WWWナビゲーション環境の試作 (2)
～ノード/リンク探索ツールの作成と評価～

3P－2

久保 信也　高野 元
NEC C&C研究所

1 はじめに

World-Wide Web(WWW) がインターネット上の情報発信システムとして普及し，公開されている情報を検索する手段としてのディレクトリ情報サービスが注目を集めている．WWWディレクトリ情報サービスとして，ドキュメントの書誌情報による検索手段や，キーワードによる検索手段を提供するためには，公開されているWWWのドキュメントを一旦獲得し，その中身を走査してタイトルなどの書誌情報とキーワードを抽出しなければならない．ドキュメントはインターネット上に分散して存在し，また日々追加・更新が行なわれているため，ドキュメントの獲得・ディレクトリ情報の作成を自動的に行なえるような手段が不可欠であり，いくつかの研究が行なわれている [1].

本稿では，WWWのノード/リンク探索ツールを作成する際に考慮しなければならない問題点とその解決方法の一例を提案している．

2 ノード/リンク探索ツールの概要

この節では，ノード/リンク探索ツールの動作概要を説明する．

WWWではHTML(Hypertext Markup Language)[2]のタグ"A"で，ハイパーリンクを記述する．またリンク先ノードは，URL(Uniform Resource Locators)[3]と呼ばれる表記方法で示す．

例　<A HREF="*URL*"> ホットワード </A>

図1にノードのHTMLドキュメントとハイパーリンクの例を示す．

ノード/リンク探索ツールは，以下に述べる手順で探索を行なう．

1. 探索開始ノードのドキュメントをHTTP(Hypertext Transfer Protocol)[4] を使って獲得する
2. ドキュメントを走査して，"A"タグで囲まれている部分を探す
3. "A"タグ中に記述されているリンク先ノード (URL) を抜き出す
4. 3で発見したノードのドキュメントをHTTPを使って獲得する
5. 2以降の繰返し

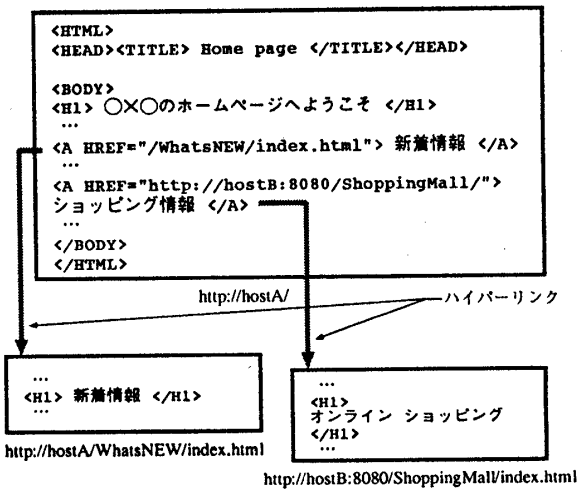

図 1: WWWでのノードとリンクの例

3 ノード/リンク探索ツールの問題点

ノード/リンク探索ツールを開発するにあたって，ネットワークのトラフィックやWWWサーバに与える負荷を不必要に増大させないために，以下に述べる点について考慮しなければならない．

1. 不要なドキュメントデータの転送

WWWではノードのドキュメントとして，HTML，プレーンテキストなどのテキストデータ以外にも様々なマルチメディアデータ，例えば静止画像データ，動画像データ，ムービーデータ，音声データを扱うことができる．これらテキスト以外のマルチメディアデータは，テキストと比較してデータサイズが大きいため，ネットワーク資源をより必要とする．本システムのように書誌情報，キーワード，ハイパーメディアの構造を抽出することを目的としている場合には，テキスト以外のデータを収集する必要がない．

2. 冗長なノード探索

ハイパーメディアのノード/リンクの構造は一般にネットワーク構造となるため，リンクを辿ってノードの探索を行なっているうちに，ループに陥ってしまうことがある．特にWWWでは，多数の情報提供者が独自にノードドキュメントを作成して他のノードにリンクを張るため，ループができることが多い．

3. 冗長なデータ転送

WWWのノードで提供されているドキュメントは，頻繁に更新されるものもあればほとんど更新されない

An Experimental Implementation of WWW Navigation Environment (2)　– An Implementation and an Evaluation of the Node/Link Traverse Program –

Nobuya Kubo and Hajime Takano
C&C Research Laboratories, NEC Corp.

