DEWS2004 5-B-02

リンク構造を利用した Web ページの更新判別手法

熊谷　英樹[†]　　山名　早人[‡]

†‡早稲田大学大学院理工学研究科　〒169-8555 東京都新宿区 3-4-1

E-mail:　†kumagai@yama.info.waseda.ac.jp，‡yamana@yama.info.waseda.ac.jp

あらまし　検索システムにおいては,データの新鮮さはその検索システムの能力を測る目安の 1 つであり，定期的に収集を行い，データの鮮度を保つ必要がある．そのために，大規模なシステムにおいては，データベースの維持にかかるコストが非常に大きなものとなっている. しかし, 我々の調査の結果では, 一ヶ月で約 10% の Web ページしか更新されていない。このため，短い間隔で再収集を行っていると考えられる商業の検索システムなどにおいては，全てのページに対して均等に再収集を行うことは非効率的である．本稿では，約 16,000 サーバ（150 万ページ）を対象として 6 ヶ月間の更新傾向を調査し，その解析結果に基づき，サーバ内のリンク構造を用いることにより，更新が行われた Web ページを効率的に再収集するための手法を提案する．

キーワード　Web クローラ，情報検索，更新判別

A Detection Technique for Modified Web pages using their Link Structure

Hideki KUMAGAI[†]　and　Hayato YAMANA[‡]

†‡Graduate School of Science and Engineering, Waseda University 3-4-1 Okubo Shinjuku-ku Tokyo, 169-8555 JAPAN

E-mail:　†kumagai@yama.info.waseda.ac.jp,　‡yamana@yama.info.waseda.ac.jp

Abstract In search engines, the freshness of data is one of the standards that measure the capability of the search engine. To maintain the freshness, periodical gathering is needed. For the reason, in the large-scale system, the cost concerning maintenance of a database is very large. However, in our experiment,　the percentage of updated pages are about ten during a month. Commercial search engines are considered to re-gather at short interval. However, equally re-gathering to all pages is inefficient. In this paper, we investigated Web page's updating tendency for six months for about 16,000 servers (1,500,000 pages). From this analysis, by using the link structure in Web servers, we propose the efficient techniques to gather updated Web pages.

Keyword　WebCrawler,　Information Retrieval,　Detection of Web pages' Change

1. はじめに

近年の WWW の急速な発達により，WWW を効率的に使うためには，ロボット型の検索システムが必須のものとなりつつある. ロボット型の検索システムなど，Web ページを対象とした検索システムにおいては，検索精度や網羅性以外に検索システム中のデータが最新のものであるかという点も重要な要素であり，一度取得した Web ページも定期的に再収集する必要がある．そのため，大規模な検索システムにおいては，データベースの維持にかかるコストは無視できないほど大きなものになっている．

また，Web ページ数の急激な増加により，必要な Web ページを収集するためのコストも大きくなってしまっている．そのため，近年では，必要な Web ページだけを効率的に収集し，収集時のコストを削減する為の手法が研究されている．

本稿では，実際の Web ページの更新を調査し，更新傾向を解析した．更新の調査は，約 16,000 サーバ，1,500,000 ページを 1 週間毎に，6 ヶ月間収集して行った．また，サーバ内のリンク構造を利用することで，更新が行われたページを効率的に再収集する為の手法を提案した．

本稿の構成は，2 節で関連研究を紹介し，3 節で調査の詳細，4 節で調査結果について述べる．また，5 節で解析結果に基づいて，Web ページを効率的に再収集する為の手法を提案する．6 節では，各手法を適用した

結果をまとめている.

2. 関連研究

WWW が発展するにつれ，大規模な Web ページのデータベースの必要性が高まり，Web ページを効率的に収集する為の様々な手法が研究されている.本節では，代表的な手法を紹介する.

2.1 重要なページのみを収集する手法

必要なページのみを対象とすることで，効率的に収集を行う手法が提案されている．代表的な手法では，PageRank[5](サーチエンジン Google[4]で用いられている)が高いページから優先して収集する手法[6]や，特定のトピックに属するページを優先的に収集する手法[7]が提案されている．しかし，これらの手法が適用できるのは，特徴的なページに特化した検索システムに限られる.

2.2 Web ページを効率的に再収集する手法

Web ページを対象としたシステムでは，定期的に再収集を行う必要がある．そのため，更新されているデータを判別することにより，効率的に再収集を行う手法が研究されている．代表的な手法を以下に示す.

2.2.1 プロトコルの拡張

過去に取得した Web ページから，変更された部分だけを取得できるように HTTP プロトコルを拡張する手法[9]や HTTP サーバの機能を拡張し，HTTP サーバ内の Web ページ内の更新されている Web ページをクライアントに通知するという手法[8]が提案されている．しかし，サーバ・クライアントに変更を強いる為，実用的ではないと考えられる.

2.2.2 各 Web ページの更新頻度の予測

Web ページは人手によって更新されている為，不規則に更新されているものが多いが，商業サイトなどでは，定期的に更新されている Web ページも多い．このことから，実際に対象となる Web ページの更新頻度を調査し，各 Web ページの更新を予測する手法[3]が提案されている．しかし，更新の傾向を予測する為には，長期間での調査が必要である．また，不定期に更新されている Web ページも多く存在すると考えられる.

