WWW におけるダングリングリンクのメンテナンス方式

6X－4　石田 和生　谷 幹也　市山 俊治
NEC ヒューマンメディア研究所

1 はじめに

WWW 上の文書データは、ネットワークを介して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たどれなくなっていることも少なくなく (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]。

リンク先が存在しなくなった場合には、その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、消失したリンクのリンク先を探索するのは一般的に容易ではない。また、リンク先の消失は通常、リンクをたどろうとした瞬間に判明するため、その時点からリンク先の探索を行っていたのでは欲しいデータを取得するまでのタイムラグが大きくなるという問題もある。

そこで本研究では、WWW ページ中に含まれるリンクを定期的にチェックし、リンク先が存在しないことが判明した時には自動的にリンク先の探索を行うリンクメンテナンス方式の提案を行い、その方式に基づいて試作したシステムについて解説する。

2 ダングリングリンク

WWW の文書は一般的には HTML で記述されることが多い。HTML では他の文書や画像データなどへのリンクをアンカータグを用いて `<A HREF="http://www.nec.co.jp/">` のように記述する。リンク先は URL (Uniform Resource Locator) で表現されているため、ローカルに存在するデータだけでなく、他のサイトのマシン上に存在するデータに対しても手軽にリンクを設定することが出来る。しかし、他のサイトのマシン上に存在している文書の管理者は一般的に、リンクを記述している文書の管理者と異なるため、リンク先の文書の変更等が、リンクを記述した文書の管理者に必ずしも通達されるとは限らない。このためリンク自体は存在しているがそのリンク先は存在しないリンク、すなわちダングリングリンクが発生することも少なくない。

A Maintenance Method for Dangling Link on WWW
Kazuo ISHIDA, Mikiya TANI, Shunji ICHIYAMA
Human Media Research Laboratories, NEC Corp.
1-4-24 Shiromi, Chuo-Ku, Osaka 540, JAPAN

3 ダングリングリンクの修復方法

ダングリングリンクの発生には大きく分けて「リンク先の文書が移動した」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 つがある。文書が削除や不可視状態にされた場合には、当然、元通りのリンクを復元することは不可能であり、この時には、代替となる文書を新たに見つけ出してリンクを設定するなどの対処が考えられるが、これは一般的には非常に困難な作業となるため今回の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、リンクの修復の 3 段階に分けることが出来る。リンク先候補文書の探索では、探索範囲が非常に広いため範囲を絞り込む手段が必要であり、また、リンク先文書の決定では、元の文書との同一性の判断がポイントとなる。本方式では、ユーザが普段アクセスしている HTML 文書の履歴を利用して絞り込みを行い、また、文書中に含まれる画像ファイルの名前や URL などの情報を利用して同一性を検証することによる、探索と同一性検証のコストの低減を特徴としている。以下、各部の処理手順を解説する。

3.1 リンク先文書の探索

消失したリンク先の文書データを探索する際、ネットワークにつながっている全てのマシン上のデータを調べることは非現実的である。そこで、本システムではユーザが普段アクセスしている HTML 文書を利用して探索を行うことにした。これは、消失したリンク先文書に関連のあるデータを同じユーザ、あるいは、同じサイトのユーザがアクセスする可能性が高いと推測されるからである。

具体的には、ユーザがアクセスした HTML 文書の URL などを記憶しておき、ダングリングリンクのリンク先の URL と記憶しておいた URL とのあいだでパターンマッチ (詳しくは後述) を行い、消失した文書の存在している可能性のある URL を推測する。

3.2 リンク先文書の決定

URL が推測されれば、実際にその文書を取得して元の文書と同一であるかどうかの判断を行う。同一文書が見つかればリンクの修復作業に入り、見つからなければ修復は失敗となる。

同一文書かどうかのチェックは、予めもとの文書の全文を保存しておき完全一致でチェックする方法が考えられるが、この方法ではチェックするリンク先の文書を全て保存しておかなければならないため膨大な記憶容量が必要になる、日付の変更

