2003－DBS－131　(72)
2003／7／18

WWW コンテンツ一貫性管理のためのリンク更新機構

中溝昌佳[†1]　森嶋厚行[†2]　有山智洋[†3]
杉本重雄[†2]　北川博之[†4]

近年，WWW は社会における重要なメディアのひとつとして大きな役割を果たしている．WWW の特徴としては，分散管理，動的な更新，リンクなどがある．これらの特徴は WWW を役立つメディアとする一方で，WWW コンテンツの一貫性管理を困難にする要因になっている．本稿では，データベースにおける一貫性制約の考え方を WWW の文脈に導入し，WWW コンテンツにおける一貫性維持を行うための機構について提案する．特にリンクの一貫性維持に焦点をあてる．提案手法の特徴は次の通りである．(1) 通常の Web アーキテクチャの自然な拡張であり，既存の Web コンテンツとも容易に組合せ可能である．(2) 単純かつ強力な制約記述言語を提供する．

A Link Updating Method for Integrity Management of WWW Contents

Akiyoshi Nakamizo,[†1] Atsuyuki Morishima,[†2]
Tomohiro Ariyama,[†3] Shigeo Sugimoto[†2]
and Hiroyuki Kitagawa [†4]

The World Wide Web (WWW) has become one of the most important media in our society. Its characteristics include distributed management, dynamic updates, and links. The characteristics are not only making the WWW a useful tool, but also making it difficult to manage the integrity of its contents. This paper proposes a method that applies to the WWW 's context the concept of integrity constraints, which is common in database contexts. We focus on the problem of managing the integrity of WWW links. The features of our proposed method are as follows: (1) The method is a natural extension of the ordinary Web architecture. (2) It provides a simple and expressive constraint description language.

1. はじめに

近年，WWW は社会における重要なメディアのひとつとして大きな役割を果たしている．WWW の特徴としては，分散管理，動的な更新，リンクなどがある．これらの特徴は WWW を役立つメディアとする一方で，コンテンツの一貫性管理を困難にする要因になっている．コンテンツに一貫性が無い具体例としては，例えば次のようなものがある．(1) 同じ組織の情報を保持する複数のページで電話番号が異なっている．(例えば，片方では 0298-59-abcd であり，もう片方では 029-859-abcd など)．これは，一方の情報だけが更新されたことによって生じる．(2) リンクをたどると，その参照先のページが存在しない (リンク切れの問題)．(3) 特定のニュースへを参照するために，ニュースページへのリンクを張ったとする．その後，しばらくたつとそのニュースの情報はバックナンバーページへ移動しまい，そのリンクの参照先が意図したものと異なってしまう．

ここでは，以上のような問題をなくすための仕組みを Web コンテンツの一貫性管理と呼ぶ．データベースの分野では一般に，一貫性の管理を行うためにデータベースが満たすべき制約 (一貫性制約[1)]) を記述するというアプローチがとられる．本稿ではこの考え方を WWW の文脈に導入する．すなわち，WWW コンテンツで成立すべき制約を記述し，それを用いて一貫性管理を行うための機構を提案する．特に，リンク切れやリンク先の内容の変更という，リンクの一貫性管理の問題に焦点を当てる．1999 年の調査[2)] によると，Web サイトの平均的なリンク切れの割合は 5.7%であ

†1 芝浦工業大学大学院 工学研究科
Grad. Sch. of Eng., Shibaura Inst. of Tech.
†2 筑波大学 知的コミュニティ基盤研究センター
RCKC, Univ. of Tsukuba
†3 図書館情報大学大学院 情報メディア研究科
Grad. Sch. of Info. and Media Studies, Univ. of Lib. and Info. Sci.
†4 筑波大学 電子・情報工学系
Inst. of Info. Sci. and Elec., Univ. of Tsukuba