2.2.3 Web サーバの更新傾向の利用

Cho 氏の研究[1]では，Web サーバごとに，Web サーバ内の Web ページの更新される割合が大きく異なることを利用している．各 Web サーバから数ページを取得して更新の有無を調べることにより，更新されている割合が高い Web サーバを判別して，優先的に収集を行う手法を提案している．この手法では，Web サーバ中の全ての Web ページを収集しているが，Web サーバの構造を利用することで，さらに効率的に収集を行えるのではないかと考えられる.

3. 更新頻度の調査

従来手法の問題点を踏まえ，長期間の調査を必要とせず，Web 全体を対象として，効率的に再収集を行う為に，本稿では Web サーバの更新傾向に着目した手法を提案する．Cho 氏[1]の研究では，各 Web サーバからランダムに Web ページを取得し，各 Web サーバの更新傾向を調査していたが，本稿では，Web サーバの構造から詳細な Web ページの更新傾向を調査し，効率的な Web ページの再収集方法を提案する．本節では本稿での更新傾向の調査方法を示し，４節において調査結果を述べる.

3.1 調査対象

大規模のデータを長期間調査するため 2003 年 6 月から 12 月の 6 ヶ月間，AC.JP，CO.JP，ED.JP，GO.JP，GR.JP, NE.JP の各ドメインからランダムに抽出した約 17,000 サーバを対象として調査を行った．各ページを再収集する間隔は一週間である．対象ページの詳細なデータを表 1 に示す．各ドメインからのサーバの抽出数は，各ドメインが全体に占める割合（JP ドメインのリスト[2]から算出）より定めた.

表 1　対象サーバの詳細

	サーバ数	ページ数	1 サーバあたりの平均ページ数
AC	90	36,526	405.8
CO	11,183	681,657	61.0
ED	463	86,193	186.2
GR	1,060	100,863	95.2
NE	2,377	347,744	146.3
OR	2,440	399,626	163.8
Total	17,613	1,652,609	93.8

3.2 調査方法

本稿で行った調査では，各サーバのトップページ（index.html）からリンクを辿ることで収集することができるページを対象とした．また，トップページから１１ 回リンクを辿った時点で収集を打ち切っている．収集されたページが更新されているかどうかは，MD5[10]によって求められた各ページのチェックサムを比較することによって判別した.

4. 調査結果

本節では，調査結果を示す.まず，全体の結果を示し，次に，サーバ単位での簡単な特徴について解析する.

4.1 全体の調査結果

まず，各ドメインの更新されたページの割合の推移を表 2 に示す．2 週間・1 ヶ月(4 週間)・3 ヶ月（12 週間）・6 ヶ月（24 週間）で区切り，調査を行った.

表 2 に示されているように，全体的に時間の経過と

共に線形に更新されたページの割合が上昇している．また，ドメイン毎に更新されたページの割合が異なり，CO.JP と OR.JP では，最終的に 10%近くの差が生じている．

表 2　更新されたページの割合

	2 週間	4 週間	12 週間	24 週間
AC	1.9%	6.1%	12.3%	18.1%
CO	4.5%	9.4%	16.8%	22.8%
ED	1.6%	8.8%	14.3%	17.7%
GR	2.9%	5.4%	11.3%	19.7%
NE	4.2%	7.6%	14.8%	23.0%
OR	2.1%	5.5%	11.5%	17.4%
Total	3.6%	7.7%	14.5%	21.0%

4.2 サーバ単位での特徴

次に，図 1 に各サーバのトップページが更新された割合を示す．図 1 に示されているように，全体の割合と比べて，高い割合でトップページが更新されている．この結果に注目し，トップページを更新しているサーバだけを収集した場合に，どの程度効率的に収集できるかを調査した結果を表 3 に示す．なお，表 3 では，トップページがフレームページであった場合には，各フレームページをトップページとして扱った．

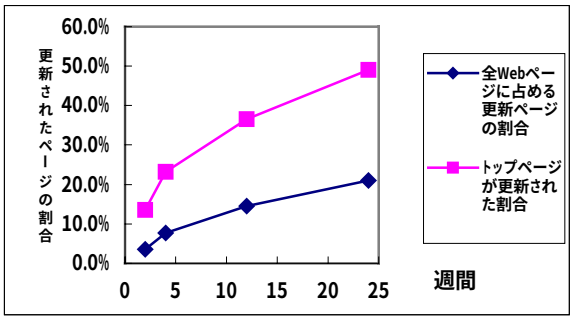

図 1　トップページの更新

表 3 に示すように，どの期間を見ても，トップページを更新しているサーバの割合は低いにも関わらず，更新されたページの大部分をカバーすることができている．この結果から，トップページをサーバの更新を判別する目安として用いることが可能であると考えられる．しかし，トップページを更新しているサーバは，更新していないサーバと比較して 1 サーバあたりのページ数が多く，収集したページ中の更新されたページの割合は，ページ全体での更新されたページの割合に対して，わずかに向上しただけである．そのため，本稿では，サーバ内のリンク構造を利用することで，より効率的に収集を行う手法を提案する．

