WWW におけるダングリングリンクの自動メンテナンス

3P−4

石田 和生　谷川 哲司　宮下 敏昭
NEC ヒューマンメディア研究所

1 はじめに

WWW 上の HTML 文書データは、ネットワークを介した別のホストに存在するページに対してのリンク付けも非常に簡単に行うことが出来るが、別のホストに存在するページはほとんどの場合、リンク元のページとは管理が独立しているため、その存在がいつでも保証されているとは限らない。実際、リンク集のページ等に登録されているリンクが時間の経過とともにたどれなくなっていくことも少なくなく (このようにたどれなくなったリンクのことを本報告ではダングリングリンクと呼ぶ)、リンクのメンテナンスが必要とされている [1]。

ダングリングリンクが発生した場合には、リンク元ドキュメントの管理者がリンクを手動で再設定する必要があるが、一般的にこの作業は容易ではない。そこで我々は、そのようなリンクのメンテナンスを自動的に行う方式の提案とシステムの試作を行った [2]。本発表では、今回新たに追加したリンク先探索のヒューリスティックスを説明し、実際にメンテナンスシステムを用いて行ったリンクメンテナンス結果について報告する。

2 リンクメンテナンスシステム

ダングリングリンクの発生原因は「リンク先の文書が移動された」と「リンク先の文書が削除、あるいは外から不可視状態にされた」の 2 つに分類できる。文書が削除や不可視状態にされた場合には、同一文書へのリンクを復元することは不可能であるので、本システムではリンク先文書が移動した場合のみを対象とする。

リンクの修復は、リンク先候補文書の探索、リンク先文書の決定、文書中のリンクの修正の 3 段階に分けることが出来る [2]。リンク先候補文書の探索は、ダングリングリンクの URL をヒューリスティックス (詳細は後述) を用いて変形することで行う。また、リンク先文書の決定は、文書中に含まれる画像ファイルの URL やメールアドレスなどの情報 (以下、キー情報と呼ぶ) が一致しているかどうかで行う。このため、文書が多少変更されていても対応が可能であり、かつ、全文を検査するのに比べ高速に判定することが出来るという特徴を持っている。リンクの修正は、ダングリングリンクのリンク先を得られたリンク先文書の候補に置き換えるものであるが、システムが文書を書き換えることに抵抗を感じるユーザも少なくないので、現在は文書の管理者に通知するのみとしている。

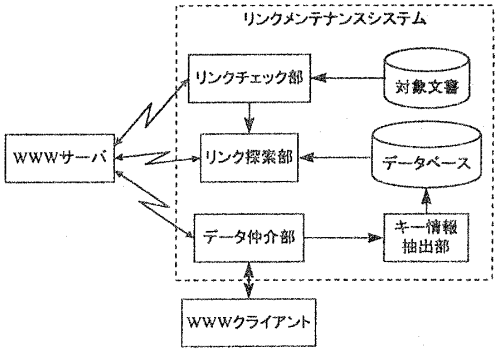

図 1: 全体構成図

システムの全体構成を図 1 に示す。システムは、キー情報抽出部 (データ仲介部)、リンクチェック部、リンク探索部からなっている。キー情報抽出部 (およびデータ仲介部) は、Proxy サーバのように、WWW サーバとクライアントの間で動作し、アクセス要求のあったページの URL とそのページのキー情報などをデータベースに蓄える。

リンクチェック部には、予めチェックするリンクの URL を登録しておき、比較的マシンの負荷が少ない時間帯などにリンク先が存在しているかどうかをチェックさせる。

リンク探索部は、リンクチェック部がダングリングリンクを検出した時に起動され、後述する方法でリンク先候補の探索を行う。探索結果は、文書の管理者に電子メールなどで報告する。

3 URL 変形のヒューリスティックス

リンク先探索の際に行われる URL の変形はいくつかのヒューリスティックスを用いて行う。ここでは本報告で新たに追加されたヒューリスティックスについて説明する。

3.1 リダイレクション

WWW 文書を移動させる際、移動したことを閲覧者に知らせるため、移動通知の文書を新しい URL とともに置いておくことや、HTTP のリダイレクション機能 [3] などを用いて新しい URL へ自動的にジャンプする設定を行うことがある。このうち、リダイレクションなどによる自動ジャンプでは、ジャンプ先の取得が容易に行えるのでリンク先探索を特に行う必要はない。一方、移動通知の文書の場合には、その文書が移動通知文書であ

An Automatic Maintenance for Dangling Link on WWW
Kazuo ISHIDA, Tetsuji TANIGAWA, Toshiaki Miyashita
Human Media Research Laboratories, NEC Corp.
8916-47 Takayama, Ikoma, Nara 630-0101, JAPAN

