

更新頻度に応じた地域WWWページの動的更新調査方法

Dynamic update investigation for regional WWW pages according to their update periods

山内慎祐† 白澤秀斗† 白川正知‡ 古川泰男‡
Shinsuke Yamauchi Hideto Shirasawa Masatomo Shirakawa Yasuo Furukawa

1. はじめに

一般にWWWページは下層に多数のリンクファイルを持っている。ページ全体としての更新の有無を知るには、下層のリンクファイル全てのヘッダ情報の更新日時を調査しなければならない。豊橋地域発のWWWページを生活情報のカテゴリ別に分類し、そのタイトル、概要、URL等をデータベース化したディレクトリ集である豊橋コミュニティ・ナビゲータは約500ページを収録している[1]。

これらは約46,000のファイルを下層に含んでいる。全ての下層ファイルのヘッダ情報を調べるのに約9時間を要し、現実的ではない。そこでファイルの更新頻度に応じて更新調査頻度を動的に変更する方法を考案し、30分以下で更新調査を可能とした。地域WWWページの更新状況を毎日調べ、ディレクトリ集の閲覧画面に更新マークを表示できるようにした。

2. 下層リンクファイルと更新調査

豊橋コミュニティ・ナビゲータに含まれる551ページが含むリンクタグから下層にある全てのHTMLファイル46,477個を抽出した。図1にファイル数に対するWWWページ数を示す。60%近くが下層ファイル数が20以下の小規模なページであるが、500以上を含む大規模なものも16ページあった。

本学LANを通じて全ての下層ファイルのヘッダ情報を読み出して、更新日時を調査したところ、平均9時間24分を必要とした。一方全てのファイルが毎日更新しているわけではないから、毎日すべてのファイルの更新調査をすることは効率が悪い。加えて多くのヘッダ情報取得に伴う対象サーバやネットワークへの負荷を考慮すると、効率の良い更新調査方法が必要である。

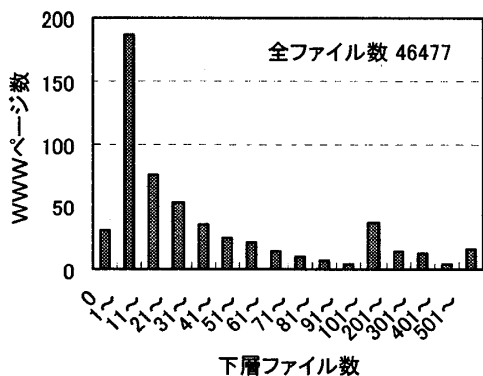


図1. 下層ファイル数に対するWWWページ数

3. 更新頻度に応じた更新調査方法

効率良く更新調査(ヘッダ情報の取得)を行うには対象ファイルの更新頻度に応じて更新調査頻度を変化させることが有効である。更新頻度の高いファイルは高い頻度で調査し、低いファイルには調査頻度を低くする。このようにして無駄な更新調査を排除できる。

そこで、ファイルの更新頻度を調査しながら、次第に更新頻度に応じて更新調査頻度が自動的に変化してゆく図2のような調査頻度可変アルゴリズムを考案した。この方法では予めファイルごとに調査頻度閾値を定めておく。そのうえで、まず更新調査日と前回調査日の差分をとり、閾値より小さいならば、更新調査は行わない。閾値以上ならば更新調査を行う。

得られた更新日時と更新調査日の差分から、調査頻度閾値が適切かを判断し、閾値の修正を行う。初めは全てのファイルを毎日調査しても、更新がされていないページは、閾値が次第に大きくなり、更新頻度が自動的に低下していく。図3にそのようなすを示す。46,467個のファイルに対して、概ね5回の調査で調査所要時間は急速に低減し、数10分になる。

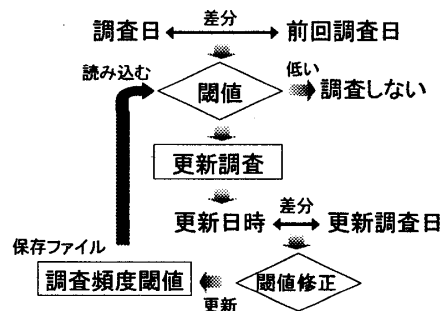


図2. 更新調査頻度可変アルゴリズム

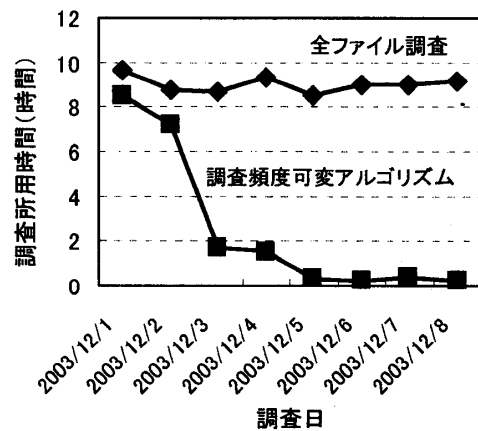


図3. 調査頻度可変アルゴリズムの効果

† 豊橋技術科学大学 大学院 工学研究科, Graduate School of Engineering, Toyohashi University of Technology

‡ 豊橋技術科学大学 未来技術流動研究センター, Research Center for Future Technology, Toyohashi University of Technology

