Webアーカイブにおける時系列閲覧：単一コレクションへの適用

Browsing Algorithm with Time Consistency in Single Collection of Web Archive

小城 正士[1] 廣瀬 信己[2] 河野 浩之[3]

Masashi KOJO Nobuki HIROSE Hiroyuki KAWANO

近年，Web 情報の文化的・社会的価値に着目し，それらを保存する試みが世界各国で進められている．我が国では，国立国会図書館インターネット資源選択的蓄積実験事業 (WARP: Web ARchiving Project) による Web アーカイブ構築が進められているが，データ収集・保存・運用に多くの技術的課題が存在する．本論文では，収集データの時系列管理に焦点を当て，Web アーカイブコレクション閲覧における時系列一貫性のあるリンク参照を，コレクションの収集期間と閲覧対象期間との関係から決定するアルゴリズムを提案する．

In many countries, web archive projects have been promoted continually for preserving cultural and social properties on web systems. In Japan, a project called WARP (Web ARchiving Project) in the National Diet Library was promoted, we have many technical issues of archiving systems such as collection, preservation and management of web archive data. In this paper, we focus on the technical issues of the time series management of web archive collections, and propose an algorithm that identifies valid navigation of links preserving the consistency while browsing web archive, which is based on the relationship between collecting and browsing periods.

1. はじめに

現在，インターネットの Web システム上に流通する情報量は，表層部に 167TB 以上，深層部に 91,850TB 以上存在すると推定される．しかし，知識流通基盤となるインターネット上の多様な情報は，従来の出版物に比べ，空間的・時間的に情報内容と存在が安定しない問題点を抱える．そこで，Web 情報を文化資産として体系的に蓄積し，将来に渡って長期保存するウェブアーカイブ (Web Archive) プロジェクトが，各国国立図書館等を中心に推進されている [1].

プロジェクト遂行に関わる主要技術は，Web データ収集・検索・保存に大別されるが，単調増加する Web ページを時系列順に長期保存するため，ページ更新時に上書きを行うサーチエンジンと異なる保存機能を必要とする．なお，我々は，長期保存・大容量保存の問題に関して，階層型ストレージシステムを用いる手法の提案と評価を行った [5].

本稿では，時系列管理に焦点を当て，アーカイブ運用における時系列一貫性をもつ閲覧手法を提案する．一般に，データベースにおけるトランザクションの一貫性は，個々のイベント発生時刻に基づいて議論する．そこで，提案アルゴリズムは，収集時に巡回できたデータ集合である個々のコレクションを一貫性があるものと定義し，各データ間の閲覧可能性を，データ更新時刻により求めるものである．

[1]正会員　日本 IBM
[2]非会員　国立国会図書館総務部
[3]正会員　南山大学数理情報学部情報通信学科
kawano@it.nanzan-u.ac.jp

2. Web アーカイブの現状と問題点

Web アーカイブは，バルク収集と選択的収集のアプローチに大別される．バルク収集とは一国全体や世界全体の Web 情報を一括収集する方法であり，選択的収集とはサイト単位や資料単位でセレクションし，著作権処理を行い収集する方法である．以下，代表的 Web アーカイブのデータ管理手法を紹介し，その閲覧における問題点を述べる．

2.1 WARP

我が国の国立国会図書館が推進する「国立国会図書館インターネット資源選択的蓄積実験事業 (WARP: Web ARchiving Project)」は[1]，電子雑誌・政府ウェブ・協力機関ウェブのコレクションからなり，著作権等を処理しながら，選択的収集を行っている．2004 年 6 月 30 日現在，電子雑誌 1,108 タイトル，政府機関 10 タイトル，協力機関 598 タイトルを所蔵する．今後，法制度面について国立国会図書館長の諮問機関である納本制度審議会を経た後，立法措置が行われ，2006 年より日本国の Web に対象を広げた収集を開始する予定である [3].

