4X-1

Web 情報のネットワーク分散型アーカイビング手法

大門 和斗† 矢吹 太朗 佐久田 博司
青山学院大学 理工学部 情報テクノロジー学科‡

1 はじめに

1.1 背景

デジタルメディアの普及とともに，インターネット上の情報を，論文やレポートの参考文献として掲載するケースが増えてきた．しかし，本や雑誌などの出版物と違い，インターネット上の情報は，永続的なアクセス可能性に問題がある．時間をおいて再度アクセスしようとしても，既にページが消されていたり，ページが書き換えられてしまっているため，同じ情報が手に入らない場合がある．

特に競争の激しい研究分野においては Web で研究成果を発表したときに誰が最初にその情報を掲載したかが争点となることがある [1]．しかし，ある情報を誰が最初に発信したのかを特定するためには，その情報の出現日時を明確にする必要がある．

これらの問題は，情報を時系列順にバックアップしていき，後で参照できるシステムがあれば解決できる．実際にインターネット黎明期より Web 情報のバックアップという試みは行われている．本研究では Web ページのデータのコピーをキャッシュと定義した．

1996 年に設立された Internet Archive という団体は，インターネット上の情報をコピーしておくことで後で参照できるようにした Wayback Machine と呼ばれるシステム（http://www.archive.org/web/web.php）を運営している．このシステムに Web ページの URL を指定すると，そのページが過去どのようなものであったかを時系列順に表示する仕組みになっている．

ウェブ魚拓（http://megalodon.jp/）は URL を指定するとそのページのコピーが当該 Web サービスのサーバに保存され，あとで参照できるようにするサービスである．しかし，この Web サービスでは，ユーザが指定したページしかコピーされないため，後になって必要であることが分かったページに関しては復元できない．

日本の国会図書館でも Web アーカイブ事業を行っている [2]．しかし，公的機関の Web サイトに関しては自動でデータを収集するが，一般の Web サイトに関しては許可を得た場合のみデータを収集するため，一般のユーザは限定的にしか Web アーカイブのメリットを享受できない．

2 目的

本研究では，ネットワーク分散型の Web アーカイブ手法について提案し，そのプロトタイプシステムを開発する．

前項で述べた Internet Archive では，Web クローラ (自動的に URL にアクセスしてデータを取得してくるロボット) を用いてキャッシュを収集し，収集されたデータは Internet Archive のシステム内に保存される．

これに対して，本研究で提案するシステムは，人が実際に閲覧した Web ページをもとにキャッシュを収集し，収集されたデータはその時利用された端末内に保存される．つまり個々の端末が Web アーカイブのための分散データベースのためのノードとなる．

3 手法

本研究で提案するシステムの概念は図 1 のように表される．

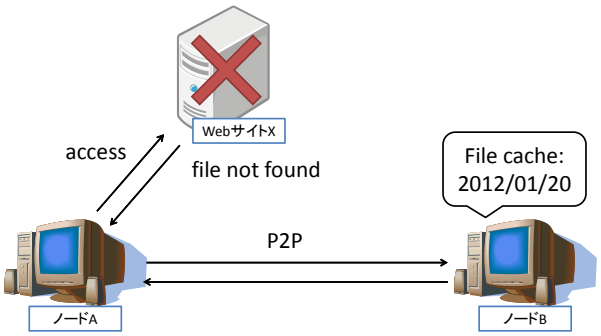

図 1 システム概念図

図 1 では，ノード A が Web サイト X にアクセスしようとしているが，ファイルが見つからなかった状況を表している．ノード B は，以前 Web サイト X にアクセスしたことがあり，その時のキャッシュを持っている．そこでノード A がノード B からキャッシュを受け取れば，サイト X を閲覧できるようになる．

このように各ノードが取得したキャッシュを共有する

Web Archiving on Distributed Network
† Kazuto DAIMON (kazuto.daimon@gmail.com)
‡ Department of Integrated Information and Technology, College of Science and Engineering, Aoyama Gakuin University

ネットワークが成立する．しかし，ユーザが閲覧したページをそのままキャッシュとして保存してしまうと，そのキャッシュにプライバシー情報が含まれてしまう危険がある．そこで，個々のノードのプライバシー情報をキャッシュしないようにするため，プライバシー情報を削除した通信で URL に再アクセスする．つまり，無料でかつ認証なしで取得できた情報のみがキャッシュされる仕組みになっている．

