惑星間ファイルシステムを使用した分散 Web アーカイブシステムの開発

中村 勇太

本研究では保存，収集に関する既存の課題を解決する Web アーカイブシステムを開発した。本研究で構築したシステム(以下本システム)は分散保存を導入し，収集したデータは単一のサーバに依存しない形で保存できる。また，ユーザの要求に応じて収集する方式により，ユーザのニーズに適合するアーカイブ構築が可能である。

現在，様々な組織で Web サイトを保存するための Web アーカイブが構築されている。これらの Web アーカイブのデータの保存および提供は運用組織に依存するという問題がある。これを解決するため，Web アーカイブを組織に依存せず永続的に保存する仕組みが求められる。また，既存の Web アーカイブはバルク型と呼ばれる Web サイトを無作為に収集する方式が中心であり，ユーザが求める Web ページの版をすべて収集できるとはかぎらない。

本システムでは，保存データを運用組織に依存する課題の解決のため，データの保存方式に分散ファイルシステムを導入した。分散ファイルシステムは，ネットワークに参加する多数のノードにデータを分散して保存する。本システムでは分散ファイルシステムの実装として，P2P ネットワークを使用した IPFS (InterPlanetary File System: 惑星間ファイルシステム)を採用した。

さらに，ユーザのニーズに適合したアーカイブを構築するため，本システムではユーザからの収集要求に応じ，その都度クローラが収集を行う仕組みを取り入れた。

本システムは要求受信機能，収集管理データベース，制御機能，収集機能，保存機能から構成されるサーバと，ユーザの Web ブラウザに組み込まれる要求送信機能からなる。まずユーザが要求送信機能を使用して Web ページの収集要求をサーバに送信する。要求受信機能がその収集要求を受信し，Web ページ管理用データベースに登録する。続いて制御機能が定期的に Web ページ管理用データベースを巡回し収集要求を処理する。最後に収集機能が Web ページを収集し，収集されたファイルを保存機能が IPFS に保存する。

本システムの評価として，複数の Web ページを収集したデータの再現度を確認した。収集したデータはテスト用の IPFS ノードに保存した。収集の対象は静的なコンテンツが主なサイトを 8 種類，アニメーション等を含む動的なサイトを 2 種類選択した。

収集結果を元ページと比較すると，URL とコンテンツが一対一対応するサイトでは，指定した URL のページ内容がほぼ再現されていた。しかし，公立図書館の蔵書検索フォーム等，ユーザの入力に応じてコンテンツを生成するサイトでは，うまく収集ができなかった。これはクローラソフトウェアである Heritrix の仕様面での限界であると考えられる。

残された課題として IPFS に保存したアーカイブを閲覧する機能の実現方法の検討があげられる。

（指導教員　阪口 哲男）