WEBサイトの階層構造の可視化による情報収集支援システム

吉村優一† 中山泰一†
†電気通信大学情報工学科

1 はじめに

インターネットに存在する情報を収集するには，計算機の習熟度に関わらず検索エンジンを用いるのが一般的である．このとき，普段あまり計算機を使用しない初心者は，検索エンジンから出力される検索結果のみを探索して，情報収集を止めてしまう傾向がある．

ほとんどのWEBサイトは，単一のページのみではなく，WEBサイト内のそれぞれのページの内容が関連しあい，図1のような“階層構造”の形式を取っている．普段から計算機を使用することの多い熟練者は，この階層構造を利用することで，検索エンジンから直接得られるページだけではなく，周辺のページも調べる手段を取る．そのため，初心者に比べて，より多くの情報を得ることができる．

そこで本研究では，WEBサイトの階層構造を取得し，可視化することによって，情報収集を支援するシステムを提案する．本システムを用いることによって，ユーザが情報を収集する際，検索エンジンから直接得られるページのみを探索する場合に比べ，より効率良く情報を収集できる．

2 関連研究

階層構造を利用した情報収集の効率化の手法として，WEBサイトの階層構造を利用したWEBページのグループ化がある[1]．検索の際に，検索結果をグループ化することによって，検索結果のリストの短縮を行っている．

また，階層構造に着目したWEBサイトの閲覧支援システムとして，ForeScope[2]がある．WEBサイトの閲覧の際に，それまでの閲覧の流れを表示することにより，WEBサイトの構成を直感的に認識できる．

WEBサイトの構造を表示するシステムとしては，

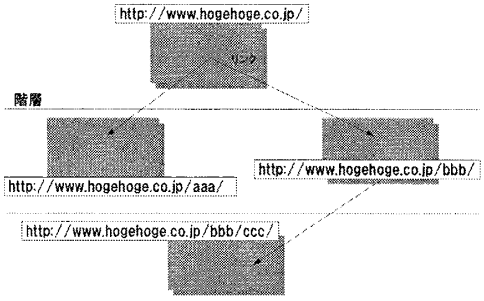

図1: WEBサイトの階層構造

Website Explorer[3]や，Firefoxの拡張機能であるScrapBook[4]がある．しかし，これらのシステムは，WEBサイトの構造を保存することを目的とするため，動作に時間がかかる．それに対して，本システムではリンクの取得のみを行うため，処理が速く，効率良く情報収集を行うことができる．

3 設計

本システムは大きく分けて2つのシステムから構成されている．内部リンクを探索する部分と，取得したリンクを描画する部分である．

3.1 リンク探索部分

内部リンクの取得は，以下の手順を再帰的に実行する．

1. WEBページのhtmlソースを取得
2. htmlソースから内部リンクを抽出
3. 抜き出した内部リンクすべてに対して1，2を実行

同じドメイン内のすべてのWEBページを取得するには処理に時間がかかりすぎてしまう．そこで，再帰する回数をユーザが指定することで，探索する深さを実行時に決定する．内部リンクの取得が完了した後，URLを階層構造ごとに分類する．情報収集の支援が目的なので，imgやjpgなどの画像ファイルへのリンク

Information Gathering Support System by the Visualization of Hierarchical Structure in Web sites
Yuichi YOSHIMURA†and Yasuichi NAKAYAMA†
†Department of Computer Science, The University of Electro Communications
182-8585, Chofu, Japan

Copyright © 2010 Information Processing Society of Japan.
All Rights Reserved.

は除いて取得する．それによって，WEB サイトの保存を目的としているシステムと比べて処理速度が向上し，情報収集の効率が良くなる．

3.2 描画部分

リンク探索部分で取得した URL を元に，階層構造の形をとるように WEB ページの描画を行う．その際，描画されたページをクリックすると，ウィンドウが対応したページに遷移できるように，画像中にリンクを埋め込んでおく．しかし，それだけでは描画されたページの情報が，そのページに遷移するまでわからない．そこで，ページ遷移をせずに描画されたページ情報がわかるように，以下の機能を追加する．

- 画像にマウスを重ねるとポップアップでタイトルと URL を表示
- 指定したページの内部リンク一覧を表示

4 実装と評価

4.1 実装

本実装では，オープンソースのブラウザである Firefox 3.5.7 の拡張機能を用いた．拡張機能の追加はドラッグ＆ドロップの簡単な作業で済むため，本システムの対象とする初心者でも行うことができる．

4.2 評価

4.2.1 実行時間の評価

実際に本システムを Google のトップページで試用した際の，インターフェース部分のスクリーンショットを図 2 に示す．また，3.1 節で述べた手順の再帰回数を変えて同ページで試用した結果，処理時間，取得リンク数は表 1 のようになった．

図 2: システムのスクリーンショット

表 1: Google トップページでの実行結果

再帰回数/回	1	2	3
処理時間/s	2.50	21.5	311
取得リンク数	88	661	6845

4.2.2 使用感の評価

評価として本システムの使用感を聞くため，計算機に触れる機会の少ない初心者に実際に使用してもらい，意見を聴取した．

本システムを被験者に使用してもらった際の感想として，「WEB サイトの構造というものを意識したことがなかったのでおもしろかった」などの好意的な意見を得られた．WEB サイトが階層構造の形式を取っているという認識を持たせることができたので，よい結果が得られたと言える．しかし，階層構造の表示方法の改良や，処理速度の向上を求める意見もあった．

5 おわりに

本研究では，現在表示されているページの内部リンクをたどり，その階層構造を可視化し表示するシステムを提案した．本システムを用いることで，ユーザが情報を収集する際，検索エンジンから直接得られるページのみを探索する場合に比べ，より効率良く情報を収集できる．

被験者の感想や意見から，今後の課題として，階層構造の表示部分を初心者でもわかりやすくすることや，階層構造の表示を高速化することが挙げられる．

参考文献

[1] 原田晶紀，佐藤進也，風間一洋：WWW ページ間の階層構造の推定と検索システムへの応用，情報処理学会研究報告，DBS118–14，pp.105–112 (1999).

[2] 深津真二，阿久津明人，外村佳伸：ForeScope：階層化コンテンツの閲覧支援システム，情報処理学会研究報告，2003–HI–102，pp.7–12 (2000).

[3] Website Explorer, http://www.umechando.com/webex/index.html

[4] ScrapBook, http://amb.vis.ne.jp/mozilla/scrapbook/index.php

Copyright © 2010 Information Processing Society of Japan.
All Rights Reserved.