WARP は，一定期間ごとに Web ページを収集し，収集日毎にコレクションを形成し，ソースファイルのリンク先アドレスをアーカイブ内のアドレスに書き換えることで，収集サイトの状態を保存する．そのため，収集日以外のサイト状態は保存されない．また，収集対象コンテンツ以外は収集しないため，他サイトへのリンク先の内容等も保存されない．その他，書き換えの困難なファイル (PDF，Flash 等) に対応出来ないこと，原本性保証などの問題がある．

2.2 Internet Archive

Internet Archive[2]は，現在公開されている最大の Web アーカイブである．ページ内リンクのクリック時に，Java Script によるアーカイブ内のアドレス付加を用いている．なお，指定アーカイブのアドレスにデータが存在しない場合，図 1(a) のように，その時点以前で最も近い時間のデータを表示する．そのため，意図したデータと全く違うデータが表示されることがある．

[1]http://warp.ndl.go.jp/
[2]http://www.archive.org/

また，ページ間のリンク参照を繰り返すと，時間を逆行する閲覧が生じる．例えば，図 1(b) において，12/20/12:00 に収集された xxx.html(1) 内の yyy.html へのリンクを参照した時に，同時間に収集されたデータが存在しない場合，最も近い 12/20/6:00 の yyy.html(2) が開かれる．ここで，(1) に戻るつもりで xxx.html へのリンクを参照しても，(1) と (2) は同時間内にないために，(2) 以前で最も近い 12/20/0:00 の (3) が開かれてしまう．以後同様に，ページ間のリンク参照を繰り返すと，(3) → (4) → (5) の時間逆行の閲覧が生じる．

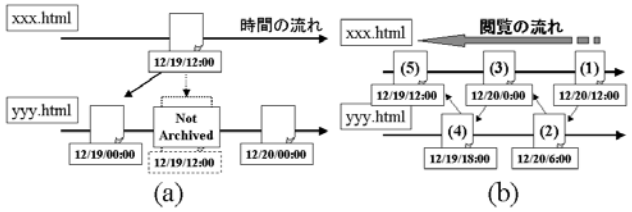

図 1: Internet Archive の手法
Fig. 1 Format in Internet Archive

2.3 NWA

NWA(Nordic Web Archive) は，北欧各国の国立図書館を主体としたプロジェクトである[3]．NWA Toolset[2] のブラウザは，同一アドレスのデータが複数バージョンある場合，図 2(a) のように，時間軸上のポインタとして表示し，指定日時の閲覧バージョンを表示する．また，軸上のポインタ以外を指定した場合，その点より前の一番近いバージョンを表示する．

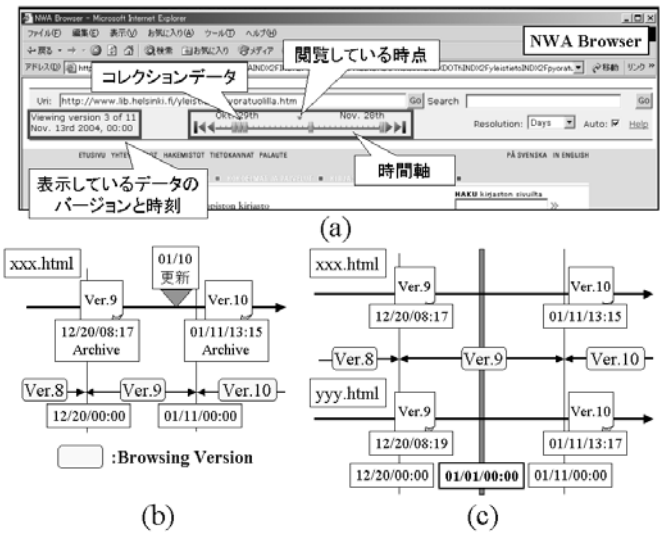

図 2: NWA の手法
Fig. 2 Format in NWA

例えば，図 2(b) において，ページ xxx.html は 12/22/08:17 に Ver. 9 が，01/11/13:15 に Ver. 10 が収集されているので，時間軸上で 12/22/00:00 から 01/11/23:59 までを指定した場合は Ver. 9 が，01/11/00:00 以降を指定した場合は Ver. 10 が表示される．また，ページ間のリンクを開く場合には，時間軸上の位置によって開くバージョンが異なるため，図 2(c) において，xxx.html から yyy.html へのリンクを開く際に，閲覧している時間が，yyy.html の Ver. 9(12/22) と Ver. 10(01/11) の中間点である 01/01/00:00 より前の時は Ver. 9 を，それ以降の時は Ver. 10 を開く問題などが生じる．