る．また，科学教育のための Web ページを調査したところ，20ヶ月後には全リンクの 18.8%が切れていたという報告もある[5]．ジョージア工科大の GVU センターにおける調査では，約 6 割のユーザが「リンク切れは WWW の利用における重大な問題の一つである」と答えており[3]，リンクの一貫性を維持することは重要な問題であると考えられる．分散管理という WWW の性質上，一貫性を完全に保証することは困難であると考えられるが，本プロジェクトでは，可能な限り一貫性を満たすよう WWW が自律的にリンクの更新を行う世界の実現を目指す．

関連システム・研究としては次のようなものがある．まず，各種のリンク切れ発見ツール (リンクチェッカ) が存在する[6][7]．これらは，指定されたページ (群) に記述されたリンクについて，参照先がアクセス可能かチェックし，レポートを作成するものである．リンクチェッカを用いてリンク切れを発見すると，Web サイト管理者にメールを送ることによってリンクの更新管理を行っているサイト[8] も存在する．また，ハイパーメディアデータベース等のコンテキストで，参照経路の一貫性の研究が行われている[4]．しかし我々の知る限り，WWW コンテンツのリンク一貫性維持の自動化の問題に取り組んだ研究は存在しない．

本研究の特徴は次の通りである．(1) 通常の Web アーキテクチャの自然な拡張であり，既存の Web コンテンツとも容易に組合せ可能である．(2) 単純かつ強力な制約記述言語を提供する．

本論文の構成は次の通りである．2 章では，WWW コンテンツの一貫性を管理するためのフレームワークを提案する．3 章では，WWW におけるリンクに関して成立する制約を記述するための言語である，WIDL (Web Integrity Description Language)/Link の説明を行う．また，制約の一種を表現するために，特別な WWW ページである Link Authority の概念を導入する．4 章では，WIDL によって記述された制約を可能な限りみたすようリンク更新を動的に行う機構の提案を行う．5 章では，Link Authority の探索について議論する．6 章では，本プロジェクトの今後の展開について述べる．

2. WIM Server を用いた WWW コンテンツ一貫性管理

我々が提案するフレームワークの重要な構成要素は，WIDL (Web Integrity Description Language) および WIM (Web Integrity Management) Server である．これらは，Web Server と ファイルシステムに格納された Web ページ群，という通常の Web アーキテクチャに追加する形式で利用する (図 1)．これは従来のアーキテクチャの自然な拡張であるため，既存の WWW コンテンツとも容易に組合せ可能である．WIDL は Web コンテンツで成立すべき制約を記述するための言語群である．それらの中で，リンクに関する制約を記述するための言語を WIDL/Link と呼ぶ．詳細は 3 章で説明するが，WIDL/Link では HTML ページや XML ページの各リンク要素の属性として制約を記述する．WIM Server は，管理下の Web コンテンツを監視し，WIDL によって記述された制約のうち Web コンテンツ変更などの理由により満たされなくなったものを発見すると，その制約を満たすように Web コンテンツを更新する (実際に変更先の探索を行うのは，WIM Server に組み込まれた WIM エンジンである)．例えば，あるリンクに対して「リンク切れがあってはならない」という制約が WIDL/Link で記述されていたとする．その場合，WIM Server がそのリンクに対してリンク切れを検出すると，代わりとなるリンク (変更先リンク) の候補群を求める．その際，WIM Server は各変更先リンク候補に対して 0 以上 1 以下の確信度を計算する．後述するが，WIDL/Link では，確信度に関する閾値を指定することによって，WIM Server に対して変更先リンクの候補群を求めるだけでなくリンクの自動更新を指示することができる．具体的には，指定された閾値を超える確信度を持つ変更先リンクが存在する場合には，制約を満たさないリンク l を，それらのうち最も高い確信度を持つ変更先リンクに自動的に更新する．そうでない場合には，変更先リンク候補一覧ページを作成し，l がその変更先リンク候補一覧ページを指すように変更する．

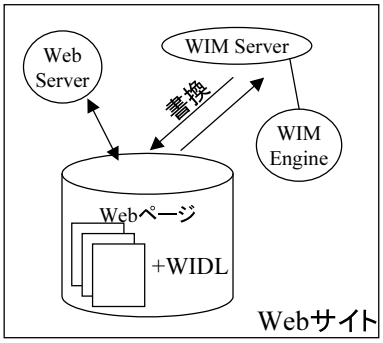

図 1 アーキテクチャ

3. 制約記述言語 WIDL/Link

WIDL (Web Integrity Description Language)/Link

属性	意味
isAlive	リンク切れでない
matches	リンク先が指定の内容とマッチする
follows	リンクが指定の link authority に従う.
threshold	自動的にリンクを更新するための閾値を与える.

