ネットワーク適合型WWW情報配送システム

WWW Information Delivery System Adapted To Network

大前 雄一朗　　北村 泰彦　　辰巳 昭治

Yuuichirou Ohmae, Yasuhiko Kitamura and Shouji Tatsumi

大阪市立大学 情報工学科

Department of Information and Communication Engineering
Faculty of Engineering Osaka City University

Abstract

As the amount of available information on the WWW increases vastly, we feel difficulty to collect information efficiently. In this paper, we propose an automated information delivery system. It is located between users and WWW resources as a proxy, and automatically collects WWW pages at designated URLs and notifies the updates to the users by using E-mail. While it collects pages, it learns the communication speed and the frequency of updates for each server and chooses a proper access strategy flexibly.

1　はじめに

現在，WWWは世界中で用いられ，多くの情報が提供されている．また，情報発信も容易になり，日々新しい情報が次々に加えられている．しかし，次第にそのデータ量は膨大なものとなり，ユーザが情報収集に費やす時間や労力が増加している．ユーザ側の負担としては代表的なものが三つ考えられる．

まず最初に，ユーザが常に最新の情報を手に入れるためには，常にサイトをチェックし情報を収集しなければならない．さらに，チェックするサイトの増加にともない，またそのサイトの更新頻度の増加にしたがって，ユーザ側の負担がさらに増加し，その結果，ユーザが情報収集を行いたいと思う全てのサイトをチェックすることがかなり困難になる．

次に，一般的にネットワークの負荷は一定ではなく時間帯によって異なる．さらに接続先の相手サーバによっても大きく異なる．具体的に述べると，接続先のサーバが企業の場合には昼頃にアクセスが集中し，プロバイダの場合には夜にアクセスが集中するなどの傾向がある．アクセスが集中しネットワークの負荷が高い時間帯にアクセスすると，情報収集に余計な時間がかかることとなる．

〒558-8585　大阪市住吉区杉本3-3-138
{ohmae,kitamura,tatsumi}@kdel.info.eng.osaka-cu.ac.jp
http://www.kdel.info.eng.osaka-cu.ac.jp/

最後に，Webページの更新頻度は様々であり，それぞれのサイトによって異なる．ニュースなどを扱うサイトは頻繁に更新を行うが，大学のトップページなどはほとんど更新されない．もしアクセスしたWebページが更新されていなかった場合には，アクセスするためにかかった労力や時間が無駄なものになり，さらにネットワークに余計な負荷をかけることになる．逆に，更新されているWebページにアクセスしないと，重要な情報を取りこぼす可能性がある．例えば，更新の間隔がWebページに記述されていればよいが，そういった更新に関する情報はほとんど記述されておらず，結局，情報を取りこぼさないために何度も更新されていないWebページにアクセスすることになる．

本稿ではこの三つのユーザ側の負担を軽減し，なおかつネットワークを効率的に利用するWWW情報配送システムについて述べる．ユーザは巡回を行いたいWebページのURLを登録するだけでよい．すると，本システムがネットワークの負荷の状態を監視し適切な時間帯に，巡回先のWebページの更新頻度に応じた適切な頻度で情報収集を行い，その結果をユーザに自動配送する．

２章で本システムと動的学習システム，適切な巡回頻度の設定方法について詳しく述べ，４章で動的学習システムの評価を行う．最後に５章において得られた諸結果の総括を行って，今度の課題について述べる．