[3]http://nwa.nb.no/

3. 時系列一貫性アルゴリズムの提案

本稿で焦点を当てる時系列管理は，同一 URL のデータが更新された時に，異なるバージョンとして蓄積されたコンテンツを管理・運用する技術であり，アーカイブが飛躍的に増大するにつれて重要となる課題である．

前述のように，既存の Web アーカイブは，収集コレクションの閲覧単位として，収集時の状態再現を行っているが，実際のユーザーによる閲覧では，収集日以外のサイト状態やサイト間のリンク関係に基づく閲覧要求が生じると予想される．したがって，単一のコレクション内に限定せず，アーカイブ全体にわたり時系列的な一貫性の高い閲覧アルゴリズムが重要である．

なお，一般のデータベースにおいて，その一貫性は，トランザクション発生時刻に基づいて決定している．よって，同様に，Web アーカイブにおいても，収集コレクション内の個々のデータ更新時刻から，異なる時点の閲覧イメージの決定が可能である．ただし，コレクション間では，その時点の実際の Web ページの状態が不明であり，閲覧イメージを簡単に決定できない．そこで，収集時に巡回できたデータ集合であるコレクションを一貫性があると定義し，データ間の一貫性を決定することを提案する．なお，一貫性を決定するパラメータとして閲覧対象期間を用いることにする．

3.1 収集コレクション内の一貫性のある閲覧

本稿で用いるパラメータを表 1 に示す．なお，コレクション $\boldsymbol{S_n}$ は，n 回目の収集で集められたデータがアーカイブされた時刻順に並んでいるものとする．

$$\boldsymbol{S_n} = \{P_0(n) \cdots P_i(n) P_j(n) \cdots\}$$
$$T_A(P_0(n)) \leq \cdots \leq T_A(P_i(n)) \leq T_A(P_j(n)) \leq \cdots \quad (1)$$

$\boldsymbol{S_B}$ はアーカイブされたデータの内，閲覧対象期間 $[T_s, T_e]$ 内において閲覧可能なデータの集合を表す．これに含まれるデータは，期間内において，単独で全て閲覧可能であるが，データ間のリンクに基づく閲覧は，$[T_s, T_e]$ の状態に依存する．データ $P_i(n)$ からデータ $P_j(m)$ への閲覧が可能であるとき，その状態を $\langle P_i(n), P_j(m)\rangle$ で表し，相互に閲覧が可能である状態を $\langle\langle P_i(n), P_j(m)\rangle\rangle$($\langle P_i(n), P_j(m)\rangle$ かつ $\langle P_j(m), P_i(n)\rangle$) で表す．また，集合 $\boldsymbol{S_B}$ における可能な閲覧状態の集合を $\boldsymbol{L(S_B)}$ で表す．

ところで，Web ページは時々刻々更新され，ある時間内で意図していた閲覧と別の時間内で意図していた閲覧は，内容更新に応じて異なると考えられる．そのため，その時点

表 1: 各種パラメータ
Table 1 Parameters

Crawling		
$[t_s(n), t_e(n)]$	:	n 回目の収集実行期間
$t_s(n)$	:	開始時刻
$t_e(n)$	:	終了時刻
$\boldsymbol{S_n}$	:	収集されたコレクション
Archive Data		
i	:	アドレス
$P_i(n)$	:	n 回目の収集において保存された i のデータ. (存在しない場合は ϕ)
$T_A(P_i(n))$	:	$P_i(n)$ が Archive された時刻
$T_U(P_i(n))$	:	収集時に取得された $P_i(n)$ の最終更新時刻
Browsing		
$[T_s, T_e]$	:	閲覧対象期間
T_s	:	起点
T_e	:	終点
$\langle P_i(n), P_j(n) \rangle$	:	$P_i(n)$ から $P_j(n)$ への閲覧
$\boldsymbol{S_B}$	:	Archive の中の閲覧可能なデータの集合
$\boldsymbol{L(S)}$	:	集合 $\boldsymbol{S}$ における一貫性のある閲覧の集合