表 3　トップページの更新による分類

	2 週間	1 ヶ月	3 ヶ月	6 ヶ月
トップページを更新したサーバ中の更新ページ数/全更新ページ数	63%	76%	84%	88%
トップページを更新したサーバ中の更新ページ数/トップページを更新したサーバ中の全ページ数	5%	9%	16%	22%
全更新ページ数/全ページ数	4%	8%	15%	21%
トップページを更新したサーバ数/全サーバ数	20%	35%	52%	66%

5. 提案手法

4 節の調査結果から，サーバのトップページは，サーバ内のページの更新を反映していると考えられる．これは，「更新のお知らせ」のようなものを，閲覧者が最も訪れやすい位置に設置しているからではないかと考えられる．しかし，大規模なサーバでは，一つのサーバに異なる種類のページが存在している．例えば，プロバイダや大学のサーバでは，明確に区切られた個人のページなどが存在する．これらのページでは，それぞれにトップページのようなページが存在し，更新の目安となっているのではないかと考えられる．そこで，本節では，サーバを小さな単位（サイト）に分割し，そのトップページを元に更新を判別する手法を提案する．

5.1 深度

まず，サーバ内のリンク構造を簡潔にするために，深度という尺度を定義する．深度とは，トップページからの近さを表す尺度で，トップページを深度 1 とし，トップページからリンクを一つ辿るたびに，深度が 1 つ上がると定義した．トップページからの経路が複数存在する場合には，最も小さい深度（トップページからの最短の経路）を適用する．例を図 2 に示す．四角がページを示し，中の数字が深度を示している．

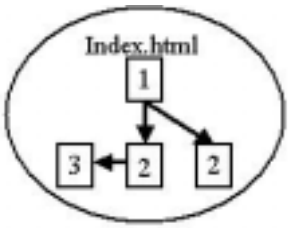

図 2　深度

5.2 ディレクトリによるサイト分割(手法 A)

サーバをサイトに分割する手法として，まず，最も単純にディレクトリ構造による分割手法を提案する．本手法では，ディレクトリ毎に 1 サイトとし，ディレクトリ内で 1 番深度が浅いページをトップページとした．サイトの記述方法は，（サイトのトップページ，メンバーページ 1，メンバーページ 2，・・・・）とする．

図 3 に示すリンク構造の場合には，ページ A をトップページとするサイト（A,B,C,D）とページ E をトップページとするサイト（E）に分割される．ディレク

トリ内で最も深度が浅いページが複数存在する場合には，該当する全てのページをそのサイトのトップページとした．

収集方針としては，各サイトのトップページをまず収集し，トップページが更新されている場合にだけ，サイト内のページを収集するとした．

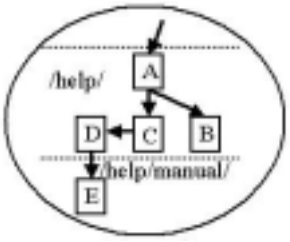

図 3　ディレクトリ分割

5.3 リンク構造によるサイト分割 (手法 B)

ディレクトリ手法では，トップページが複数ページとなり，チェックにかかるコストが大きくなってしまう為，本節では，リンク構造による分割手法を提案する．リンク分割手法では，ディレクトリによるサイト分割を行った後に，リンク構造によって更に細かくサイトを分割する．

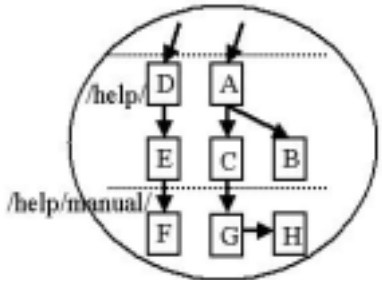

図 4　リンク構造によるサイト分割

まず，自分自身とは異なるディレクトリからリンクされているページをトップページの候補とする．図 4 に示すリンク構造の場合には，A,D,F,G が候補となる．次に，トップページからリンクが張られている同じディレクトリ内のページをサイトのメンバーとする．図 4 の場合では，（A,B,C），(D,E)，(F)，(G,H)となる．しかし，F のようにメンバーが 1 ページしかないサイトを生成してしまうと，収集が非効率的になるため，子のページを持たないページはトップページとしないこととした．その結果，F の親である E がトップページとなり，(F)の代わりに(E,F)が生成される．最終的に生成されるサイトは，（A,B,C），(D,E)，(E,F)，(G,H)となる．

5.4 リンク構造によるサイト分割の改良 (手法 C)

サイトの総数を少なくすることにより，チェックにかかるコストを削減することが可能である.そのため，本手法では，「上位のサイトが更新されていない場合は，下位のサイトのチェックを行わない」とした．上位のサイトとは，下位のサイトのトップページにリンクを張っているページが属するサイトであり，図 4 の場合では，(G,H)の上位サイトは(A,B,C)である．この場合には，A が更新されていない場合には，G のチェックは行われず, (G,H)のサイトは収集されない. また, (A,B,C)の上位サイトが更新されていない場合には，A も G もチェックされないとした．