2 WWW情報配送システム

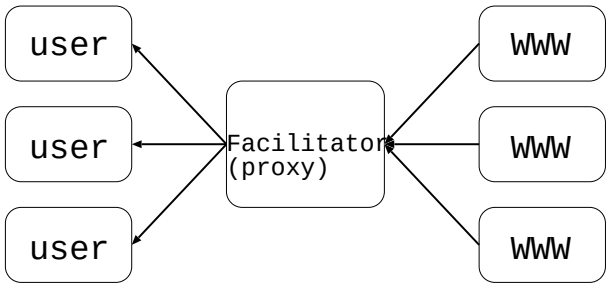

図 1: システムの概念

まず本システムの全体像について述べる．

このシステムはユーザーとWWWの間にFacilitatorとして存在し，ユーザはブラウザを介して要求を述べることができる

具体的には，最初にプロキシを設定し，スタート画面にアクセスする．スタート画面でIDとパスワードを入力すれば，ブラウジング画面になる．

図 2: ブラウジング画面

ブラウジング画面はフレームにより上下に分割されていて，上段がコマンドフレーム，下段がブラウジングフレームとなっている．上段のコマンドフレーム内のGOTOの欄にURLを入力することで，ブラウジングを行うことができる．現在表示されているWebページの巡回を行いたい場合には，SAVE THIS PAGEのボタンを押す．そうすると，そのWebページのURLが巡回リストとして保存される．また，POSTを使ったCGIの場合，クエリーに入力した文字列も記憶している．従って，データベースでの特定のキーワードに対する検索結果の更新状況なども調べることができる．ユーザが巡回先を登録すると，あとは本システムが適切な時間帯と頻度で巡回を行う．巡回を行う時間帯と頻度の設定方法は後で述べる．

巡回を行って，更新のあったWebページはサーバ内にキャッシュする．そしてそのキャッシュしたWebページのURLをユーザに電子メールで通達する．ユーザはブラウザでアクセスすることにより，情報を得ることができる．

2.1 動的学習システム

一般的にアクセスが集中する時間帯に一定の傾向があるのは，前述の通りである．またネットワークの負荷は変化するものである．負荷の低い時間帯にアクセスし，ネットワークの負荷を動的に監視すれば，ネットワークを効率的に利用することができる．

従って，本システムは動的学習を行う．まず，その際に残しておくデータの内容について述べる．このデータには，情報収集を行ったサーバ名とその際の通信速度を時間帯別にして記録されている．分類をWebページ毎ではなくサーバ毎に行ったのは，通信速度の記録はWebページごとに行うことが考えられるが，あるWebページへの通信速度は，そのWebページのサーバによって決まると考えられるからである．

次に学習方法について述べる．本システムは一定期間ごとに学習を繰り返す．

data

server name	0:00	1:00	.....	22:00	23:00
www.asahi.co.jp	4050	3454	.....	2194	1593
www.yomiuri.co.jp	6594	7505	.....	5785	5234
	.				
	.				
	.				

図 3: データの例

学習したデータは図3のように，それぞれのサーバにおける一日の間の時間帯別通信速度 [bps] を一つのまとまりとする．その学習したデータ (記憶) は記憶領域 (memory) に保持される．記憶領域にはある一定の記憶量が設定されている．もし学習した記憶がその記憶量を超えた場合には，古い記憶から忘却していく．そして，その記憶領域内の記憶の平均値を計算し，サーバ毎に巡回時刻を決定する．例えば，記憶領域にD1，D2，D3が保持されていたとする．今，新しくD4を学習してきた場合には，一番古いD1が削除 (忘却) され，その代わりにD4が挿入される．このことを図4に示す．

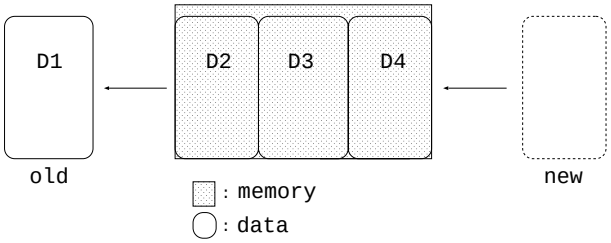

図 4: 更新方法

これらのことにより，ネットワークの負荷の低い時間帯にアクセスすることができる．従って，特定の時間帯にアクセスが集中するのを避けて，ネットワークの負荷の均衡化を行うことができる．