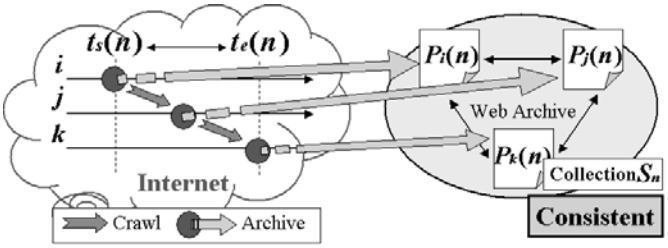

図 3: コレクション内の一貫性
Fig. 3 Consistency in a Collection

の正確な Web の状態は，アーカイブデータに基づいて再現出来ない．しかしながら，Crawler により収集されたコレクションは，少なくとも Crawler の収集時間範囲に存在し得た正当な状態であり，そのリンク構造は収集時間内において一貫性があると言える．よって，図 3 のように Crawler によって収集されたコレクション $\boldsymbol{S_n}$ を，その収集が行われた時間内 $[t_s(n), t_e(n)]$ において一貫性がある (Consistent) と定義し，$\boldsymbol{S_n}$ 内のデータ間の閲覧状態は全て正当 (閲覧可能) と考える．

$$\boldsymbol{L(S_n)} = \{\langle\langle P_i(n), P_j(n)\rangle\rangle\}, (\forall i, j \ P_i(n), P_j(n) \in \boldsymbol{S}_n) \quad (2)$$

以下，このコレクションにおける一貫性のある閲覧状態の集合を基準に，個々のデータ間に対する閲覧の一貫性を，閲覧対象期間 $[T_s, T_e]$ との関係を用いて議論する．

3.2 閲覧可能なデータ

閲覧対象期間 $[\ T_s, T_e\ ]$ を指定した場合，ブラウザ上で閲覧されるデータは，その期間内に存在した可能性のあるデータでなければならない．アーカイブされた $P_i(n)$ が，その時点での i のデータであったと確実に言える期間は，そのデータの最終更新時刻から収集時刻まで，つまり期間 $[\ T_U(P_i(n)), T_A(P_i(n))\ ]$ である．この期間を “*Determinable*” な期間と呼ぶ．$T_A(P_i(n))$ から次の $T_U(P_i(n+1))$ までの間は，収集が行われておらず，更新の有無が不明であるために，その間の i のデータは $P_i(n)$ であると言い切れない．よって，この期間を “*Indeterminable*”な期間と呼ぶ．この期間のうち，少なくとも一部 (もしくは全部) の期間においては，i のデータは $P_i(n)$ であったと推測される．

ここで，前述の議論で，収集時に収集したコレクションを一貫性があると定義したのと同様，Crawler が収集を行った時間を閲覧対象期間に含む場合，その時点から次に分かっている更新時刻までの間の i のデータを，$P_i(n)$ とする．つまり，データ $P_i(n)$ の *Determinable* な期間 $[\ T_U(P_i(n)), T_A(P_i(n))\ ]$ が $[\ T_s, T_e\ ]$ に含まれるとき，その後にくる *Indeterminable* な期間 $[\ T_U(P_i(n)), T_U(P_i(n+1)))$ においても，$P_i(n)$ を閲覧可能とする．また，閲覧可能なデータが存在しない期間は，その期間内において i のデータはアーカイブされていないものとして，閲覧不可とし，i への閲覧は `Not Found` を返すものとする．

