Web 空間解析のためのリンクデータベースの設計と実装

松 永 吉 広† 廣川 佐千男††

Design and implementation of the LinkDatabase for the Web space analysis

Yoshihiro Matsunaga† and Sachio Hirokawa††

1. はじめに

爆発的な広がりを見せている WWW の基本表現形式は文書ファイルのある部分と別の文書ファイルを関連付けるといったいわゆるハイパーリンクの機能を持ったハイパーテキストである．したがって，このようなハイパーリンクの情報，つまりリンク情報が WWW の構造を決定しているといえる．

4) で述べられているようにこのリンク情報を用いた Web グラフや Web の空間分析などに関する様々な研究が行われている．また，検索エンジン Google は検索結果の順位付けにこのリンク情報を用いる手法 5) を導入することで高い精度の検索結果を得ることに成功し，高い人気を集めている．

このような Web 空間の解析にはある文書が他の文書を参照する順リンクの情報とある文書が他の文書から参照される逆リンクの情報が必要となる．

これらのリンク情報を Web から集めた大量のデータ中から提供するシステムとしては，Baharat 等の Connectivity Server2) がある．また，リンク情報を用いて Web サイトの評価を行うといった WebRatingService3) のように，Web サイト内を対象とした比較的小規模のリンク情報でも十分有用なデータとなりうる．

リンク情報を用いた Web 空間解析には，このようなリンク情報を効率良く利用するためのシステムを自前で持たなければならない．そうでなければ，一回ごとの実験に時間がかかるだけでなく，実験の度にネットワーク資源を浪費することになる．

我々の研究室ではこれまで BerkeleyDB を用いたリンクデータベースを実装し，Web ページの収集リンク情報抽出，検索エンジンによる逆リンク情報取得と保存等を行い，Web 空間の分析を行ってきた．本研究ではデータベースとして PostgreSQL を利用し，頑健で安定的に実験が行えるシステムを設計，実装した．本稿ではその概略と，それを用いた学内 Web ページ群の解析結果について述べる．

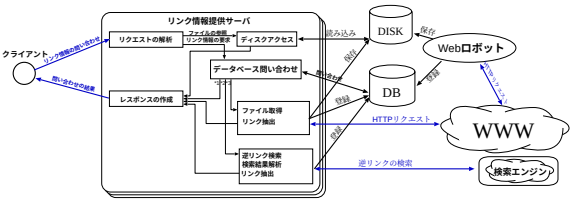

図 1 システムの概略図

「データベース問い合わせ」から下への処理はそれぞれ
登録済みのリンク情報の要求の処理,
未登録の逆リンクの要求の処理,
未登録の順リンクの要求の処理である．

2. リンク情報データベース

図 1 に本研究で作成したリンク情報データベースシステムの概略を示す．

3. データベース

本研究ではリンク情報を格納するためにフリーソフトウェアである PostgreSQL6) を使用した．関係データベースシステムを利用するにあたって，リンク情報を登録するためのスキーマを考える必要がある．ある文書が他の文書を参照しているリンク情報をそのまま関係と考えるとスキーマは，リンク情報（ある文書を示す URL，ある文書が参照している文書を示す URL）になる．また，関係データベースとしてはふさわしくないのだが，順リンクのリンク情報 (ある文書 A を示す URL，文書 A が参照している文書を示す URL のリスト)，逆リンクのリンク情報 (ある文書 A を示す URL，文書 A を参照している文書を示す URL のリスト) のスキーマも考えることができる．そこで，この２つのスキーマの比較を行うために人工的なリンク情報を作成し，検索にかかる応答時間を測定した．その結果，２００のリンクを持つノードに対してリンクを問い合わせたときにかかる応答時間の差が 1.1 倍程度であり，Barabashi 等 1) がいう

† 九州大学システム情報科学府
Graduate School of Information Science and Electrical Engineering, Kyushu University
†† 九州大学情報基盤センター
Computing and Communications Center, Kyushu University

リンク情報（リンク元 URL，リンク先 URL，アンカー文字列，リンク先の拡張子）
ファイル情報（取得したファイルの URL，ファイルサイズ，取得日時，リンクの個数）

図 2 リンクデータベースのスキーマ

ようにリンク数はべき分布に従い，非常に多くの順リンクまたは，逆リンクをもつような Web ページは数少ないこと，関係データベースとしての利便性等を考え，前者のスキーマを採用することにした．

最終的には，リンク情報を再実験，分析等で使われる可能性がある他の情報とともにデータベースに登録することとし，スキーマを図 2 のように定めた．下線部は主キーを表す．また，取得した Web ページのデータは，別にディスク上に圧縮して保存している．