るという判断と、文書中から新しいURLの抽出を行う作業が必要となる。このため本研究では、まず、20ページ分の移動通知文書を収集し、文書中のキーワードと文書のサイズについて調査を行った。その結果、90 %の文書に、「移動」「知らせ」「変更」「引越し」「ジャンプ」といったキーワードが2種類以上含まれていることが判明した。また、文書サイズは、ひとつの文書を除き4 KBytes以下であった。そこで本システムでは、1) 文書サイズが4 KBytes以下で、かつ2) 移動を示すキーワードを2種類以上含むページを移動通知文書として判断することにした。また、新しいURLの抽出は、移動を示すキーワードの近辺に存在するURLを探索するという方法が考えられるが、文書サイズが比較的小さいことから、単純に文書中のURLをひとつ抜きだす簡易抽出方法を採用した。

3.2 URLのディレクトリ遡り

経験的に、`http://host/B/C/test.html`というURLで示される文書は、`http://host/B/C/`あるいは`http://host/B/`といった場所に存在する文書からリンクがはられていることが多い。そこで、ダングリングリンクが発生した場合に、そのURLの最後の/以下を取り除いたURLを生成し、そのURLで示される文書中に含まれるハイパーリンクをたどる。たどった結果得られた文書が消失した文書と同一であれば探索は終了。同一文書が見つからない場合には、新たに得られた文書中に存在するURLを同様にたどり、文書の探索を行う。ただし、再帰的に全てのリンクをたどると探索空間が膨大に広がってしまうため、a) リンクを再帰的にたどる回数の制限 (本システムでは、もとの文書が存在したディレクトリの深さ+3としている)、b) 違うドメインのURLは無視、を行うことで探索空間の絞り込みを行っている。これらの制限は、a) 文書を移動させる場合、`http://host/A/test.html`を`http://host/A/old/test.html`のように階層的に近い場所にする、b) ホストが変更されることがあっても別ドメインに移ることは少ない、といったことが多いと考えられることからきている。

4 メンテナンスの実行例

前章までで述べたメンテナンス方式を組み込んだリンクメンテナンスシステムを試作し、実際にメンテナンスを実行した結果を表1に示す。なお、表1中の「ホスト置換」というのは、文書のURLのホスト名の一部分を書き換える変形パターンである ([2])。また、メンテナンス対象となるURLには、著者らが保持していたブックマークと、検索エンジンにより検索したコンピュータとモバイル関係のリンク集に含まれるURLから、既にたどれなくなっているリンク50個をピックアップして用いている。このため、実験を開始した時点ですでにリンク先文書の取得が不可能で、キー情報の抽出を行うことが出来ない。そこで、今回の実験では、メンテナンスの結果得られた文書がもとの文書と同一であるかどうかのチェックはタイトルなどから主観的に行った。

表1: メンテナンス結果

メンテナンス結果	URL変形パターン	数
成功	ホスト置換	11
	リダイレクション	11
	ディレクトリ遡り	2
	ホスト置換+リダイレクション	2
	ホスト置換+ディレクトリ遡り	1
失敗	—	23
合計		50

表1の結果を見ると、約半数のリンクについてはメンテナンスが成功していることが分かる。成功した例の大半は、ホスト置換やリダイレクションといった比較的シンプルな手法によるものであるが、ディレクトリの遡りや複数パターンの組合せなど、手作業で行うと非常に労力のかかるメンテナンスも全ダングリングリンク中約20 %含まれており、これらのリンクが自動的にメンテナンス出来たことの効果は大きいと考える。

残りのメンテナンスに失敗した文書のURLは、そのほとんどがプロバイダーや学校などで用意された個人ページであり、ユーザアカウントごと削除された可能性が高い。この場合には、メンテナンス作業が不可能であると考えられる。

5 おわりに

本報告では、WWW上のHTML文書で発生するダングリングリンクの自動修復について、その手法と試作したシステムの実験結果について述べた。本システムを用いることで、従来人手で行っていたHTML文書中のリンクのメンテナンス作業を自動的に行うことが出来るようになり、メンテナンスに要する労力の低減が可能となる。

本研究は日本情報処理開発協会 (JIPDEC) による次世代電子図書館システム研究開発事業の一環として、次世代電子図書館システム実現のための個別技術とその実装技術の開発のために行っている。

参考文献

[1] M. S. Ackerman, R. T. Fielding, Collection Maintenance in the Digital Library, http://csdl.tamu.edu/DL95/papers/ackerman/ackerman.html, 1995.

[2] 石田, 谷, 市山, WWWにおけるダングリングリンクのメンテナンス方式, 第55回情処全大, 3-343, 1997.

[3] Hypertext Transfer Protocol — HTTP/1.1, http://www.rfc-editor.org/rfc/rfc2616.txt, 1999.