6. 実験結果

本節では，提案手法の評価と従来手法との比較を行う．

6.1 評価指標

提案手法を評価する為の指標として，更新率とカバー率の 2 つの指標を定義する．更新率は，効率的に再収集が行えているかを示す指標であり，カバー率は，更新されたページをどの程度網羅的に再収集できているかを示している．以下にそれぞれの定義を示す．

$$\text{更新率}(\%) = \frac{\text{収集したページ中の更新されたページ数}}{\text{収集した全ページ数}}$$

$$\text{カバー率}(\%) = \frac{\text{収集したページ中の更新されたページ数}}{\text{全更新ページ数}}$$

6.2 節では，上記の指標を用いて各手法の比較を行う．目標は，更新された Web 全てを効率よく再収集することであり，更新率・カバー率ともに 100%に近づくほどその手法が優れていることになる．しかし，更新率とカバー率はトレードオフの関係にあるため，両指標でバランスよく高い値を出すことができる手法が求められる．

6.2 各手法の評価

本節では，各手法を 3 節で収集したデータに適用した結果を示す．図 5 では，Web ページ全体と各手法の更新率を，図 6 では，Web ページ全体と各手法のカバー率を比較している．

図 5 に示すようにどの期間を見ても，各手法が Web ページ全体より高い更新率を示しており，特に短い期間では，Web ページ全体より 2〜3 倍高い値を出している．特に提案手法 C は，平均的に Web ページ全体の 2 倍程度の値を出している．

しかし，図 6 に示すように，更新率が高い手法ほどカバー率は低くなっており，手法 C の場合のカバー率は 70〜80%である．

上記の結果から，収集できるページ数が少ない場合には提案手法 C を用い，カバー率を高く設定したい場

合には提案手法 A を用いるなどの使い分けをすることで効率的な再収集が行えると考えられる．また，提案手法 C を用いて収集したのちに，提案手法 C には合致しないが提案手法 B には合致するようなサイトを収集することで効率的に再収集が行えると考えられる．

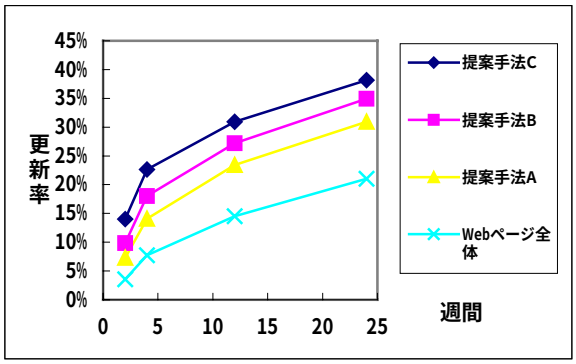

図 5　各手法の更新率の比較

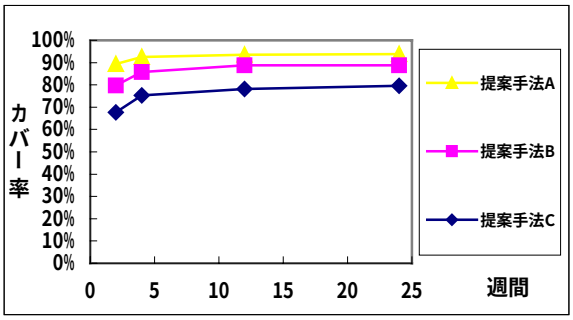

図 6　各手法のカバー率の比較

次に，図 7 にドメイン毎に分類した各手法の更新率を示し，図 8 にドメイン毎に分類したカバー率を示す．どの期間を見ても値に差がなかったため，4 週間の時点での値を用いた．

図 7 に示すように，どのドメインでも手法 C＞手法 B＞手法 A＞全体の順に更新率が高い．特に，ED ドメインでは手法 C は他の手法と比べて著しく更新率が高かった．逆に，GR ドメインでは手法 C と手法 B は大きな差がなかった．

また，図 8 に示すように，手法 C のカバー率はドメイン毎に大きく異なる値を示した．手法 A や手法 B では，ドメインに関わらず平均的に高い値を示しているが，手法 C では AC と OR ドメインでのカバー率が他の手法より著しく下がっていた．逆に，ED ドメインでは他の手法と大差ないカバー率を示している．上記の結果から，ED ドメインでは，更新率・カバー率ともに高い値を示しており，手法 C は ED ドメインを対象とした場合に，特に効果を発揮すると考えられる．CO や GR のドメインでは，企業のサーバが多く，表 1 に示すように比較的小規模なサーバが多い．小規模のサーバでは，ディレクトリに基づいてサイトを形成していない場合が多く，本稿で提案した手法が適用できない場合が多いのではないかと考えられる．