3.1 HTML ファイルからのリンク情報抽出

HTML で，ある文書から他の文書へのリンクを表現する場合，一般にアンカータグを用いる．しかし，本研究ではこの他にもフレームタグ，クリッカブルマップ，メタタグでのリフレッシュについてもリンクであると考え，リンクデータベースに登録することにした．特にフレームは Web サイトのトップページに使われることが多く，その影響は大きいと思われる．実際，九州大学内の約 8 万個の HTML ファイル中，フレームタグ，クリッカブルマップ，メタタグによるジャンプがそれぞれ 2234 個，853 個，215 個あることを確認した．

3.2 リンク情報提供サーバ

リンク情報提供サーバは，クライアントのリクエストによって指定された URL の順リンク，または逆リンクを返すデータベースのインターフェースとして使用されることを想定し，設計，実装を行った．サーバは要求されたリンク情報がデータベースに登録されていない場合，自動的に Web ページを取得し，順リンクを得る，または検索エンジンを利用し，逆リンクを得る．そして，データベースに登録，クライアントに結果を返す．

リンク情報の要求のみであれば，このサーバを使用することで，クライアントは，要求するリンク情報がデータベースに登録されているかどうかに関わらず，リンク情報のリクエストを送信することで，リンク情報を得ることができる．ここで，図 2 のデータベース内の他の情報については関係データベースが提供する SQL 等で直接検索を行うことで得ることができる．

逆リンクに関して要求されたリンク情報が未登録の場合は検索エンジンを利用する．そこで，逆リンクの検索サービスを行っている代表的な検索エンジンである AltaVista と Google について逆リンクで得られた Web ページが実際に存在する割合を調べた．具体的には 10 個の URL それぞれについて逆リンク検索で得られる上位 50 件について実際に存在するかどうかを調べた．その結果，AltaVista では 35%，Google では 83%となったので Google を利用することにした．

4. システムの使用例

動作テストを兼ねて九州大学のトップページ (http://www.kyushu-u.ac.jp/) からたどれる学内の Web ページを収集するロボットを作成し，2001 年 1 月 11 日から 2 月 4 日の間に 85336 個の HTML ファイルを収集した．図 3 は，x 軸にリンク数，y 軸にその数だけ順リンクを持つ HTML ファイルの個数をとり，対数プロットしたものである．1) で知られているように，リンク構造は単純なランダムなものではなく，べき分布に従っていることが確認できた．図 4 は各 HTML ファイル d について，それから出るリンク数 out(d) とそのページへの逆リンク数 in(d) を座標として対数軸でプロットしたものである．対角線上に密集したしたページ群や，四分円の円周上に密集したページ群を発見できた．このようなページのクラスタの解析は今後の課題である．

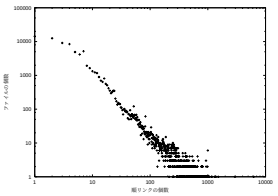

図 3 順リンクの個数とそのファイル数の関係

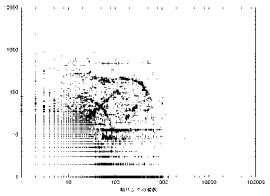

図 4 順リンクと逆リンクの個数による散布図

5. ま と め

本研究ではリンク情報を蓄えるためにデータベースシステムとして PostgreSQL を用いて，リンクデータベースの設計と実装を行った．本システムではアンカータグだけでなく，フレームタグや，メタタグなどについてもリンクとみなした．クライアントからのリクエストに応じてリンク情報を返すサーバを設計，実装し，九州大学内の Web ページをデータベースに登録し，システムの使用例としてリンク情報の分析を行った．このシステムの開発によってリンク情報解析の実験を安定的に行うことができるようになった．

参 考 文 献

1) Albert,R.,Jeong,H.,Barabasi,A., ”Diameter of the World Wide Web”, Nature,401,130-131,1999.
2) Bharat,K.,Broder,A.,Henzinger,M., Kummar,P., Venkatasubramanian,S., “The Connectivity Server: fast access to linkage informatio n on the Web”, 7th Intl World Wide Web Conference, 1998.
3) Hiraishi,H., Kato,H., Ohtsuka,N., Mizoguchi,F., “Web Site Rating and Improvement Based on Hyperlink Structure”, Discovery Science,429-434,2001.
4) 廣川佐千男, 池田大輔, “Web グラフの構造解析”, 人工知能学会誌 Vol.16,No.4, 525–529, 2001.
5) Page,L.,Brin,S.,Montwani,R.,Winogrand,T., “The PageRank Citation Ranking:Bringing Order to the Web”,
6) PostgreSQL,INC., “PostgreSQL”, http://www.postgresql.org/