ものもある．定期的にノード/リンクの探索を行なってドキュメントデータの獲得を行なうような運用形態の場合，前回獲得時から更新されていないデータを転送するのは無駄である．

4. 特定サーバへのアクセスの集中

あるノードから張られているハイパーリンクのリンク先ノードは，特定のWWWサーバ上に存在している場合が多い．そのため，同一のWWWサーバに対して連続してアクセスが発生する傾向がある．

5. 探索の終了方法

無限に探索を行なうことを許すと，インターネット全体に存在するノードの探索が行なわれる可能性がある．

4 実装方法

この節では，前節で述べた問題点に対応したノード/リンク探索ツールの実装方法について述べる．探索結果は，データベースに格納してディレクトリ情報として利用するとともに，ノード/リンク探索ツール自身も探索中に利用する [5].

1. ドキュメント データ タイプの判別

不要なデータ転送を行なわないために，ノード ドキュメントのデータ タイプを判別する．

初めて探索するノードの場合，最初に"HEAD"メソッドを用いたHTTPリクエストをWWWサーバに送り，メッセージ ヘッダのみのHTTPレスポンスを要求して，ヘッダ中に記述されているデータ タイプを利用して判別する．テキスト (HTMLを含む) 以外のドキュメントであれば，以後はドキュメントの転送要求はしない．

得られたデータタイプは，そのノードの属性情報としてデータベースに格納する．

2. ノードの探索時刻

WWWサーバからのレスポンス メッセージ ヘッダ中に記述されているレスポンスメッセージの送信時刻を"ノードの探索時刻"としてデータベースに格納する．

あるノードを探索する前に現在時刻とデータベース中の探索時刻とを比較し，指定した時刻が経過していなかった場合には探索経路にループがあったと判断し，そのノードは探索しない．

3. ノードの最終更新時刻

WWWサーバからのレスポンス メッセージ ヘッダ中に記述されているノードURLの最終更新時刻を，"ノードの最終更新時刻"としてデータベースに格納する．

ノード ドキュメントの転送を要求する時に，メッセージ ヘッダの"If-Modified-Since"フィールドに前回ノード ドキュメントを獲得した時に得た最終更新時刻を設定する．WWWサーバは指定された時刻以降にノード ドキュメントが更新された場合のみドキュメントを送り返す．

4. アクセスの分散

同一サーバに対するアクセスが連続しないように，探索開始ノード (複数あって良い) を頂点とする階層構造で，幅優先探索で行なう (図2).

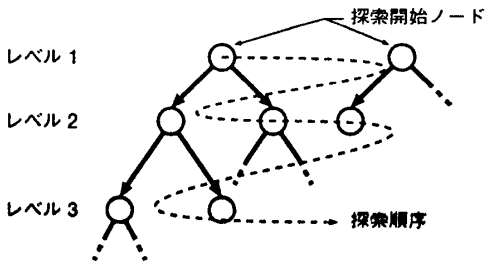

図 2: ノードの幅優先探索

5. 探索範囲の指定

探索の範囲をノード/リンク階層のレベルとノードが存在するサーバのロケーション (ホスト名，ドメイン名) で指定できるようにする．

幅優先探索中にレベルが指定した値に達した時に，探索動作は終了する．

5 評価

今回試作したノード/リンク探索ツールは，Perl(バージョン 5.001m) と dbm ライブラリ (GNU dbm 1.7.3) を使用して実装し，EWS4800/260 上で実行した．

社内5部門のWWWサーバを探索範囲として設定し，探索を行なった結果は以下の通りである．なお試作版では幅優先探索の替わりに，同一ホストへのアクセスの集中を防ぐために，連続して同一ホストにアクセスする場合には3秒間待つという実装によって負荷の集中を避けている．デバッグ用のログを解析した結果，所要時間の大部分は発見したURL中のホスト名の正式名をDNSで引くために要した時間であった．

WWWサーバの台数	11
発見したノード (URL) の個数	19,273
探索対象となったノードの個数	1,529
所要時間	3:36'42"

6 おわりに

Perl と DBM ライブラリを用いた試作ツールでの検証を元にして，現在C言語とRDBを用いて性能面を強化したノード/リンク探索ツールの開発を行なっている．

参考文献

[1] 田村 健人，村岡 洋一. WWWにおける広域検索システム. 情報処理学会 第51回全国大会, September 1995.

[2] T. Berners-Lee and D. Connolly. "Hypertext Markup Language - 2.0", November 1995. RFC1866.

[3] L. Masinter T. Berners-Lee and M. McCahill. "Uniform Resource Locators (URL)", December 1994. RFC1738.

[4] R. Fielding T. Berners-Lee and H. Frystyk. "Hypertext Transfer Protocol – HTTP/1.0", October 1995. Internet-Drafts.

[5] 高野 元，久保 信也，島村 栄. WWWナビゲーション環境の試作 (1) ～ディレクトリサービスの構成～. 情報処理学会 第52回全国大会, March 1996.