1F－6

WWWにおけるユーザ主体の情報差分提供システム*

松下 康之†　Santi Saeyor†　石塚 満†

東京大学工学部電子情報工学科‡

1　はじめに

近年のWWWの普及により、誰でも個人で情報発信を行うことが出来るようになった。その情報量は増え続けている。また、WWW情報空間が非均質でダイナミックな性格をもつため、WWW全体を把握することは不可能であるといえる。

こうした中で、ユーザが自分が求める情報に効率よくアクセスする事は非常に難しい。

そこで、本研究ではウェブページの更新状況を定期的にチェックし差分情報データベースを作り、ユーザの求める情報をもとにフィルタリングを施したものをユーザに提示することで、ユーザの情報入手の効率化を図るSite Managerの提案をする。

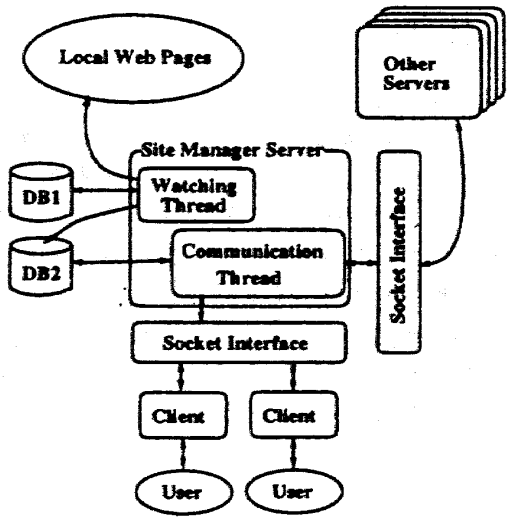

図 1: 本研究で想定するシステム概要

2　本システムの目的と利点

本システムの目的は、ユーザの求める情報へのアクセスの効率化である。情報の差分を提供することで、ユーザは何がどのように変化したのかをすぐに知ることができ、新しい情報を効率的に得られるようになる。また情報の差分提供をHTTPd側で行うことによって、個人ユーザ毎にロボットを動かし重複した更新確認要求をHTTPdへ送ることも不必要となる。以下、そのシステムの概要を述べる。

3　本システムの概要

本稿のシステムは、HTTPdと並列に常駐されるサーバ部分とユーザ個人がもつクライアント部分に分けられる。こうすることのメリットは、サーバが常に情報の更新を監視する際にネットワークを介す必要が無くなることである。以下、その構造とウェブページの情報の差分の扱いについて述べる。

3.1　クライアント側の機構

クライアントはユーザが更新をチェックしたいウェブページのURL及び、そのページに関してどの程度の変化を更新と見なすかという情報を付加してサーバへ更新情報の提示要求をする。ここでいう変化の程度とは、ウェブページの変化の量、種類によっていくつかの段階的な分類を行いそれに基づいてユーザの求める更新を知るためのものである。それについては以下の差分情報の扱いで述べる。また、ユーザが更新をチェックしたいウェブページの1リンク先のウェブページにもユーザの興味のある情報が多く含まれる可能性が高くそれをチェックするか否かを指定する。

---

* User-Oriented Information Differencial Providing System on WWW

† Yasuyuki Matsushita , Santi Saeyor , Mitsuru Ishizuka

‡ Dept. of Information and Communication Engineering, Faculty of Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113, Japan
e-mail : yasuyuki@miv.t.u-tokyo.ac.jp

3.2 サーバ側の機構

サーバはHTTPdが動作しているのと同じマシン上で動作させ、定時間隔でウェブページの更新状況をチェックし変化があればその差分をもとに生成する差分情報をデータベースへ追加する。この時、更新前のウェブページの情報も保持しておりそれとの比較で差分を抽出する。比較はまず最長共通文字列を検索しその前後で2つの文書に分け、それにより分けられる2つの部分に関して同じ操作を繰り返すことで差分を抽出した。
クライアントからの要求があればその要求にあてはまる更新情報をデータベースから取り出し、クライアントへ返す。ここで、クライアントからの要求が他サイトのウェブページの更新状況を含む場合には、サーバ間通信により情報を所定のサイトのサーバから取り寄せてクライアントへ返す。ここでクライアントからの要求はクライアント固有のものとなっておりそれをもとにフィルタリングをここで施してそれを応答として返す。

3.3 差分情報の扱い

単に最終更新時刻をチェックするのみでは、ユーザにとって意味のある情報の更新が行われたとは判断できない。そこで、ユーザにどの程度の変化を更新と見なすかという情報を要求する。いまの段階では、変化の量については更新されたテキストの量（タグを除く）に関しての分類を行い、変化の種類に関してはリンク先の増減を対象としている。これらの分類に基づき、どのような更新を意味のある更新と見なすかということをクライアント側でユーザが指定する。

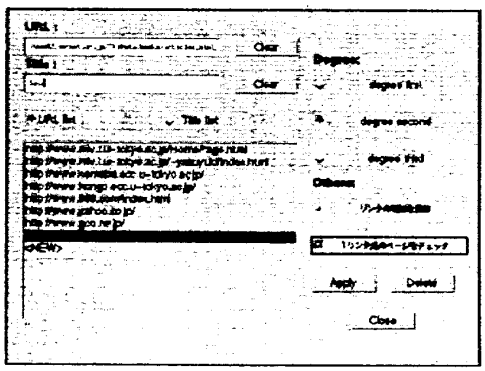

図 2: クライアント側のリクエスト作成画面

3.4 処理の流れ

上記のように、クライアント側では更新をチェックしたいURLとそれに付随する情報を作成し、それをサーバへ送る。サーバは受け取ったURLの更新状況をデータベースより調べてそれがユーザの求めるような更新であった場合、それを差分情報としてクライアントへ送る。クライアントから送られてくる情報の中に、最後にチェックした時刻が含まれており、それ以降の差分情報の書き換えが無かったならば更新は無いということになる。サーバによる更新の監視は、クライアントの要求とは独立して定時間隔で行われる。

3.5 差分情報の提示

ウェブページの差分情報は、そのウェブページをベースに更新されて新しくなった部分に色付けをし、削除された部分は本文とは少し離した形で一番下にまとめて提示することにした。これにより、ユーザはどこがどのように更新されたのかを一目で知ることができる。また、削除された部分に関してはそれほど重要度が高くないと思われるため本文とは無関係のところにまとめて記述するという方法をとった。

4 おわりに

本稿ではWWW上の情報資源に対する効率的なアクセスのための情報差分提供システムを提案し、そのシステムを作成した。ただ、サーバ間通信に関しては十分な実験が出来ず、現段階では評価が難しい。部分的にネットワークを介して情報収集するロボットと連動させて実験を続けたい。今後は差分抽出の効率的な手法について検討し、ユーザの個人情報をより多く利用して、より個人にマッチした情報差分提供システムを作成したい。

参考文献

[1] UMBC AgentWeb – Agent Examples
http://www.cs.umbc.edu/agents/agents.shtml

[2] UMBC AgentWeb – Agent related projects
http://www.cs.umbc.edu/agents/projects/