社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－FI－73 (10)
2003／11／13

情報発信組織主導のWebアーカイブシステム

柊和佑*, 阪口哲男**, 杉本重雄**, 田畑孝一**
*図書館情報大学
**筑波大学図書館情報学系
*, **〒 305-8550 茨城県つくば市春日 1-2
*E-mail: ragi@ulis.ac.jp
**E-mail: {saka, sugimoto, tabata}@slis.tsukuba.ac.jp

概要

World Wide Web (以下 Web) により提供される情報の重要度が増し、それに伴って過去に Web で提供されていた情報を保存し、将来にわたって提供する必要性が高まっている。しかし、Web ページの更新作業に加えて過去のものを維持管理するのは大きな負担となる。現在、インターネットで過去の情報を保存し提供する Web アーカイブシステムが存在しているが、それらは情報発信組織と独立に行っているものであり、情報発信組織の意向をくんだアーカイブを行っていない。そこで本研究では、情報発信組織が自ら責任をもってアーカイブを行うためのシステムを構築する。本システムは、更新作業と収集蓄積を連携させることで取りこぼしの無いアーカイブを可能にし、アーカイブ内で Web ページの URL 変遷を追跡できるようにし、ゲートウェイを利用し収集時そのままの状態を再現する。また本システムは、各情報発信組織が主導してアーカイブを管理提供するために、収集、提供時に情報発信組織の意向に基づいて動作する。本稿では構築するシステムと、その開発状況について述べる。

キーワード

インターネット, Web アーカイブ, リンク切れ

A Web Archiving System for Information Providers

Wasuke Hiiragi*, Tetsuo Sakaguchi**, Shigeo Sugimoto**, Koichi Tabata**
*University of Library and Information Science
**Institute of Library and Information Science, University of Tsukuba
*, **1-2, Kasuga, Tsukuba, Ibaraki, 305-8550, Japan
*E-mail: ragi@ulis.ac.jp
**E-mail: {saka, sugimoto, tabata}@slis.tsukuba.ac.jp

Abstract

The spread of the Internet makes information provided by the World Wide Web (Web) important, and providing old information that archived from the Web is more needed. However, update of current Web pages and maintenance of old pages are heavy load for information providers. Currently, some archives of Web pages can be used to get old web pages and they do not have relation with information providers of the Web pages. This paper describes the archiving system, which is designed for information provider's uses. The system consists of a collection sub-system that is cooperating with Web servers of information providers and a retrieval sub-system that provides the same contents as originals. Archiving policies that are described by information providers are used to collect Web pages and to provide archived contents. Information providers can manage the archiving system by the policy description. This paper also describes the development of the archiving system.

Keywords

Internet, Web archiving system, dead link

※本論文は第24回ディジタル図書館ワークショップの論文です.

1. はじめに

近年、インターネットの普及により World Wide Web (Web) の利用者が増加している。Web は、情報発信組織が情報を発信者自身によって迅速に利用者に届けることができる。しかし、Web を用いて情報を発信するということは、それだけで負担となるため、発信される情報は常に現在のことが優先され、負担の大きい過去の情報を将来に渡って責任をもって提供し続けることは行われていない場合も多い。さらに、該当する Web ページの URL の変更や、削除されてしまい、いわゆるリンク切れにより過去の情報が提供されなくなることもある。Web ページの利用者にとって Web ページが更新される前の情報を永続的に得ることは難しいのである。

最近は、このような更新される以前の Web ページを取得するために、Web アーカイブというシステムを利用することができる。Web アーカイブとは、Internet Archive [1] に代表される、インターネット上の Web ページを将来においても利用可能な形で蓄積し提供を行うサービスである。しかし、従来の Web アーカイブは情報発信組織から独立した組織によって行われており、収集方法、蓄積方法、提供方法は、アーカイブを管理している組織の考え方に基づいて行われている。そのため、提供の可否や更新状況に応じた取りこぼしのない収集等、情報発信組織のアーカイブに対する要求に対応することは難しい。また、URL の変更のように情報発信者でなければ解らないことがおきた場合への対応は難しい。