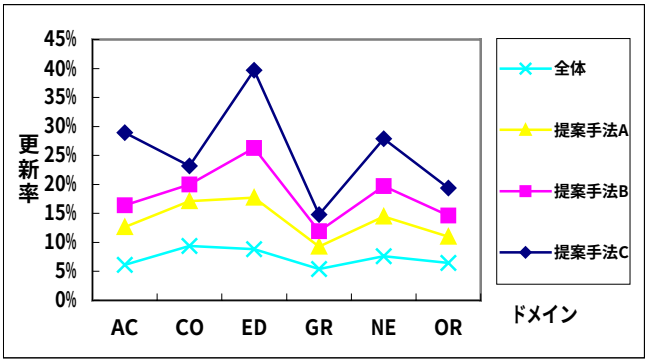

図 7　ドメイン毎に分類した各手法の更新率

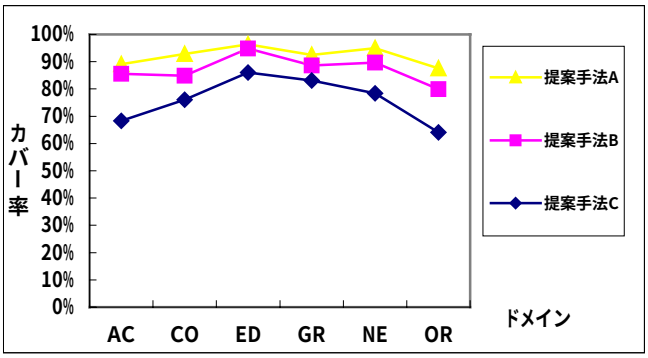

図 8　ドメイン毎に分類した各手法のカバー率

6.3 従来手法との比較

本節では，従来提案されている手法との比較を行う．

6.3.1 幅優先の収集方法との比較

まず本節では，最も広く用いられている幅優先の収集方法との比較を行う．幅優先の収集方法は，再収集のために考えられた手法ではないが，収集には最もよく用いられている手法であり，再収集の際にも用いられることが多い．幅優先の手法では，各サーバのトップページからリンクを辿り，5.1 節で定義した深度の浅いページから収集を行う．

図 9 に提案手法と幅優先手法との比較を示す．どの期間でも値に大差がなかったため，最も特徴の出ている 4 週間の時点での値を図 9 に示した．x 軸は更新したページの何%をカバーしているかを示し，y 軸はその時点で，全ページの何%を収集しているかを示す．深度の系列の点は，x 軸の値が低い順に深度 1 から深度 12 を表している．深度 1 の点は深度 1 のページのみを収集した場合，深度 2 の点は深度 2 までのページを収集した場合を示す．また，基準線はランダムにページを再収集した場合を示している．

目的としては，少ない収集ページ数で更新されたペ

ージを多く集められることが求められるので，サーバのトップページから順に収集をしていった場合でもランダムに再収集した場合よりは高い成果を得られていることが分かる．また，本論文で提案したどの手法もサーバのトップページから順に収集をしていく場合（幅優先の収集方法）より高い成果を挙げていることが分かる．

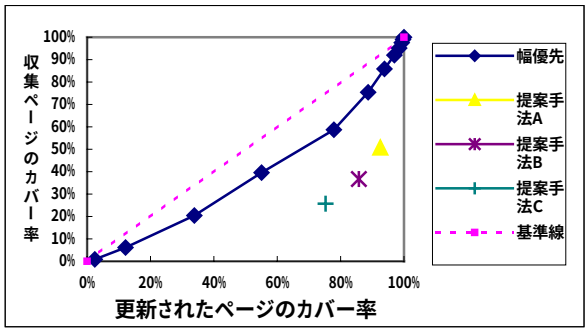

図 9　幅優先手法との比較

6.3.2 更新履歴を用いる手法の考察

まず，更新間隔に基づいた予測手法の有効性を考察する為に，図 10 に全 Web ページの更新回数の分布を示す．また，図 11 に更新された Web ページのみを対象として，更新回数の分布を示す．

図 10 に示すように，8 割以上のページは全く更新されていなかった．また，図 11 に示すように，更新されたページでは期間内に 1 度だけ更新されているページが 5 割以上を占めている．また，毎週必ず更新されているページ（4 週間時点での 3 回，12 週時点での 11 回，24 週時点での 23 回）の割合が高かった．例えば，24 週時点でのデータでは，9～22 回更新しているページのページ数は 2000～4000 ページ程度だが，23 回更新しているページは約 7000 ページ存在している．これはプログラムなどで自動的に更新を行っているページではないかと考えられる．

更新間隔に基づいた手法では更新が行われた Web ページを優先的に再収集するため，複数回更新されているページしか効率的に再収集することはできない．しかし，上記の結果から，2 回以上更新されたページは更新されたページ全体中の 5 割以下でしかなく，カバー率は 50%にも満たない．

また，対象期間が短くなればなるほど，1 度しか更新されていないページの割合は高くなっており，更新間隔に基づいた手法は長期間の調査を必要とすることがわかる（4 週間時点では 2 回以上更新されたページは全体の 36%だが，24 週間時点では 47%存在する）．対して，本稿で提案した手法では，2 週間～24 週間のどの期間でも幅優先で収集した場合の 2 倍以上の更新率で収集することが可能である．

