リンク切れ対応機能をもつ HTTP プロキシの開発

中溝 昌佳 † 有山 智洋 †† 森嶋 厚行 ††† 杉本 重雄 ††† 北川 博之 ††††

† 芝浦工業大学大学院 工学研究科 †† 図書館情報大学大学院 情報メディア研究科
††† 筑波大学 知的コミュニティ基盤研究センター・図書館情報学系 †††† 筑波大学 電子・情報工学系

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．これらの特徴により，WWW は社会に不可欠なメディアとなるまでに発展したが，一方でコンテンツの一貫性が必ずしも保証されていないという問題点がある．

我々は，リンク切れやリンク先の内容の変化への対応という，リンクの一貫性維持の問題に焦点を当てた研究を行っている．ある調査[3] によると，WWW の利用における重大な問題の一つとして，約 6 割のユーザが「リンク切れ」と答えており，リンクの一貫性維持は重要な問題と考えられる．我々は，これまでにリンクの一貫性を維持するための LIM(Link Integrity Maintenance) サーバの設計・開発を行ってきている[1)2)]．LIM サーバは Web サーバ＋ファイルシステムという通常の Web サイトアーキテクチャに追加することにより，その Web サイトのコンテンツに現れるリンクの一貫性を維持しようと言うものである (図 1)．具体的には，リンク切れなどを発見すると，変更先なるリンク先候補を自動的に求めるシステムである．LIM サーバの目的は，Web サイト管理者の支援を行うことであり，次の性質を持つ．(a)LIM サーバをインストールする必要がある．(b) 一貫性維持の対象は，インストールした先の特定の Web サイトである．(c) リンク切れになったリンクを自動的に更新可能である．

本稿では，このような Web サイト管理者のためのシステムではなく，Web の一般閲覧者のためのシステムを提案する．このシステムは一種の HTTP プロキシとして実装するため，LIM プロキシと呼ぶ (図 2)．管理者のための LIM サーバとは異なり，LIM プロキシは次の特徴を持つ．(a) 利用者はどこかにインストールされている LIM プロキシを利用するよう，ブラウザを設定すればよい．(b) プロキシを通じてアクセスする全ての Web ページ中のリンクが対象である．(c) リンク切れを発見すると，プロキシが代わりのリンク先を探してくれるが，コンテンツを勝手に更新することは無い．

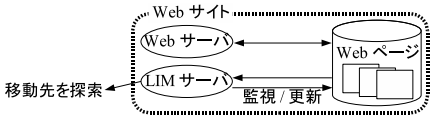

図 1 LIM サーバ

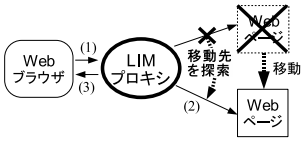

図 2 LIM プロキシ

LIM プロキシの動作は図 2 のようになる．(1) ページへのアクセスの指示がプロキシに伝わる．(2) そのページが存在しない (リンク切れ) ことを発見すると，システムが替わりとなる新しいリンク先の候補を探索する．(3) リンク切れ表示の代わりに，新しいリンク先の候補の一覧をブラウザに表示する．なお，リンク切れで無い場合は，通常の HTTP プロキシとして動作する．

関連研究としては次のようなものがある．まず，リンクの一貫性維持に関連したシステムとしては，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．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，参照先がリンク切れ等でないかチェックし，レポートを作成するものである．また，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も存在する[4)]．これらは全て Web サイトの管理支援が目的となっており，LIM プロキシのような Web の一般閲覧者を支援することを目的としたシステムとは異なる．

2. LIM サーバにおけるリンク探索アルゴリズム

LIM プロキシは，LIM サーバの技術をベースに設計する．したがって，ここではまず LIM サーバがリンク切れを発見したとき，代わりのリンクを探索するためのアルゴリズム[2)] の概要を説明する．以下では，対象となるリンク (URL) を u とする．アルゴリズムは，大きく分けて次の 3 つの構成要素からなる．

(1) 定期的にリンク u をチェックし，リンク切れでなければ u が指すページのコンテンツを LIM サーバにキャッシュする．もし，u がリンク切れであることを発見すると，次の (2)(3) を実行する．

(2) u が指していたページの移動先の候補となる Web ページ (URL) のリスト U を作成する．

(3) 各候補ページ (の URL)$u_i \in U$ に，ある基準に基

Development of an HTTP Proxy with the Ability to Cope with Broken Links

Akiyoshi Nakamizo† Tomohiro Ariyama†† Atsuyuki Morishima††† Shigeo Sugimoto††† Hiroyuki Kitagawa††††

†Graduate School of Eng., Shibaura Inst. of Tech.
††Grad. Sch. of Info. and Media Studies, Univ. of Lib. and Info. Sci.
†††RCKC, Univ. of Tsukuba.
††††Inst. of Info. Sci. and Elec., Univ. of Tsukuba.