そこで、本研究は、Web を用いて情報を発信している人間や組織自身が運用することにより、自らの Web ページを自分たちの方針に基づいて収集、維持管理することで、過去の Web ページを将来にわたっても利用可能な形で提供するためシステムの開発を目的とする。

2. Web アーカイブ

2.1 Web ページの構成とそのアーカイブ

通常 1 つの URL を指定すると、Web ブラウザには 1 つの Web ページが表示される。1 つの Web ページは、Web ページの HTML データ、インラインイメージとして表示される画像データ、音楽を流すための音楽データといったような様々な要素から構成されている。そのため、Web ページを蓄積するためには、それらのデータをすべて取得し、整理して蓄積する必要がある。さらに、現在インターネットに公開されている Web ページは、HTML データ形式のものだけではなく、プログラムによって出力されるものや、Flash のようなアニメーションによって構成されたもの、PDF や画像ファイルだけのものなど、様々な種類が存在する。

静的な Web ページであれば HTML データと、それに付随するイメージや音楽のデータ等を収集すれば問題はない。しかし、プログラムによって生成される Web ページについては、サーバに保存されているのはプログラムコードであり、利用者に Web ページとして示されるのはプログラムの実行により生成されたデータである。この場合、プログラムコードと生成されたデータ、どちらを蓄積するか区別する必要がある。なお、本稿では前者を内部形式、後者を外部形式と呼ぶ。

同一の URL を時間をおいて複数回収集すると、1 つの URL に対して複数の Web ページが収集した回数蓄積される。そのため、URL だけでなくいつ取得したのか、という情報を使わなければ蓄積した時点の Web ページを再現することはできない。そのため、アーカイブでは各データには取得日時という情報が付加される。そして、本システムではこのように取得日時まで特定した Web ページのことを「リソース」と呼ぶ。

通常の Web ページは、前述のように様々なデータを組み合わせたものになるため、リソースの取得時点における各構成要素を取得することになる。このリソースを構成する要素を「コンポーネント」と呼ぶ。

コンポーネントもURLと取得日時によって識別される。また、リソースは取得日時ごとに存在するため、結果として1つのURLで複数のリソースを示すことになる。このような、更新された日時の異なる複数のリソースをまとめて「リソース集合」と呼ぶ。

さらに、論文の各ページを画像として提供している場合の1ページのように、単体では意味をなさないリソースやリソース集合を、コンテンツ作成者やアーカイブ管理者の判断によって、まとめて扱うことができるようにする。このまとまりを「グループ」と呼ぶ。