図 2 WIDL/Link で用意されている属性

は，HTML ページもしくは XML ページ (以下 Web ページ) で記述されるリンクに関する制約を記述するための言語である．リンクの制約の記述は，Web ページのリンクに WIDL/Link で規定された属性を追加することによって行われる．次に WIDL/Link を用いた制約記述の例を示す．

```
<A wi:matches="*WIDL の開発が終了しました*"
 href="http://www.mlab.info/news.html">
   ニュース</>
```

ここで，wi は WIDL/Link の名前空間であり，matches=e はリンク先のページの内容がパターン e にマッチすることを表す制約である．したがって，この例の制約はリンクが指す先のページに「WIDL の開発が終了しました」という文が含まれていることを表している．WIDL/Link では 4 つの属性を定義している．図 2 にそれらの一覧を示す．属性は 2 種類に分類される．第一のグループ (isAlive, matches, follows) は制約を表すための属性である．残りの 1 つ (threshold) はリンクが制約を満たさなくなった場合，システムが適切なページを探した後，どの基準で自動的にリンクを更新するかを指定する閾値である．

3.1 isAlive

属性 isAlive は,「リンク切れでない」という制約を表す．次に例を示す．

```
<A wi:isAlive
 href="http://ho-expo.org/">
   東大宮博覧会</>
```

isAlive が成立しない状況とは，ページが移動した場合や，ページが無くなった場合などがある．したがって，isAlive が条件として指定されているとき，システムはリンク切れを検知するとそのリンクが制約を満たさなくなったと判断する．

3.2 matches

先に説明したとおり，matches は「リンク先のページが特定の内容にマッチする」という制約を表す．これは次のように，関数と組合わせて利用することもできる．

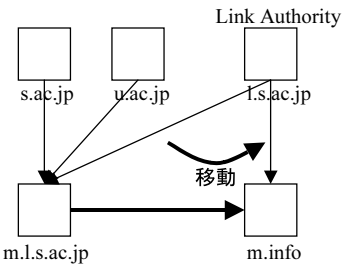

図 3 Link Authority

```
<A wi:matches=snapshot()
 href="http://www.mlab.info/">
    素敵な日々</>
```

ここで，snapshot() は最初にこの制約を WIM Server が認識した時点での，リンク先のページの内容を文字列としてそのまま返す関数である．したがって，この例のように制約が指定されているとき，システムはリンク先の内容に少しでも変更が加えられると，リンクが制約が満たされなくなったと判断する．

3.3 follows

WIDL/Link では，*Link Authority* という概念を導入する☆．ある Web ページ p は次の条件を満たすとき，別の Web ページ q に関する Link Authority であると言う．

(1) p が q へのリンクを持っており，かつ
(2) q が q' に移動すると，必ず p 中の q へのリンクは q' へのリンクに変更される．

例を図 3 に示す．まず，ある大学 A の研究室の Web ページ q_1 が存在し，このアドレスが m.l.s.ac.jp であるとする．q_1 は，s.ac.jp，l.s.ac.jp，u.ac.jp という複数のページからリンクされている．これらのうち，p_1(l.s.ac.jp) はその研究室が所属する学科の研究室一覧ページである．このとき，一般には l.s.ac.jp は q_1 に関する Link Authority である．したがって，例えば次のような状況が生じる．q_1 が q_2(アドレスは m.info) に移動したとする．すると，q_1 を参照しているページではリンク切れが存在するが，通常 l.s.ac.jp だけは q_1 へのリンクを q_2 に張り替えるはずである．

