5R－2

ハイパーテキスト構造解析と サーチエンジンへの応用

牧　大介

電気通信大学大学院情報システム学研究科

1.はじめに

インターネット World Wide Web で主に用いられるようになったハイパーテキストは、文書と文書をリンクとして結びつける事により、従来の文書では表現し得なかった構造を持つことが可能となった。しかし、この新しいメディアにおいては、従来の紙ベースでの文書で用いられた「起承転結」などの論理構造がリンクという新たな結びつけにより成立しなくなったとも言える。本研究では、ハイパーテキストの中でも HTML 文書を取り上げ、その論理構造を解析し、適切な文書情報構造(目次)を自動生成する。そしてこの目次を用い、全文検索するサーチエンジンを作成し、全テキストを対象とする従来型よりも軽量なサーチエンジンを提案する。

2.構造解析

本研究ではハイパーテキストとして HTML 文書を取り上げる。もっともよく目にし、作成が簡単で、既存の情報量が多い。しかし、「適当にリンクを張っただけ」のものが多く閲覧に支障をきたす事も頻繁である。構造解析では HTML 文書群の意味的な目次（章・節・項）を作成する。

まずリンクからの解析を行い、サイト全体の俯瞰図を作成する。具体的にはある HTML ファイルから別のファイルへのリンクを探し出し、多次元配列に On/Off で記録する。この俯瞰図を元に、インデックスとなっているファイルをファイル名・パスをヒントにして選び出し、このインデックスとなっているファイルを基点としてリンク構造をツリー状に展開していく。インデックスファイルごとにトピック（章）が存在するとして考えるので、いくつかのトピックに同じ HTML ファイルが出現する可能性はある。リファレンスとしてのリンク（この内容に関してはココを見ろ、など）は主に片方向のリンクとなるので取り除く。双方向のリンクを重視し、双方向のリンクが張られているという事は相関関係があり、どちらかがどちらかに含まれるトピック（節）であると考える。また、ほぼ全部のファイルに表などの形で同じリンクを提供している場合（各ページにタイトル・ What's New ・注文ページなどのリンクを用意している場合）は、そのリンクを Template.html というファイルに記録しそのフォルダ

Hyper-Text Link Analyzing, and Link Analyzing Search Engine

Daisuke Maki

University of Electro-Communications

1-5-1 Choufugaoka, Choufu, Tokyo, Japan

に置いておけば、解析の際に取り除くようにしている。最終的にツリービューには取り出した章とその下位の節が表示され、この段階で「リンクという観点から見た」目次が生成される。

次にタグからの解析を行う。まずファイルを一つ取りだし、先ほどのリンク解析結果からそのファイルへリンクを持つファイルも用意する。最初に取り出したファイルから平文（どのタグにも挟まれていない部分。普通の文章だと考える）を取り除き、いくつかのタイトル判断タグ（<LI><B><I><Hn>など）で挟まれた部分をタイトル（項目タイトル）だと考えて残す。この内容を元に先ほどのツリー構造を展開していく。リンク構造解析でのツリービューが「全体のリンク構造という観点から抽出した章・節」だとしたら、タグ解析ではさらに各ページを解析することでこれに「項」という概念が追加される。結果としてそのサイト全体の「章・節・項」が構成され目次が出来あがる。

3.サーチエンジンへの応用

本研究ではリンク構造解析の応用例として、自サイトサーチエンジン（あるサイト内に限定されたサーチエンジン）へリンク構造解析を適用してみる。つまり、リンク構造解析の「目次を自動生成する」という特質を用いて、情報提供型サイト(ある内容{JavaScript について、など}に関しての情報を提供するサイト)で、「キーワードに関しての説明を検索する事」に特化したサイト内サーチエンジンを提案する。つまり、ディレクトリ型のようにゴミの少ない適切な文書を検索し、全文検索型のように更新作業が自動化できるサーチエンジンである。

実装例としては、そのサーバー上に公開されている HTML ファイルをフォルダごとコピー、これに構造構造解析をかけて平文部分やいらないタグの部分を削除し、上位のページから情報（章・節のタイトルにあたる）を差し込む。こうして各 HTML ファイルはそのタイトルや目次のような情報のみを残した状態になる。これをフィルタリングしたのちインデクサにかけてインデクシングする。つまりフィルタ部のみにユニットを追加し、残りの部分は全てサーチエンジン本来のユニットを使用する事になる。

結果として平文部を切り捨て、目次の内容だけを残すため、インデックスが大幅に圧縮される（60のサンプルサイトに対し、平均の圧縮率は 34%)。これにより「本文中に出現しているだけ」のキーワードに関しては HIT しなくなり、検索結果にゴミが減る。

4. おわりに

情報検索関連技術の問題として「テキストに書かれている内容をおおまかに把握することのできる技術」というのが求められている。文書群の量が増加すればするほど、情報検索には精密さがより求められるようになる。このためには、各テキストの内容をシャープに把握する事ができる技術が必要不可欠である。このために既存のハイパーテキストから目次を自動生成するという方法は一つの提案になりうると考えられる。