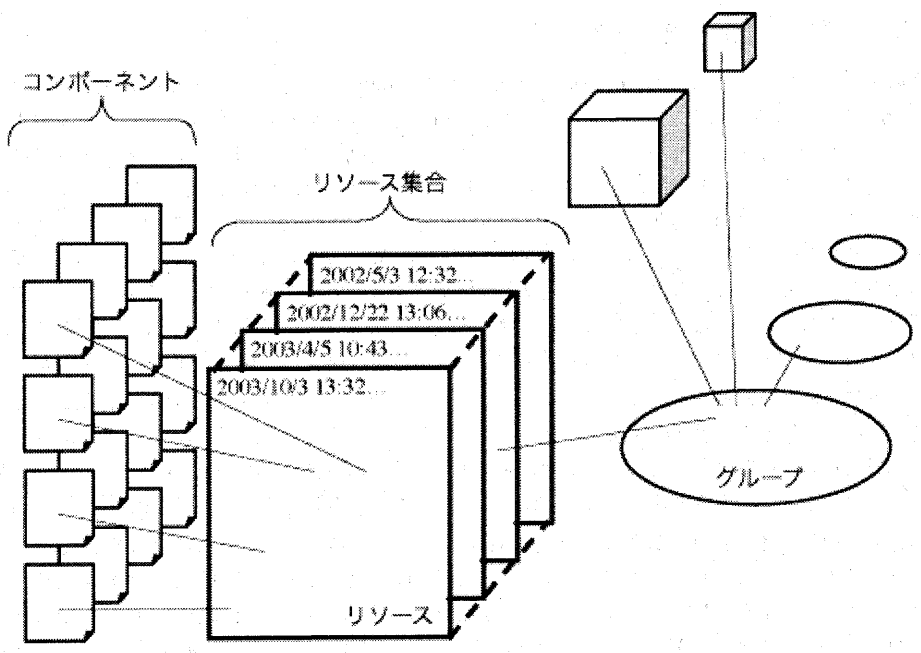

図1　Webページの構成とそのアーカイブ

2.2 リンク切れとアーカイブの利用

Webページを閲覧していると、URLが変更になったというメッセージを頻繁に見かける。これは、情報発信組織の改変でドメイン名が変わった場合など、情報発信組織の都合によってWebページの構成が変わったときなどに発生する現象である。その場合、情報発信組織は新しいWebページに誘導するための情報を発信するが、その情報も一定期間後に削除されてしまうことがある。このように、URLがすでに存在していないWebページを示している場合をいわゆるリンク切れと呼び、その対策が問題となっている[3]。

最近は、既存のWebアーカイブシステムを利用することにより、リンクが切れる以前のWebページを発見することは可能である。しかし、当該Webページが現在どのURLになっているかはわからない。また、あるURLのWebページが以前は別のURLであったとしても、それがどのようなURLであったのかもわからないのである。これは、現在のWebの仕組みでは、URLの変更を示す情報を情報発信者が明示的に示さないと確認できないことから生じる問題である。この問題を解決するためには、過去にどのようなURLであったのか、現在はどんなURLになっているのか、という情報を記録し、提供することが必要

である。閲覧者はその情報を見ることによって、どのようにURLが変遷していったのか確認することが可能となる。

本システムでは、リソース集合に変更前のURLや、いつ消滅したのか、という情報を付与し、URLの変更があってもアーカイブ内でその追跡を可能とする。また、消滅した日時を記述することにより、変更等によって一旦使われなくなったURLが全く別の内容のWebページのURLとして再利用される場合、そのURLについて新規のリソース集合として扱うことにする。

2.3 Webページの収集蓄積とポリシー

既存のアーカイブサービスには、Internet ArchiveやWARP[2]等がある。基本的に自動収集プログラム(クローラ)によってWebページを収集している。自動で行われるため大量のWebページを一定のタイミングで収集していくことができる。しかし、発信されている情報によっては、ニュースのように数分から数日で更新されるものや、企業情報など一度公開されてしまえば1年に数回も更新されないWebページも存在する。そのため、自動収集型のアーカイブシステムではWebページによって収集のタイミングを変えるものも存在する。これは、それぞれのWebページの更新頻度を推測し、それに見合った周期で収集を行うものである。しかし、この方法でもクローラだけでは更新されていくWebページを確実に収集するためには不十分である。取りこぼしのないように収集の頻度を上げることで対応することも可能だが、そのために増加するサーバやネットワークへの負荷の問題やリソースの取得にかかる時間による限界がある。

クローラによる収集が難しいのは、アーカイブシステムがWebページの更新とは独立に収集していることにある。情報発信組織の更新作業がいつ行われたかが判れば、そのタイミングにあわせてWebページを収集すればもれのない収集ができる。そこで、本システムでは情報発信組織がWebページの更新に利用するコンテンツ管理システム等と連携することにより、アーカイブの収集タイミングを更新作業に合わせる。

この方法で収集する対象は更新された時点のWebページである。プログラムによって生成されるWebページは、そのもととなるデータの更新に様々な方法を用いるため、更新時点にその内容を収集することはできないうえ、内部形式はプログラム等により独自のものとなるため、外部形式で収集しなければ将来の利用が難しい。そのようなWebページのためには、クローラによる定期的な収集を行う必要がある。

このように、情報発信組織はアーカイブに内部形式、外部形式を区別して収集させたり、対象外のWebページについてはアーカイブに収集しないよう指示する必要がある。このような指示を行うため、本システムでは一定の書式に基づいて記述を行い、事前にアーカイブに登録しておく。これを「収集ポリシー」と呼ぶ。

2.4 アーカイブの提供

現在、Webページを閲覧するために様々なWebブラウザ(ブラウザ)が使われている。通常、Webアーカイブの利用についてもブラウザを用いることになる。通常のブラウザは現在のWebページを閲覧することが目的であるため、URLのみを指定して現在のWebページを取得する機能しか持っていない。しかし、アーカイブではURLだけでなく取得日時も指定しなければリソースを指定できない。そのために、なんらかの方法でURLと取得日時をアーカイブ内を示すものに書き換えなければならない。