このとき，例えば u.ac.jp では次のように制約を記述する．

☆ Google などにおける Authority ページとは異なる概念である．

```
<A wi:follows="l.s.ac.jp"
 href="http://m.l.s.ac.jp/">
    某研究室</>
```

この例のように制約が指定された場合，follows による制約が満たされなくなるのは，指定された Link Authority において m.l.s.ac.jp のアドレスが変更されるか，もしくは Link Authority そのものが存在しなくなった場合である．

一般に，あるページ q に関する Link Authority は複数存在し，それぞれ異なる (暗黙の) 制約を表す．例えば，上の例における q_1 に関する Link Authority として，q_1 の研究室の教員 B が学生時代に所属していた研究室の OB リンク集ページ p_2 があるとする．そのとき，p_2 は (きちんとメンテナンスされていれば)「教員 B が運営する研究室」に関する Link Authority であり，p_1 は「大学 A の某研究室」に関する Link Authority である．このように，Link Authority を適切に選ぶことにより，多様な (暗黙の) 制約を記述することができる．

3.4 threshold

属性 threshold は自動的にリンクを更新するための閾値を指定する．閾値は 0 以上 1 以下で記述する．例を次に示す．

```
<A wi:isAlive wi:threshold="1"
 href="http://toyosu-it.ac.jp/">
    豊洲工大</>
```

指定された閾値を x とする．この場合，システムが発見した変更先リンクの確信度が x 以上の時のみ，WIM Server はリンクをその変更先リンクに自動的に書き換える．確信度が 1 になる場合の例としては，ページのリダイレクトが存在した場合がある．閾値以上の変更先候補が存在しない場合には，変更先のリンク候補のランキングページを作成し，そこへのリンクに変更する．

3.5 一般的な規則

WIDL/Link では以上の属性を組み合わせて制約を記述する．ただし，次の規則がある．

(1) デフォルトで isAlive の指定が存在
(2) デフォルトで threshold="1"の指定が存在

したがって，次の例と 3.4 節の例は同じである．

```
<A href="http://toyosu-it.ac.jp/">
    豊洲工大</>
```

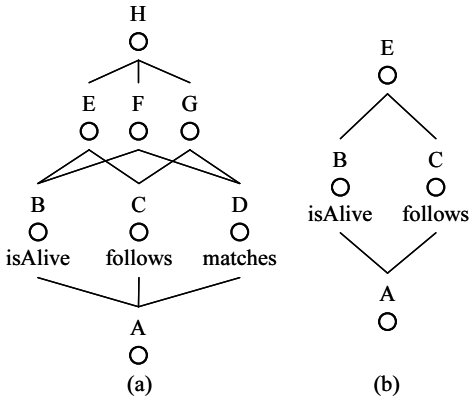

図 4 制約の組合せ間の関係 (a) とその部分グラフ (b)

このようなデフォルトの制約の存在によって，既存の WWW コンテンツであっても，WIM Server の機能を利用可能である．

4. 制約を満たすリンクの探索

4.1 最上位アルゴリズム

WIDL/Link が記述可能な制約の組合せ間の関係を図 4(a) に示す．図中の辺は，より厳しい制約である関係を表し，上位のものほど厳しい制約であることを表す．WIM Server はあるリンク l の制約が満たされなくなっていることを発見すると，次のような手順でリンクの探索を行う．

(1) l に関して指定されている制約の組合せが最上位ノードとなるよう，図 4(a) の部分グラフ G を抽出する．例えば，isAlive と follows が指定されている場合は図 4(b) となる．
(2) G の上位ノードから順に，その条件を満たす変更先リンクの探索を行う．各ページの確信度を，G のトポロジカルソートに矛盾しない順に割り当てる．特に確信度 1 の候補は，必ず G の最大元に対応する制約を満たすものに割り当てる．
(3) 変更先リンクの候補を確信度順に並べ，1 列のリンク列を作成する．