3.1 ページのアーカイビングと復元

3.1.1 ページのアーカイビング

ユーザが Web ページを閲覧するとシステムがその URL を取得する．前述のように，システムはプライバシー情報を削除し，再び同じ URL にアクセスする．その際に得られたデータをローカルファイルシステムに保存し，ローカルデータベースにキャッシュのメタ情報 (URL・取得日時・コンテンツタイプ・ファイルシステム上での格納ディレクトリ) を記録する．その後，中央サーバ（後述）にキャッシュしたことを通知する．

3.1.2 ページの復元

ページを復元したいときは，そのページの URL を中央サーバに問い合わせる．そのページのキャッシュ情報があれば，それを返す．キャッシュ情報を受け取ったノード（図 1 のノード A）は，そこに記録されたノード（図 1 のノード B）にキャッシュを要求する．ローカルデータベースからキャッシュを復元し，要求元のノードに送信する．

4 システム構成

本研究では図 2 のような機能構成でクライアントソフトを実装した．

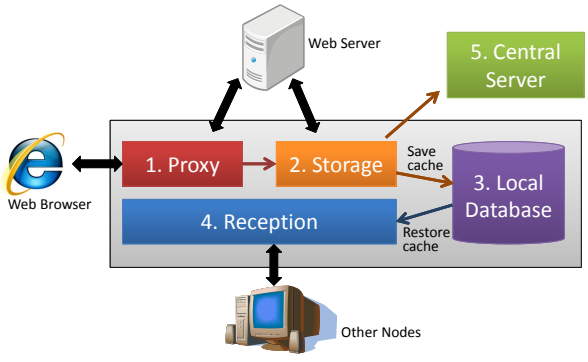

図 2　システム構成図

4.1 データの保存機能

図 2 中 1 の Proxy はユーザ端末上のプロキシサーバとして機能し，ブラウザからの HTTP リクエストを受け取る．Proxy はブラウザからの HTTP リクエストを 2 の Storage に渡す．Storage はそのときにリクエストに含まれていた Authorization ヘッダや Cookie ヘッダ等のプライバシー情報を削除し，再度 URL にアクセスする．そして，キャッシュをローカルファイルシステム上に保存し，3 の Local Database を更新する．最後に 5 の Central Server にキャッシュ情報を送信するとデータの保存が完成する．

4.2 データの送信機能

図中の Other Nodes から過去のキャッシュに関して問い合わせがあった場合，4 の Reception 機能が 3 の Local Dabase よりキャッシュのメタ情報を検索し，該当するファイルをファイルシステムから取得する．そのデータを Other Nodes に返し，データの送信が完了する．

5 考察

Web アーカイブに関しては著作権問題をはじめ，多くの問題を抱えており様々な議論がなされている [3]．

インターネット上で Web サイトを公開した人物がそのメディアを削除したとしても，Web アーカイブシステムが既に収集しデータをコピーしている可能性も考えられる．これに対し，Internet Archive は，パブリックアクセスで得られたデータのため，問題ないと解釈している．これらのデータは著作者が削除申請をすればアーカイブを削除するとしている．

他にもホームページの最新のものだけを表示させたく，古いものは表示させたくないという人がいるかもしれない．これに対して Internet Archive は日付順に表示させ，それを見るユーザにその情報がアーカイブであることを認識してもらえば問題ないとしている．

これらの問題に関しては，インターネットを使うユーザのネチケットに依存してくるのではないだろうか．Web アーカイブのメリットをより多くの人が肯定するようになるとさらなる発展が考えられる．

6 おわりに

本研究では，ネットワーク分散型の Web アーカイブ方法について提案した．従来の Internet Archive と違い，人が見たページをキャッシュし，分散して管理することにより，データの収集量および，システムの可用性が高まったといえる．

参考文献

[1] 鵜川義弘. Internet archive. http://edb.miyakyo-u.ac.jp/ugawa/20001201/iArchive.html, 2001.
[2] マイナビ. 国会図書館、公的機関サイトの自動収集開始 - web アーカイブ事業, 2010. http://news.mynavi.jp/news/2010/04/02/006/index.html.
[3] dhr. インターネットアーカイブの諸問題. http://www3.ocn.ne.jp/~dhrname/simple_webarchive2.htm.