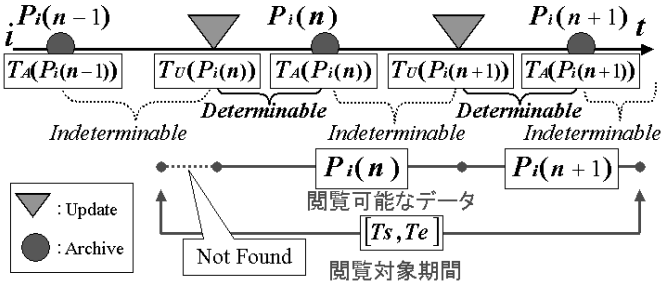

図 4: 閲覧可能なデータ
Fig. 4 Browsable Data

例えば，図 4 では，$P_i(n), P_i(n+1)$ の *Determinable* な期間が $[T_s, T_e]$ に含まれているので，それぞれ期間 $[T_U(P_i(n)), T_U(P_i(n+1)))$,$[T_U(P_i(n+1)), T_e]$ において閲覧可能としているが，$P_i(n-1)$ は *Determinable* な期間が含まれていないので閲覧不可であり，期間 $[T_s, T_U(P_i(n)))$ において i は `Not Found` である．

以上，閲覧対象期間内での個々のデータの閲覧可能性を論じた．しかし，図 4 のように，$[T_s, T_e]$ を広く取った場合，一つのアドレスについて閲覧可能なデータが複数存在する．そこで，複数ある閲覧可能データ中から参照するデータを自動的に決定するために，データ間での一貫性のある閲覧を定義する．

3.3 データ間での一貫性のある閲覧

$[\ T_s, T_e\ ]$ 内でデータ $P_i(n)$ から j へのリンクをクリック時に，データ $P_j(m)$ の閲覧が妥当であるならば，閲覧可能状態の集合 $\boldsymbol{L(S_B)}$ に状態 $\langle P_i(n), P_j(m)\rangle$ を含める．以後，様々な状態の $\boldsymbol{L(S_B)}$ を示す．ここでは，誌面の制約上，単

一コレクションの場合を記述するが，複数コレクションについても [6] で議論を行っている．

単一コレクションに収まる場合

単一コレクションに収まる場合とは，$[T_s, T_e]$ が $n-1, n+1$ 回目の全てのコレクションデータの *Determinable* な期間を含まず，全ての $P_i(n-1), P_i(n+1)$ が閲覧不可な状態を表す．$P_i(n)$ の閲覧可能性は $[T_s, T_e]$ の状態によって異なる．

まず，相互リンクを持つ二つのデータ $P_i(n)$ と $P_j(n)$ 間のリンクについて考えるが，その際に注目すべき点は，各データの更新時刻 $T_U(P_i(n)), T_U(P_j(n))$(常に $T_U(P_j(n)) < T_U(P_i(n))$ であるとする) と $[T_s, T_e]$ の関係である．それぞれの間で成り立つ関係は図 5 で表される．

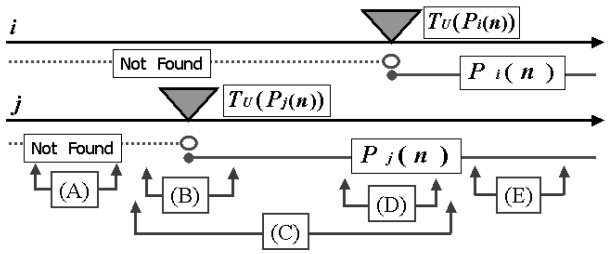

図 5: 閲覧対象期間と最終更新時刻の状態
Fig. 5 Statuses in Browsable Duration and Last Modified Time

Case A. $\boldsymbol{T_s < T_e < T_U(P_j(n)) < T_U(P_i(n))}$
閲覧可能なデータがないので，i, j 共に `Not Found`.

$$\boldsymbol{S_B} = \{\phi\} \tag{3}$$

Case B. $\boldsymbol{T_s \leq T_U(P_j(n)) \leq T_e < T_U(P_i(n))}$
$P_j(n)$ は閲覧可能だが，i の閲覧可能なデータがないため，$P_j(n)$ から i への閲覧は `Not Found`.