実際に制約を満たさなくなったリンクが与えられたとき，本アルゴリズムはその制約の組合せに応じてそれぞれ処理を行う．

4.2 各場合における処理

以下では図 4 における E,B,A に対応する制約の組合せが満たされなかった場合の処理の概要を説明する．重要なポイントは，どの場合においても，制約を満たす範囲で「以前指していたページの変更先を探す」という暗黙の仮定があることである．したがって，各処

理では制約を満たすページというだけでなく，この点を考慮して変更リンク候補の選択が行われる．

***E* の場合**　これは，`isAlive` および `follows` の指定が行われている場合である．この制約が満たされない時，満たされない条件によって 5 つの場合に分ける事ができる．以下で，$C1$ は「`isAlive` が満たされている」$C2$ は「`follows` の指定先が存在する」$C3$ は「指定された Link Authority とリンクが矛盾しない」ということとする．

(1)　$C1 \wedge \neg C2$: 処理 1
(2)　$C1 \wedge C2 \wedge \neg C3$:　処理 2
(3)　$\neg C1 \wedge C2 \wedge C3$: 処理 3
(4)　$\neg C1 \wedge \neg C2$: 処理 1
(5)　$\neg C1 \wedge C2 \wedge \neg C3$: 処理 2

処理 1　Link Authority が移動したと仮定し移動先の探索を行う．移動先の探索については 5 章で説明する．移動先の候補が見つかった場合，それを Link Authority とみなして制約の再評価を行う．見つからなかった場合には，`isAlive` の真偽に応じて，現在のリンクを候補とするかもしくは場合 B の処理に移る．

処理 2　指定された Link Authority で，旧リンクが新たなリンクに変更されていれば，それを変更先リンク候補とする．Link Authority は存在するが，旧リンクの代わりとなる新たなリンクが見つからない場合は処理 1 を行う．

処理 3　現在指している (制約を満たさない) リンクを変更先候補とすると同時に，場合 B の処理に移る．

***B* の場合**　リンク切れを起こすと制約を満たさなくなったと判断され，システムは現在アクセス可能なページであり，かつ変更先であると考えられるページを探索する．具体的には，次のヒューリスティクスに基づいて変更先のリンクを探索する．(1) 同じ Web ページは，時間が近いものほど内容が似ている傾向にある．(2)Web ページが移動するとき，同じサイト内で移動する可能性が高い．(3)Web ページの移動先がリダイレクトされている場合，移動先のページの URL がわかる．(4)Web 検索エンジンなどで逆引きすると，Link Authority を発見できる可能性がある．

具体的には次の処理を行う．まず，リンクが切れた場合に備えて，システムはリンク先のページを定期的にキャッシュする．リンクが切れた場合，次の処理を行い変更先リンクを探索する．(a) リダイレクト先が保存されている場合，そのページにリンクを書き換える．(b) リダイレクト先が保存されていない場合，次の二つの処理を行う．(b1) 検索エンジンなどを用いてリンクの逆引きを行う．結果のページの中から，変更先のリンクが存在しないか探索する．(b2) キャッシュしたページを用いて探索を行う．まず同じ Web サイト内のページを探索し，それでも見つからなければ，検索エンジン用いて探索する．

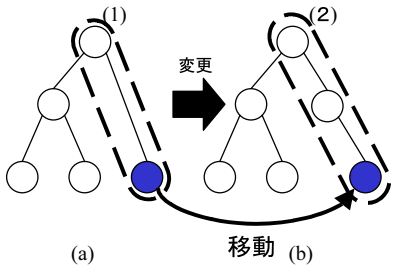

図 5　ある大学のサイトの構造．変更前 (a) と変更後 (b)

***A* の場合**　この場合，現在アクセス可能なページでなくても良いので，WIM Server がキャッシュしているページ，もしくは Internet Archive や各種検索エンジンなどに格納されているページを変更先リンク候補とする．

5. Link Authority の探索

前述したとおり，`follows` 属性で指定された Link Authority がアクセス不可能になった場合などに，WIM Server は Link Authority の探索を行う．本節ではある大学の Web サイトを例として探索手法の概要を説明する．