そこで、本システムではゲートウェイを用意し、ブラウザのリクエストをアーカイブに合ったものに変更することにした。これにより、リソースに手を加えることなく閲覧者に提示し、現在利用されている様々なWebページの閲覧ツール等をそのまま利用することができる。

また、アーカイブの内容によっては全ての閲覧者に提供しない性質のものもある。情報発信組織のイントラネット内だけで閲覧するものや、一定期間以上前のものは提供したくないといったものである。本シ

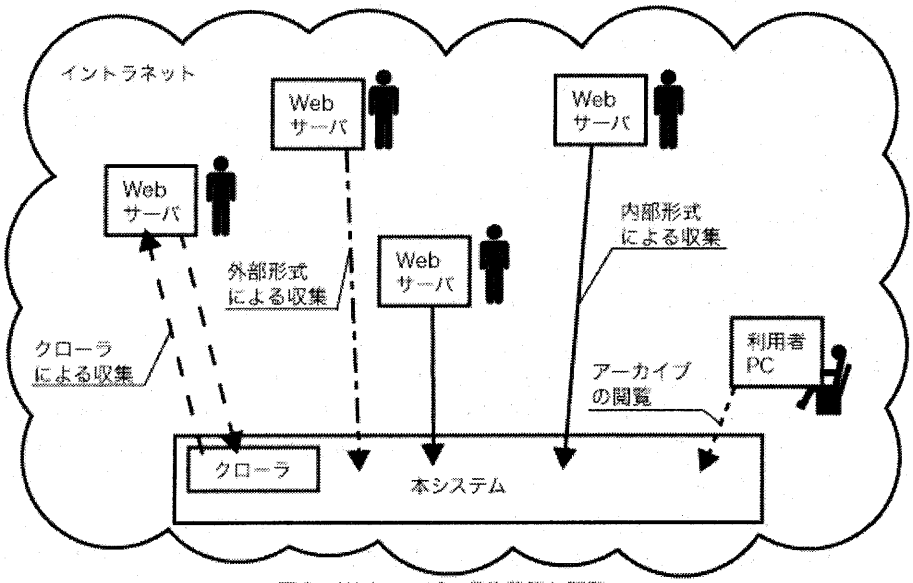

図2　Webページの収集蓄積と閲覧

ステムでは、前述した収集ポリシーと同様に、提供に関する記述も行えるようにする。この、情報発信組織が前もって指定しておく提供に関する制限についての指示を「提供ポリシー」と呼ぶ。

3. 本システムの機能

3.1 収集

本システムでは、アーカイブを行う組織は情報発信組織、情報発信組織と何らかの関係をもつものとし、コンテンツ管理システムを利用して Web ページの更新を行った際にアーカイブ内のデータも更新する方法をとる。静的な Web ページに関しては更新作業ごとに、更新のあった各コンポーネントを収集し、アーカイブに蓄積する。この場合、各コンポーネントは内部形式のままアーカイブされる。また、プログラムによって生成されるような動的な Web ページを収集するために、外部形式での収集も行えるようにする。また、掲示板のような動的な Web ページの場合その内容を、更新作業とは別に利用者がブラウザを利用して新たに加えることがあるため、クローラによる一定期間ごとの収集も行えなければならない。

本システムでは、これらの収集方法を情報発信組織が収集ポリシーを組み合わせることにより自由に設定出来るようにする。ポリシーには URL のパターンを示し、そのパターンごとに指示をおこなう。その指示には、前述のものに加え、示された URL を収集の可否も含む。

3.2 蓄積

本システムでは蓄積しているコンポーネントに収集時の URL と収集日時をメタデータとして付加している。メタデータはコンポーネント、リソース、リソース集合、グループのそれぞれを単位として付与する。実際に保存されるデータはコンポーネント単位であり、リソースとリソース集合、グループについてはメタデータによってその構成を表す。