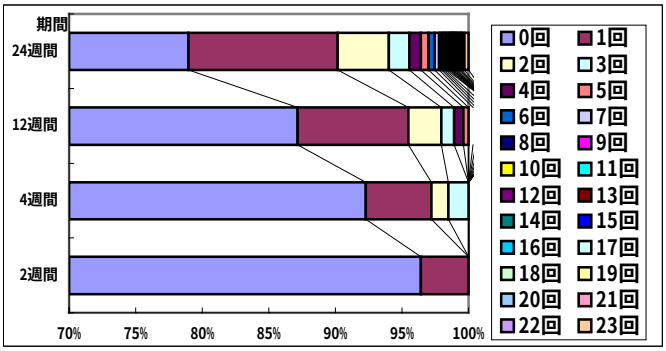

図 10　更新回数の分布（全 Web ページ）

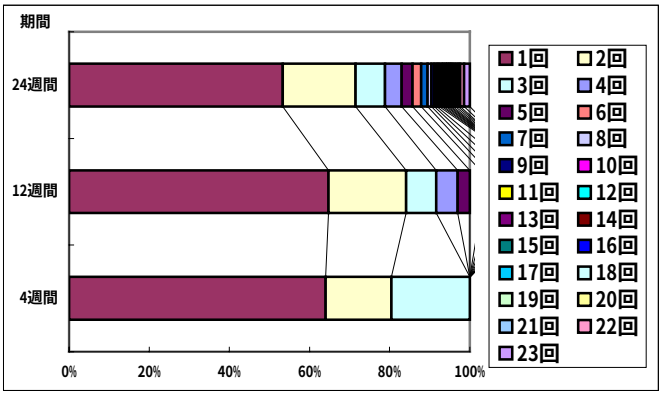

図 11　更新回数の分布(更新された Web ページのみ)

6.3.3 Cho 氏の手法[8]との比較

次に，Cho 氏が提案したサンプリングを用いた手法[8](2.2.3 節で説明)との比較を行う．図 12 に Cho 氏の手法に従い，更新率が高いサーバから再収集を行った場合の更新率を示し，図 13 にカバー率を示す．Cho 氏の手法では，ランダムにサンプリングを行った結果から各サーバの更新率を求めているが，今回の実験では，実際の各サーバの更新率を用いた．図 12,1 では，x 軸がサーバの更新率を表し，90%以上の場合には，サーバの更新率が 90%以上のサーバだけを再収集した場合，80%以上の場合には，更新率が 80%以上のサーバ全てを再収集した結果を示している．また，図中の基準戦は理論的な下限値を表す．

図 12 に示すように，Cho 氏の手法はサーバの更新率が高いサーバだけを収集している場合には更新率は非常に高く，図 13 に示すように，サーバの更新率が 10%以上のサーバだけを収集した場合は，60～70%をカバーすることができると同時に 40%以上の更新率で再収集を行うことができる．Cho 氏の手法はあくまでランダムのサンプリングをベースとしているので常に安定した値を期待することはできないが，効率的に再収集が行えると考えられる．

次に，図 14-17 に各期間での Cho 氏の手法と本論文で提案した手法との比較を示す．各図の X 軸は更新さ

れたページのカバー率を示し，Y 軸は全体中の何割のページを収集したかを示す．各図から，長い期間では提案手法と Cho 氏の手法では差がないが，短い期間では Cho 氏の手法の方が優れていることが分かった．