大学 s の学科 l のページ p があり，別のあるページ r のリンクが p を Link Authority と設定していたとする．大学 s のサイトの構造は図 5(a) のようになっている．ある時，大学 s の Web サイトの構成が大幅に変更されたために，p もこの変更に影響を受けた．図 5-(b) は変更後のサイトの構造を表す．ここで変更後に p のアドレスが変わってしまった場合，ページ r に記述されていたリンクの `follows` 属性で指定していた Link Authoriy が無くなってしまう．このような場合，システムは新たな Link Authority p' の探索を行う．

我々はシステムが p' を発見する方法として以下の方法を用いることとした．

(1)　システムは p とサイト (ドメイン) のルートから順番に，p' となりうるものを探索する．
(2)　探索を行う際に，変更前のページ p と 探索対

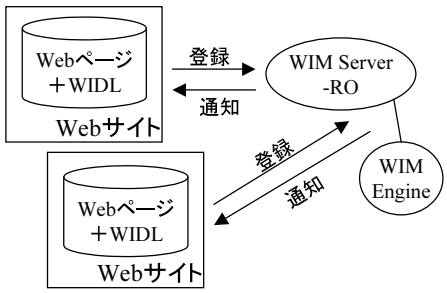

図 6 WIM Server-RO

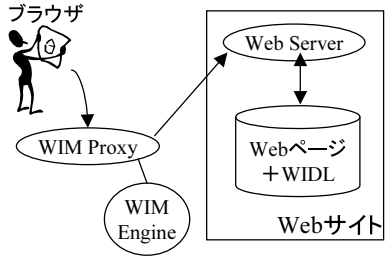

図 7 WIM Proxy

象のページについての類似度を測り，この類似度の近いものを p' の候補とする．

類似度は，TF/IDF 法でキーワードの重み付けを行い，コサイン距離を用いて計算することで求める．ただし，p と p' の類似度を直接比較するのではなく，図 5 における (1) の内容と，(2) の内容を全てまとめた内容を比較し計算する．また，以前の Link Authority が指していたリンクには高い重み付けを行う．最後に，発見された p' の候補を類似度に応じてランキングする．

6. 今後の展開

我々は，今後の展開として次の開発を検討している．

- WIM Server-RO
- WIM Proxy
- Link Authority Discovery Engine

6.1 WIM Server-RO

図 1 のように，Web サイト管理者は WIM Server をインストールすることにより，リンクの一貫性管理の支援を受けることができる．このインストールされた WIM Server は管理している Web ページの読み書きの権限が共に与えられるため，本章では WIM Server-RW (Read Write) と呼ぶ．しかし，管理者によっては，新たなソフトウェアのインストールを望まない場合や，そもそもホスティングサービスの利用などにより，WIM Server-RW を設置できない場合等も考えられる．我々は，このようなユーザの為に WIM Server-RO(Read Only) を用意する (図 6)．

WIM Server-RO は，管理対象の Web サイトとは別の計算機環境に設置される．利用するためにはサイトの管理者が WIM Server-RO に管理対象のサーバを登録する必要がある．WIM Server-RO は WIM Server-RW と同じアルゴリズムを実装するエンジンを持つが，コンテンツの自動的な書き換えを行わないという点が異なる．その代わり，WIM Server-RO は定期的に登録されたサイトを調べ，WIDL/Link による制約が満たされていないリンクを発見すると，WIM Server-RW と同じ処理を行い，候補ページの一覧をメールなどの手段でサイト管理者に通知する．サイト管理者はこの一覧を参照し，リンクを更新する．