などの本質的でない変更への対処を別途必要とする、という問題点があるため、今回はHTML 文書中に存在するキー情報を予め保存しておき、同じキー情報が含まれているかどうかで判断することとした。ここでキー情報とは、HTML 文書中に存在し、かつ、大幅に変更される可能性の低い情報のことで、例えば以下のようなものがあげられる。

1. 文書の URL
2. 文書中で設定されているリンク
3. 文書中に含まれている画像ファイルの名前
4. 文書から切り出したキーワード
5. 文書中に含まれている E-mail アドレス

キー情報に基づいた比較方法だと異なる文書でも同一文書としてしまうことがあり得るが、URL が似通った文書どうしでこのようなことが起こる確率は低いと考えられるので、実用上問題はないと考える。

3.3 リンクの修復

ダングリングリンクのリンク先の探索に成功した場合、リンクの修復作業に入るのであるが、その方法としては、1) 探索結果のリンク先を文書の管理者に通知して管理者に修正を任せる、2) ダングリングリンクのリンク先を探索結果に置き換える、などの方法があり、いずれの方法も実装自体は容易である。しかし、システムが勝手に文書を書き換えてしまうことに抵抗を感じるユーザも少なくないと思われるため、今回のシステムでは管理者に通知するのみとしている。最終的には両方の手段を用意しておき、その判断をユーザに任せるようにする予定である。

3.4 URL のパターンマッチ

リンク先探索の際にポイントとなる URL のパターンマッチは、リンク先文書の移動パターンによっていくつかのヒューリスティックな方法を用いている。

文書の移動パターンの代表的なものとしては、マシンのホスト名の変更、ディレクトリ構成の変更、ファイル名の変更、の 3 つがある。ホスト名の変更の場合、対象文書 (例として、ホスト名 `host1.nec.co.jp` の文書 `doc.html`) の URL は、

```
http://host1.nec.co.jp/doc.html
                ↓
http://host2.nec.co.jp/doc.html
```

のように変更される。この場合には、記憶しておいた URL のホスト名部分で`*.nec.co.jp` にマッチするようなホストを探索して、マッチするホストが存在した場合には (例えば`host3.nec.co.jp`)、そのホストに `doc.html` が存在しているかどうかをチェックする。これを `doc.html` が見つかるか、あるいは、マッチするホストが存在しなくなるまで繰り返す。

ディレクトリ構成の変更やファイル名の変更の場合も同様に、ダングリングリンクのリンク先 URL の一部をワイルドカードに置き換えて、記憶している URL のパス部分とマッチングをとることで探索を行う。

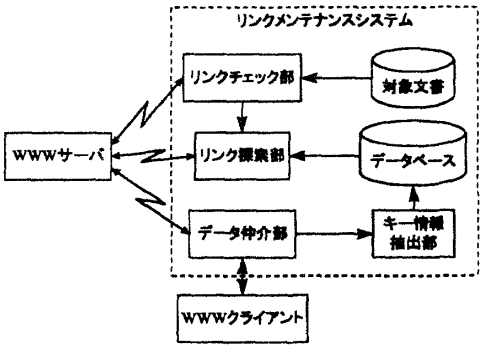

図 1: 全体構成図

4 システムの構成

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 (図 1)。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 などの情報をデータベースに蓄える。このようにシステムを Proxy 上に構築したのは、複数のユーザ間でデータベースの共有化を行うためである。

リンクチェック部は比較的マシンの負荷が少ない時間帯などに、予め登録しておいたメンテナンス対象の文書中にあるリンクのリンク先が存在しているかどうかをチェックする。

リンク探索部は、リンクチェック部がチェックしたリンクが存在しなかった時に起動され、前章で説明したような方法でリンク先の探索を行い、その結果を文書の管理者に電子メールなどで報告する部分である。

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を説明した。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、ユーザの情報活用がスムーズに行えるようになる。しかし、本システムで対応できるリンク修復は、まだ一部の限られた場合のみである。特に、リンク先の文書が削除などによって完全に消失した場合には全く対処できない。今後は、本システムをベースに、消失したリンクの探索方法などについてさらに詳しく検討を行うとともに、削除されてしまったページの代替ページの探索手法についても検討を行う予定である。

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.