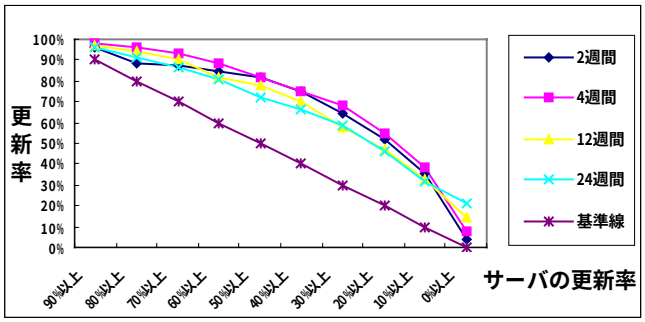

図 12　サーバの更新率に基づく収集方法の更新率

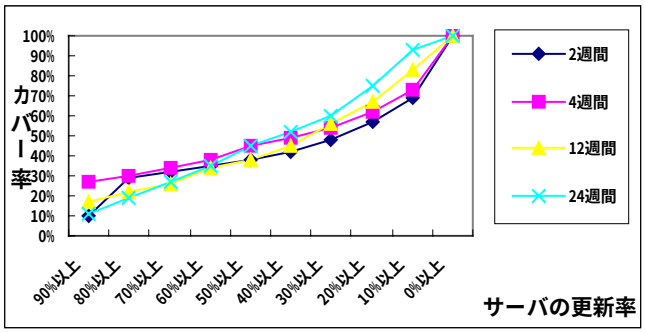

図 13　サーバの更新率に基づく手法のカバー率

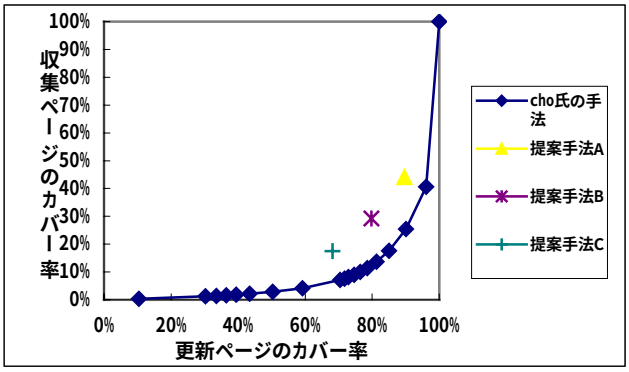

図 14　Cho 氏の手法と提案手法の比較（2 週間時点）

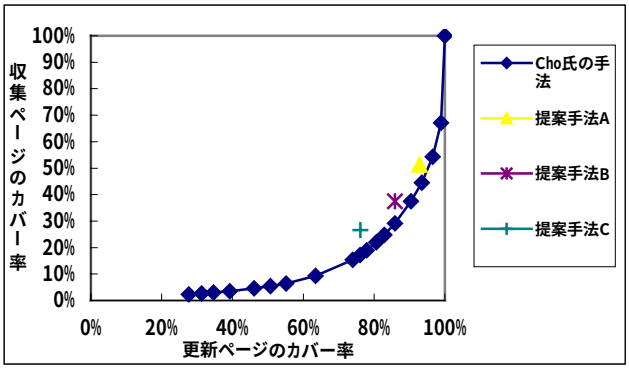

図 15　Cho 氏の手法と提案手法の比較（4 週間時点）

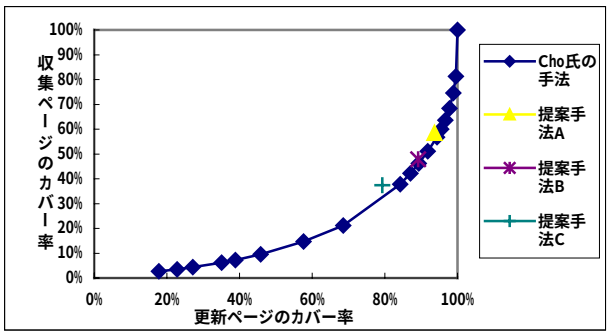

図 16　Cho 氏の手法と提案手法の比較（12 週間時点）

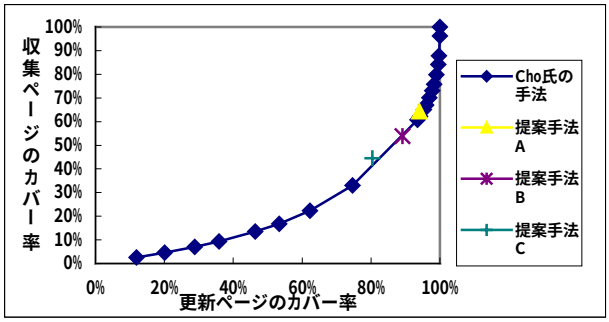

図 17　Cho 氏の手法と提案手法の比較（24 週間時点）

6.3.4 Cho 氏の手法＋本論文で提案した手法

本稿で提案した手法と Cho 氏の手法は相反するものではないので，次に，両者を組み合わせて適用した結果を図 18 に示す．

図 18 に示すように，2 週間の時点では，提案手法を組み合わせた場合よりも Cho 氏の手法を単独で用いた方が効率が良い．しかし， 4 週間の時点では提案手法 A を組み合わせた場合の方が効率的に再収集を行うことができる．また， 12 週間・24 週間の時点では，Cho 氏の手法を単独で用いるより提案手法 A や提案手法 B を組み合わせた場合の方が効率的に再収集を行うことができると考えられる．

提案手法 C を組み合わせた場合には効率が悪くなっているが，提案手法 C は更新率に重点を置いており，対象を絞って収集を行っているため，もともと更新率が高いサーバに適用した場合には効果を発揮できないのではないかと考えられる．

上記の結果から，Cho 氏の手法だけでも効率的に再収集を行うことができるが，対象の期間が 4 週間以上の場合には，本論文の提案手法（提案手法 A・提案手法 B）を組み合わせることによって，より効率的に再収集が行えることがわかった．