$$\boldsymbol{S_B} = \{P_j(n)\}, \boldsymbol{L(S_B)} = \{\phi\} \tag{4}$$

Case C. $\boldsymbol{T_s < T_U(P_j(n)) < T_U(P_i(n)) \leq T_e}$
i と j 共に $P_i(n), P_j(n)$ が閲覧可能である．また，i の更新が行われているので，$P_j(n)$ の意図していた参照と異なっている可能性があるが，i のそれ以前のデータが閲覧不可であるので，$P_i(n)$ を閲覧するのが最も妥当である．

$$\boldsymbol{S_B} = \{P_i(n), P_j(n)\}, \boldsymbol{L(S_B)} = \{\langle\langle P_i(n), P_j(n)\rangle\rangle\} \tag{5}$$

Case D. $\boldsymbol{T_U(P_j(n)) \leq T_s < T_U(P_i(n)) \leq T_e}$
Case C. と同様.

$$\boldsymbol{S_B} = \{P_i(n), P_j(n)\}, \boldsymbol{L(S_B)} = \{\langle\langle P_i(n), P_j(n)\rangle\rangle\} \tag{6}$$

Case E. $\boldsymbol{T_U(P_j(n)) < T_U(P_i(n)) \leq T_s < T_e}$
$P_i(n), P_j(n)$ 共に閲覧可能で，更新も行われていないので，相互に閲覧可能である．

$$\boldsymbol{S_B} = \{P_i(n), P_j(n)\}, \boldsymbol{L(S_B)} = \{\langle\langle P_i(n), P_j(n)\rangle\rangle\} \tag{7}$$

4. むすび

本稿では，Web アーカイブコレクション閲覧に際して，時系列一貫性のあるリンク参照を，コレクションの収集期間と閲覧対象区間との関係から決定するアルゴリズムを提案した．また，提案手法により，既存の Web アーカイブの幾つかの問題を解決できることを述べた．今後，実際のアーカイブへの実装，提案手法の特性に基づく Web データ収集戦略の決定などが必要である．

[謝辞]

本稿の一部は，文部省科学研究費 (16016248) の研究成果による．

[文献]

[1] Abiteboul, S., Cobena, G., Masanes, J., and Sedrati, G., “A First Experience in Archiving the French Web,” Research and Advanced Technology for Digital Libraries, Springer, 2002.

[2] Hallgrimsson, P. and Bang, S., “Nordic Web Archive,” 3rd ECDL Workshop on Web Archives, 2003.8, (online), available from *http://nwatoolset.sourceforge.net/docs/nwa@ecdl2003.pdf*, (accessed 2005.5.9).

[3] 廣瀬信己, “国立国会図書館におけるウェブ・アーカイビングの実践と課題,” 情報処理学会研究報告，Vol.2003, No.51, pp.95-111, 2003.

[4] 河野浩之, 川原稔, “Web 検索におけるテキストマイニング,” 人工知能学会誌, Vol.16, No.2, pp.212-218, 2001.

[5] 小城正士，廣瀬信己，河野浩之，“Web アーカイブにおける長期ストレージシステムの提案,” DBWeb2004，Vol.2004, No.14，pp.33-40，2004.

[6] 小城正士，廣瀬信己，河野浩之，“ Web アーカイブにおける時系列参照アルゴリズムの提案,” DEWS2005，2005. (online), available from *http://www.digitalcity.gr.jp/~satoh/DEWS2005/procs/papers/6A-o4.pdf*, (accessed 2005.5.9).

小城 正士　Masashi KOJO
日本 IBM．2005 京都大学大学院情報学研究科博士前期課程修了．2003 京都大学工学部情報学科卒業．日本データベース学会会員.
廣瀬 信己　Nobuki HIROSE
国立国会図書館総務部．1997 日本長期信用銀行金融商品開発部．1997 東京大学経済学部経済学科卒業．日本図書館協会個人会員.
河野 浩之　Hiroyuki KAWANO
南山大学数理情報学部教授．1997 京都大学大学院情報学研究科システム科学専攻助教授．1990 京都大学大学院工学研究科数理工学専攻博士後期課程研究指導認定退学．情報処理学会，電子情報意通信学会など所属.