コンポーネントメタデータは、コンポーネントを特定するための以下の4要素から成る。コンポーネントのURL(componentURL)と収集された日時(componentDate)があり、この二つでコンポーネントの識別子となる。またコンポーネントが、どのような種類のデータであるかを示す要素(componentType)は、HTTPのContent-Typeを記述する。コンポーネントを見るために必要な条件を記述したメモ(componentMemo)は、収集ポリシーに情報発信者が手作業で記述しておくことで自動的に付加される。

リソースメタデータには、リソースの識別のための要素と、リソースを構成しているコンポーネントが記述されている。閲覧者から要求されたリソースの構成コンポーネントを特定することが主な目的となる。リソースはURL(recourceURL)と取得日時(resourceDate)を識別子としている。また、リソースを構成するコンポーネントの識別子を列挙する要素(resourceComponent)をもつ。

リソース集合メタデータは、リソース集合の時間的な変遷を表したメタデータである。リソース集合の発生した日時(timeResourceBirth)と消滅した日時(timeResourceDeath)と、過去のURLの変遷(timeResourcePast)を記録しており、これによって過去にURLが変更されているようなリソース集合も、遡って参照することができるようになっている。また、リソース集合の識別子はURL(timeResourceURL)と、発生日時の組み合わせである。

グループメタデータには、情報発信組織の定めたリソース、リソース集合間の関係が記述されている。メタデータには、リソース集合およびリソースの持っている、複数で1つのことを表している関係や原文翻訳文関係等についての記述(groupType)と、所属しているリソースおよびリソース集合の識別子(groupResource)、このグループの所属する親グループについての記述(groupMother)がある。

3.3 提供

アーカイブ内のリソースを閲覧する場合、日時を指定してWebページを取得する必要がある。既存のブラウザはURLのみを指定するので、本システムではそのURLに加えて日時の指定を行う専用のユーザインタフェースを準備する。

ユーザインタフェースに閲覧したい日時と利用者情報を入力し、ブラウザを利用してブラウジングを行えば、指定した日時のWebページが表示される。その際、入力された利用者情報は情報発信組織が前もって指定しておいた「提供ポリシー」と照らし合わされ、提供に関しての制限などに用いられる。提供ポリシーには許可するユーザやホストとアーカイブ提供期間についての記述を行う。利用者によってリソースが要求されると、リソースのURLと記述された指示をもとに、提供してもよいリソースなのかを判断する。現在考えている指示は、閲覧可能ユーザの限定、現時点で見せてもよいリソースか、一部しか見せないのか全部見せるのか、の3つがある。収集ポリシーと同様に、指示は組み合わせて記述する。

4. システムの実現

本システムは現在以下のような環境、構成で開発を進めている。

4.1 開発環境

現在、本システムはOSとしてFreeBSD、Webサーバおよびサーブレットの利用のためTomcatを使用し、JAVAの開発環境としてJavaSDK1.4.1を利用している。データベースにはPostgreSQLを使用した。閲覧用プログラムの開発には、OSとしてMacOS10.2.8、JAVAの開発環境としてJavaSDK1.4.1を使用している。

4.2 システムの構成

本システムの構成を図3に示す。本システムはいくつかのサブシステムで構成されている。サブシステムには、情報発信者の更新作業に合わせてWebページを収集し、メタデータを付与したうえでリソースとしてデータベースに蓄積する収集蓄積サブシステムと、利用者の要求に合わせてアーカイブされたリソースを特定し利用者に提示する提供サブシステムが存在する。Webサーバにはコンテンツ管理システムを置き、収集サブシステムと連携する。また、情報発信組織が記述する各ポリシーは、各サブシステムとは独立した状態で置かれている。