6.2 WIM Proxy

提案手法を手軽に利用可能にするために開発されるのが WIM Proxy である (図 7)．WIM Server-RO/RW が Web サイト管理者のためのソフトウェアであるのに対し，WIM Proxy は WWW の一般利用者のためのソフトウェアである．これは基本的に HTTP Proxy として機能する．利用するためには，ブラウザなどで Proxy の設定を行うだけでよい．通常の HTTP Proxy と異なる点は，制約を満たさないリンクを発見したとき，WIM Engine を用いて変更先リンクの候補を提示することである．WIM Proxy が WIM Server-RW および RO と異なる点は，対象が不特定多数の Web サイトであるということである．したがって，WIM Engine のアルゴリズムの変更が多少必要になる．例えば，不特定多数のリンク先の情報をあらかじめキャッシュしておくことができないため，4.2 節 (B の場合) のアルゴリズムがそのままでは利用できない．したがって，WIM Proxy では Google によって用意されているキャッシュ機能や，Internet Archive に保存されているキャッシュ機能を利用することを検討している．不特定多数のページを対象とするため，必ずしも WIDL/Link による制約が記述されているとは限らないが，3.5 節で説明したように，デフォルトの制約を持つと解釈されるため，全てのページに対して少なくともある程度 (図 4 の B,A) のサポートは提供される．

6.3 Link Authority Discovery Engine

3 章で説明した Link Authority は単純かつ強力な概念であるが，属性 `follows` の指定を行うためには，あらかじめどのページが Link Authority であるかを知っている必要がある．Link Authoity Discovery Engine (LADE) は，多数の Web ページの中から，指定され

たページの Link Authority を発見するためのソフトウェアである．通常の Web ページ検索エンジンと同じように，他のソフトウェアや利用者からネットワークを通じて利用可能となるように設置される．これは，WIM Server や WIM Proxy と異なり，提案フレームワークを間接的に支援するものである．例えば，Web サイト管理者が WIDL/Link による制約を記述する際の参考にしたり，WIM Server が変更先のリンクを探すためなどに利用するといったことが可能である．また，将来的には `follows` 属性の自動作成などに利用することも想定している．

LADE はページを発見するという意味では通常の Web ページ検索エンジンに似ているが，ページ選択の基準が全く異なるため，内部的には全く異なるアルゴリズムを実装する必要がある．例えば，次のような手法が考えられる．(1) あるページへのリンクを持つページを長期的に観察し，リンク先ページの移動に従ってすぐに更新された場合，候補として高く評価する．(2) あるページへのリンクが `follows` 属性を持つことを発見すれば，指定されているページを候補として高く評価する．(3) リンクのメンテナンスがよく行われている (常にリンク切れが一定の割合以下の)Web サイトを発見し，そのサイトのページは全て候補として高く評価する．

7. おわりに

本稿では，一貫性制約の考え方を WWW の文脈に導入し，WWW コンテンツのリンク構造に関する一貫性管理を行うための手法を提案した．提案手法は既存の Web アーキテクチャの自然な拡張であり，また，単純かつ強力な制約記述言語を提供する．今後の課題としては，更新リンク候補のランキング手法のより詳細な検討，提案手法の実装と評価，それに基づくアルゴリズムの改良などがあげられる．

謝　辞

ゼミなどでご議論いただきました筑波大学図書館情報学系の田畑孝一教授と阪口哲男助教授に感謝いたします．

参考文献

1) S. Abiteboul, R. Hull, V. Vianu: Foundations of Databases. Addison-Wesley 1995.
2) All Things Web. State of the Web Survey. http://www.pantos.org/atw/35654.html
3) Georgia Institute of Technology GVU Center. GVU's 8th WWW User Survey. http://www.gvu.gatech.edu/user_surveys/survey-1997-10/.
4) Eitetsu Oomoto, Youichi Shima: Integrity Constraints for Reference Links in Hypermedia Database Systems. CODAS 1996: 182-185.
5) Science Education Broken Links: http://www-class.unl.edu/biochem/url/broken_links.html
6) Xenu's Link Sleuth (TM): http://home.snafu.de/tilman/xenulink.html
7) Link Check with LinkAlarm : http://linkalarm.com/
8) Planet SOSIG - A spring-clean for SOSIG: a systematic approach to collection management: http://www.ariadne.ac.uk/issue33/planet-sosig/