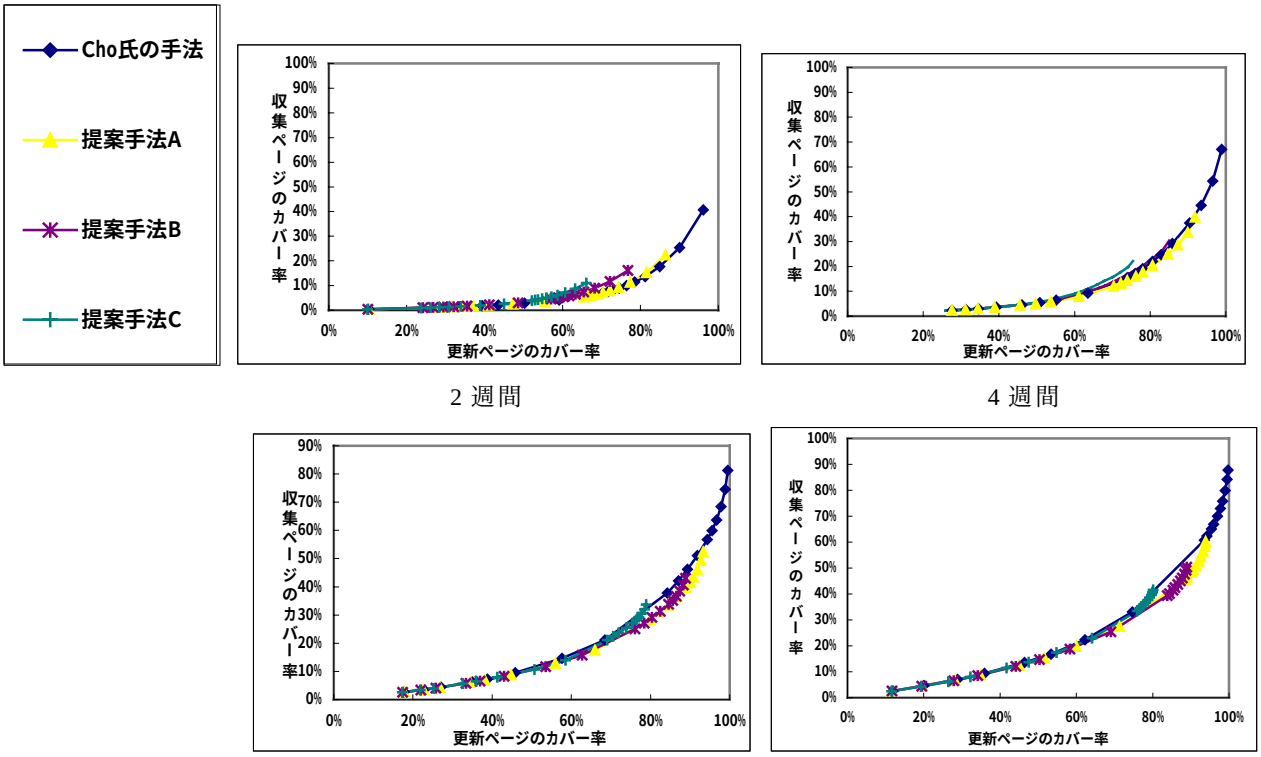

図 18　Cho 氏の手法と提案手法を組み合わせた結果

7. まとめ

本稿では，サーバ内のリンク構造を用いた Web ページの効率的な再収集手法を提案した．実験結果から，本手法を用いることで最も一般的な手法である幅優先の収集方法より 2〜3 倍高い更新率で再収集を行えることがわかった．また，ED ドメインでは他のドメインを対象とした場合より効果が大きいことが分かった．

6 章で行った従来手法との比較から，更新履歴を用いる手法は長期間の調査を要するが，高い効果は得られないと考えられる．また，Cho 氏の手法は，本稿での実験でも高い効果を得られているが，本稿で提案した手法と組み合わせることによってより効率的に再収集が行えることがわかった．

8. 今後の課題

本手法は大部分のサーバでは有効だったが，本手法での更新率･カバー率が著しく低いサーバも存在した．これらのサーバでは，ディレクトリをサイトの区切りとして用いていないことが特徴として挙げられる．このようなサーバに対してディレクトリを用いないサイト分割手法を用いることによって，より効率的に再収集が行えると考えられる．

謝辞

本研究の一部は，科学技術振興費「e-Society」，及び，文科省 21 世紀 COE「プロダクティブ ICT アカデミア」によるものである．

参考文献

[1] Junghoo Cho,Alexandros Ntoulas:"Effctive Change Detection Using Sampling",Proc of the 28th VLDB Conference 2002

[2] -:”!JP 日本レジストリサービス” http://jprs.jp/

[3] Jenny Edwards, Kevin McCurley, John Tomlin:"An Adaptive Model for Optimizing Performance of an Incremental Web Crawler",The 10th World Wide Web Conference:http://www10.org/

[4] -:”Google”http://www.google.com

[5] S.Brin,L.Page:"The anatomy of a large-scale hypertextual web search engine",Proceedings of the 7th International World Wide Web Conference,(1998)

[6] Junghoo Cho, Hector Garcia -Molina,Lawrence Page:"Efficient crawling throuth URL ordering",Proceedings of International World Wide Web Conference,p.p. 161-172(1998)

[7] Jeffrey Dean,Monika R.Henzinger:"Finding Related Pages in the World Wide Web",the 8th World Wide Web Conference:http://www8.org

[8] O.Brandman,Junghoo Cho,Hector Garcia-Molina,Narayanan Shivakumar:"Crawler-Friendly WebServers",In Workshop on Performance and Architecture of Web Servers (2000.6)

[9] Jeffrey C. Mogul, Fred Douglis, Anja Feldmann, Balachander Krishnamurthy:“Potential Benefits of Delta Encoding and Data Compression for HTTP”,SIGCOMM, pp. 181-194(1997.9).

[10] R. Rivest:“The MD5 Message-Digest Algorithm”, Network Working Group Request for Comments: 1321