づきスコアを割り当てランキングする．

本アルゴリズムは，候補ページの収集や各候補へのスコアの割り当てのために，次のヒューリスティクスを利用する．

H1 同じページは時間が近いと内容が似る傾向がある．本アルゴリズムでは，検索エンジンのキーワード検索により類似ページを収集し，LIM サーバが持つキャッシュと類似度を測り，類似度の高いページのスコアを高くする．

H2 ページが移動するとき，同一サイト内で移動する可能性が高い．本アルゴリズムでは，同一サイト内で LIM サーバのキャッシュと類似度の高いページを探索し，サイト内で発見したページのスコアを高くする．

H3 ページの移動先がリダイレクトされている場合，移動先ページの URL がわかる．リダイレクト先のページのスコアを高くする．

H4 検索エンジンなどで，リンク u の逆引き検索を行うと，その結果 (u へのリンクを持つページの集合) に含まれているページが，リンクの変更先を持つ可能性がある．これは，Web ページの更新と検索エンジンの更新に時間差があることに着目したヒューリスティクスである．

3. LIM プロキシの設計と実装

LIM プロキシと LIM サーバは利用形態が全く異なるため，上記の探索手法をそのまま利用するには，次のような問題がある．(問題点 1) LIM プロキシは不特定多数のページを対象とする．したがって，特定のページを監視対象とする LIM サーバと異なり，ページ全てをキャッシュする事は非現実的である．(問題点 2) LIM プロキシはインタラクティブなシステムのため，代わりのリンクの探索に長時間かけるわけにはいかない．我々の実験では，**H2** に基づくサイト内探索では，場合によっては数十分の探索を要する．(問題点 3) LIM プロキシは多人数に利用され，かつリンクの更新を行わないので，何度も同じリンクの探索を行う可能性がある．

以上の問題点に対処するため，LIM プロキシの開発においては次の工夫を行った．(工夫 1) 探索の手がかりとするキャッシュとして次のいずれかを利用する．(1)Web 検索エンジン (google など) のキャッシュ機能 (2)Internet Archive (3) もしリンク切れのサイトに LIM サーバ設置されていれば，その LIM サーバが保存しているキャッシュ．(工夫 2)**H2** に基づく同一サイト内の探索以外をまず行い，その時点での結果を利用者に提示する．もし結果に不満でありかつ待つ気があれば，利用者が指示を行うことにより同一サイト内の探索を行う．その際，タイムアウトを指定することも出来る．(工夫 3) 一度探索した結果を LIM プロキシにキャッシュしておき，他の利用者による探索時に再利用する．監視対象ページのコンテンツを保存する LIM サーバのキャッシュと異なり，保存するのはリンク切れが起こった場合の，代わりとなるリンクの候補だけである．

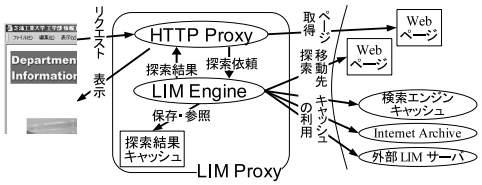

図 3 LIM プロキシのアーキテクチャ

LIM プロキシのアーキテクチャを図 3 に示す．主要な構成要素は HTTP Proxy 部と LIM Engine 部である．

HTTP Proxy 部: ブラウザからのリクエスト処理を行い，Web ページを取得する．アクセス時にページが存在すればページの内容をそのままブラウザに表示し，存在しない場合は LIM Engine 部に新しいリンク先ページの探索を依頼し，その探索結果をブラウザに表示する．

LIM Engine 部: アクセス先の Web ページが存在しなかった場合に，ページの移動先を探索し，ランキングを行う．探索の手がかりとするキャッシュを入手するために，検索エンジンキャッシュ，Internet Archive, 外部 LIM サーバと通信を行う．一度探索を行った結果は探索結果キャッシュに格納し，再利用を行う．

本システムは Java 1.4.2 により実装した．また，検索エンジンとしては google を利用した．

4. おわりに

本稿では，リンク切れ対応機能を持つ HTTP プロキシである LIM プロキシの開発について述べた．また，代わりとなるリンクを探索する仕組みの概要と，LIM プロキシ固有の問題への対象方法について説明した．今後の課題としては，探索精度向上のための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芝浦工業大学工学部の古宮誠一教授と筑波大学図書館情報学系の永森光晴講師に感謝致します．本研究の一部は文部科学省科学研究費補助金若手研究 (B)(課題番号 15700108) による．

参 考 文 献

1) 中溝昌佳, 森嶋厚行, 有山智洋, 杉本重雄, 北川博之．WWW コンテンツ一貫性維持のためのリンク更新機構の提案．日本データベース学会 Letters, Vol. 2, No. 2, pp. 65-68, 2003 年 10 月
2) 中溝昌佳, 有山智洋, 森嶋厚行, 杉本重雄, 北川博之．WWW におけるリンク一貫性維持支援システムの開発 (投稿中)
3) Georgia Institute of Technology GVU Center. GVU's 8th WWW User Survey. http://www.gvu.gatech.edu/user_surveys/survey-1997-10/.
4) Planet SOSIG - A spring-clean for SOSIG: a systematic approach to collection management : http://www.ariadne.ac.uk/issue33/planet-sosig/