2.2 適切な巡回頻度

それぞれのWebページによって，更新頻度は様々である．ネットワークを効率的に利用するためには，更新頻度の高いページに対しては巡回の回数を増やし，逆に更新頻度の低いページに対しては巡回の回数を減らす必要がある．そのため，本システムでは，Webページ毎に一日に巡回する回数を設定し，そのWebページの更新頻度に応じて巡回の回数を変化させている．

具体的には，一日に一度巡回する場合は最も通信速度の速い時間帯を選ぶ．一日に数回の巡回を行う場合には，Webページの更新が一日に均等に行われると仮定して，２４時間で均等に巡回を行う．例えば，一日に３回の巡回を行う場合には，0,8,16時，1,9,17時，2,10,18時…のように巡回する時間帯の組み合わせを決めた上で，これらの組み合わせのうちで最も効率的に巡回できる時間帯，つまり３回分の合計通信速度が最も速くなる時間帯を選択する．

また，巡回頻度は一日おきに変更する．翌日の巡回頻度は，巡回した回数の内，実際にWebページが更新された回数で３通りに場合分けを行って設定する．

具体的に一日N回の巡回を行う場合を考えてみる．

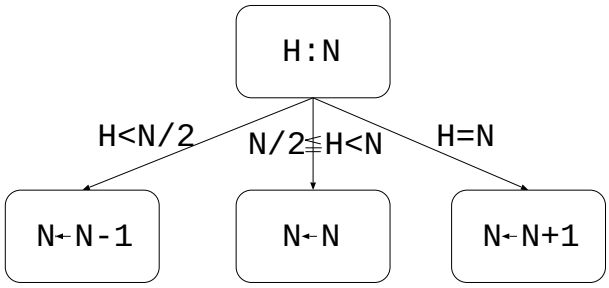

図 5: 更新ヒット回数による場合分け

図5の中のH [回] とは，更新ヒット回数 (一日の間に，巡回を行った際に巡回先のWebページが更新されていた回数) である．この図のように，Webページの更新がｎ／２回未満であった場合には，巡回の回数をｎ－１回とする．Webページがｎ回とも更新されていた場合には，翌日の巡回の回数をｎ＋１回とする．それ以外の場合には，巡回の回数はｎ回のままで変更を行わない．

3 動的学習システムの評価

ここでは本システムの動的学習システムの有用性を評価するための実験をおこなう．

動的学習システムにより，ネットワークの負荷の低い時間帯にアクセスできるかどうかを実験するため，以下の二つのモデルの通信速度を比較した．

(1) 動的学習システムを利用し巡回時刻を決めるモデル

(2) 定時 (0時) に巡回を行うモデル

一日の巡回頻度を一回と設定し，(1) のモデルは先に一週間の学習を行い，その後の一週間を実験期間とした．巡回先のWebページは，ニュースを扱うサイト，企業のサイト，大学のサイトから，３つずつ，合計９つ選んだ．

表 1: 実験結果

	通信速度 (bytes/sec.)
(1)	139380
(2)	125679

上の表１にあるとおり，上段の動的学習システムを利用したモデルのほうが，下段の定時に巡回を行うモデルより通信速度が高い．これは動的学習システムによる効果である．

4 結論

本稿ではユーザの代わりに情報を収集し，その結果を配送するWWW情報配送システムについて述べた．動的学習システムを提案し，その動的学習システムにより学習した記憶などに基づいて，適切にアクセスできる時間帯と巡回頻度を決定した．それにより，ネットワークの負荷を抑え，負荷の集中する時間帯を分散させることが可能となる．

今後の課題としては，更新箇所の明示が挙げられる．これは更新箇所が少なかったりサイズが大きいWebページなどでは，どこが更新されたかすぐに判断できない場合があるからである．

参考文献

[1] 浅井宣和, 通信状態と更新頻度を考慮したWWW情報自動配送システムに関する研究, 大阪市立大学工学部情報工学科卒業論文, 1998