4. 更新調査

このような更新調査方法により、豊橋コミュニティ・ナビゲータに収録されているWWWページの更新調査を毎日定時に実施した。調査頻度を定める閾値として、表1に示すような7つの段階を設定した。これらは日、週、月、半年、年に概ね一度更新するファイルの捕捉を目的にしたものである。

図4にその結果を示す。トップページのヘッダ情報からのみ更新を調査した場合と下層ファイルのヘッダ情報を含めて調査した場合とは、後者の方が多くの更新を捕捉していることが分かる。ここでは下層ファイルが一つでも更新されていれば、そのページは更新しているものと見なしている。すなわちトップページのヘッダ情報の更新日時からだけでは、ページ全体の更新状況が正確に捕捉できないことを示しており、提案方法の有効性が分かる。

表1. 更新調査頻度の閾値

調査頻度フラグ	調査頻度閾値	更新状況捕捉対象
1	2日未満	毎日更新のファイル
2	10日未満	週に一度更新
3	40日未満	月に一度更新
4	200日未満	半年に一度更新
5	400日未満	年に一度更新
6	765日未満	二年に一度更新
7	765日以上	二年以上更新なし

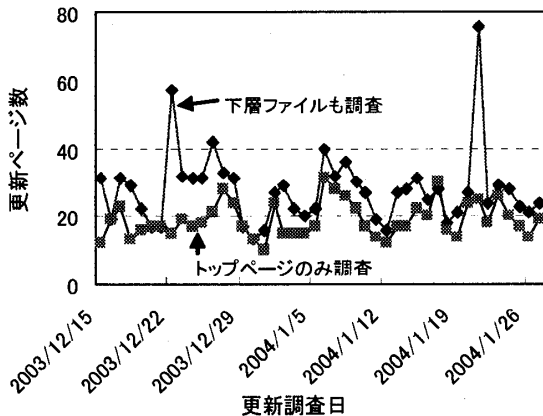


図4. 更新調査の結果

図5には調査したファイル数と更新ページ数の関係を示す。調査したファイル数は調査頻度可変アルゴリズムの特徴により日毎に変動する。図中のA, B, Cは表1で示した調査頻度フラグ2の閾値10日未満のファイルの調査周期と推定される。またDは40日未満に対応するファイルの調査周期と推定される。44日という短い期間のデータであるが、調査したファイル数に概ね応じて更新ページが捕捉されていることが分かる。

図6には調査所要時間を示す。調査したファイル数に応じて調査時間が長くなっている領域(大括弧の部分)と必ずしもそうとはいえない大きく食い違う領域とがある。その原因についてはネットワークの状況等も考慮しなくてはならず、解明できていない。

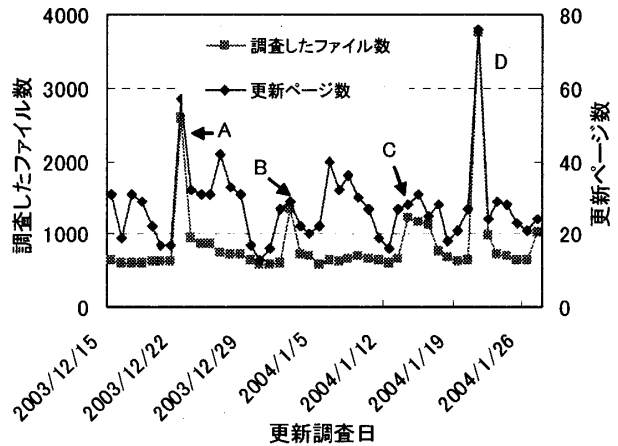


図5. 調査したファイル数と更新ページ数

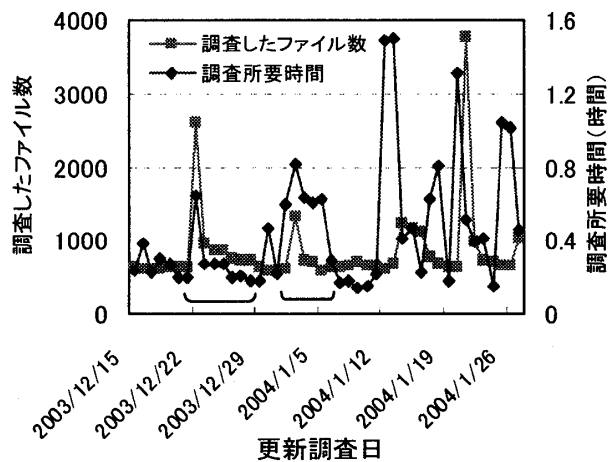


図6. 更新調査所要時間

5. 考察

更新調査頻度可変アルゴリズムにより下層ファイルも含めた大量のページの更新状況を効率良く捕捉可能になった。この方法は更新の見逃し(更新を検出できずに次ぎの更新がなされる)、更新検出の遅れ(更新の検出が更新日より遅れる)等の問題がある。ファイルの更新発生を確率的にモデル化して、提案方法がどの程度正確に更新を捕捉しているかを評価する必要がある。

6. おわりに

本研究は地域発のWWWページがどの程度新鮮な情報を提供しているかを把握し、更新状況を地域情報利用者に知らせる目的をもっている[2]。トップページからのみでは日々更新されているページは全体の5%前後、下層ファイルの更新を考慮しても10%未満と、更新状況は高いものではない。また情報利用の点から下層ファイルの更新がトップページに反映される仕組が望ましい。

参考文献

[1]古川: 情報処理学会第60回全国大会, 4-243(2000).
 [2]山内ほか: FIT2002, L-15(2002).