ただしblogのみから得られる情報にも限界があるため，今後はニュースサイト，掲示版等，さまざまなメディアの境界における変化を扱う必要が出てくると思われる．

参考文献

1）Fetterly, D., Manasse, M., Najork, M. and Wiener, J.: A Large-Scale Study of the Evolution of Web Pages, Proceedings of the 12th International World Wide Web Conference (2003).
2）Ntoulas, A., Cho, J. and Olston, C.: What's New on the Web? The Evolution of the Web from a Search Engine Perspective, Proceedings of the 13th International World Wide Web Conference, pp.1-12 (2004).
3）Patterson, A.: CobWeb Search, http://ia00406.archive.org/cobwebsearch.ppt.
4）Toyoda, M. and Kitsuregawa, M.: Extracting Evolution of Web Communities from a Series of Web Archives, Proceedings of the 14th ACM Conference on Hypertext and Hypermedia (Hypertext 03), pp.28-37 (2003).
5）Toyoda, M. and Kitsuregawa, M.: Creating a Web Community Chart for Navigating Related Communities, Proceedings of the 12th ACM Conference on Hypertext and Hypermedia (Hypertext 2001), pp.103-112 (2001).
6）小山直子，増永良文：Companionーを用いたジェンダー関連Webコミュニティの詳細分析，情報処理学会研究報告 2004-DBS-134(II), pp.477-484 (2004).

（平成16年12月1日受付）