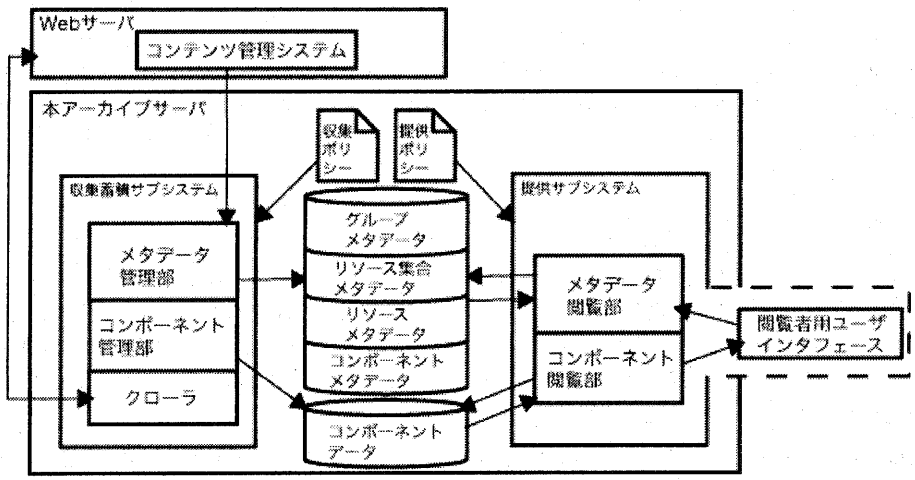

図3　システムの構成

4.3 収集蓄積サブシステム

収集蓄積サブシステムは内部にコンテンツ管理部、コンポーネント管理部、メタデータ管理部とクローラを備える。

コンテンツ管理システムは、情報発信組織によるWebページの編集を仲介するシステムで、コンポーネントと各メタデータを、情報発信組織の更新作業に合わせてコンポーネント管理部とメタデータ管理部に渡す部分を担当し、更新・削除といった操作をメタデータ用に解釈する。なお、情報発信組織が操作するのはコンテンツ管理用に用意した管理用Webページとなる。クローラは、収集ポリシーによって指定された、定期的な収集が必要なリソースを収集する部分である。コンポーネント管理部、メタデータ管理部は、コンテンツ管理部およびクローラが取得したリソースとメタデータをデータベースに登録する部分である。その際、送られて来たコンポーネントについては収集ポリシーによって蓄積するかしないかの指定が行われることもある。リソースの更新の場合はコンポーネントの保存、コンポーネントメタデータの生成、リソースメタデータの生成を行う。コンポーネント管理部より、削除やグルーピング等の指示が出された場合にはリソース集合メタデータ、グループメタデータの変更が行われる。

4.4 提供サブシステム

提供サブシステムは内部にメタデータ閲覧部、コンポーネント閲覧部をもつ。また、閲覧者の環境に設置されるユーザインタフェースもこのサブシステムに含まれる。

メタデータ閲覧部、コンポーネント閲覧部は、閲覧者によって示された日時情報付きのURLを、メタデータを参照してリソースに直し、閲覧者に提示する部分である。日時指定ユーザインタフェースは閲覧者のPC上にあり、日時の指定用インタフェースとプロキシの機能をもっている。なお、プロキシは、アーカイブシステムとブラウザの橋渡しをするゲートウェイの役割をする。

コンポーネント閲覧部から送られて来たコンポーネント群はそのままブラウザに渡され、閲覧者にWebページとして提示される。このように専用のゲートウェイを利用するのは、アーカイブされたリソースを収集時と変わらない状態で提供するためである。現在のシステムはブラウザを利用した人間による閲覧を中心にしているが、ゲートウェイを利用しているため、計算機による自動処理にも対応していけると考えている。

5 おわりに

本研究では、情報発信組織が自ら維持管理するWebアーカイブシステムの構築を目的としている。本稿では、更新作業と連携した収集蓄積方法と、ゲートウェイを利用した提供方法を中心に、システムの概要を説明した。また、情報発信者の収集と提供の条件を指定するポリシーのアイデアを示した。Webページを再現することを目的としてメタデータの定義も行い、それにより収集蓄積システムだけでなくURLの変遷を追跡管理する方法も示した。

今後は、各メタデータを利用して、どのようにアーカイブされたリソースを提示していくか、実装方法含めて検討し開発を進めていくつもりである。

参考文献

[1] Internet Archive. http://www.archive.org/. 2003/10/8

[2] WARP. http://warp.ndl.go.jp/. 国立国会図書館, 2003/10/8

[3] 柊和佑, 阪口哲男. WWWページ検索結果の選択における利用者支援. 情報技術レターズ. Vol.1. 2002, p.221-222.