WWW リンク情報によるコンテンツ空間の構成方式

4 L － 1

武藤 哲幸
mutou@slab.ntt.co.jp
NTT ソフトウェア研究所

1. はじめに

近年、Internet 上における爆発的な情報量の増大に対応するため、さまざまな検索方式や、ナビゲーション方式が提案されてきている。一般的にこれらの多くは、数多くの候補の中からユーザが必要とするものを取捨選択する過程を支援するための、情報フィルタとしてとらえることができる。

本研究では、これとは逆に現在ユーザが注目している情報の周辺にあるもの（＝場情報）を積極的にユーザに提示するような情報サービスの提供をめざしている。WWW における場情報とは、たとえば InfoGather[*1]などによる利用者の挙動やアクセス履歴、利用者プロファイリング情報などの動的な情報や、hyperlink による HTML 文書間の関係など、HTML で表現された情報以外のものを指す。

本稿では、hyperlink を用いた基本的な場情報の提供方式と、周辺情報の提示機能を持った新しい WWW ブラウザを提案する。

2. Hyperlink アナライザ

WWW において、もっとも基本的な場情報として考えられるものが、コンテンツ間の hyperlink である。一般的に、hyperlink の調査は Web ロボットなどによって行われるが、適用例としてはサイトマップの作成やリンクチェッカ、プリフェッチャ、検索エンジンのドキュメント収集部としての利用が主である。

本稿では、ある束縛条件を与えた Web ロボットを利用して複数を hyperlink 調査によってカテゴライズし、周辺情報として利用する手法を述べる。

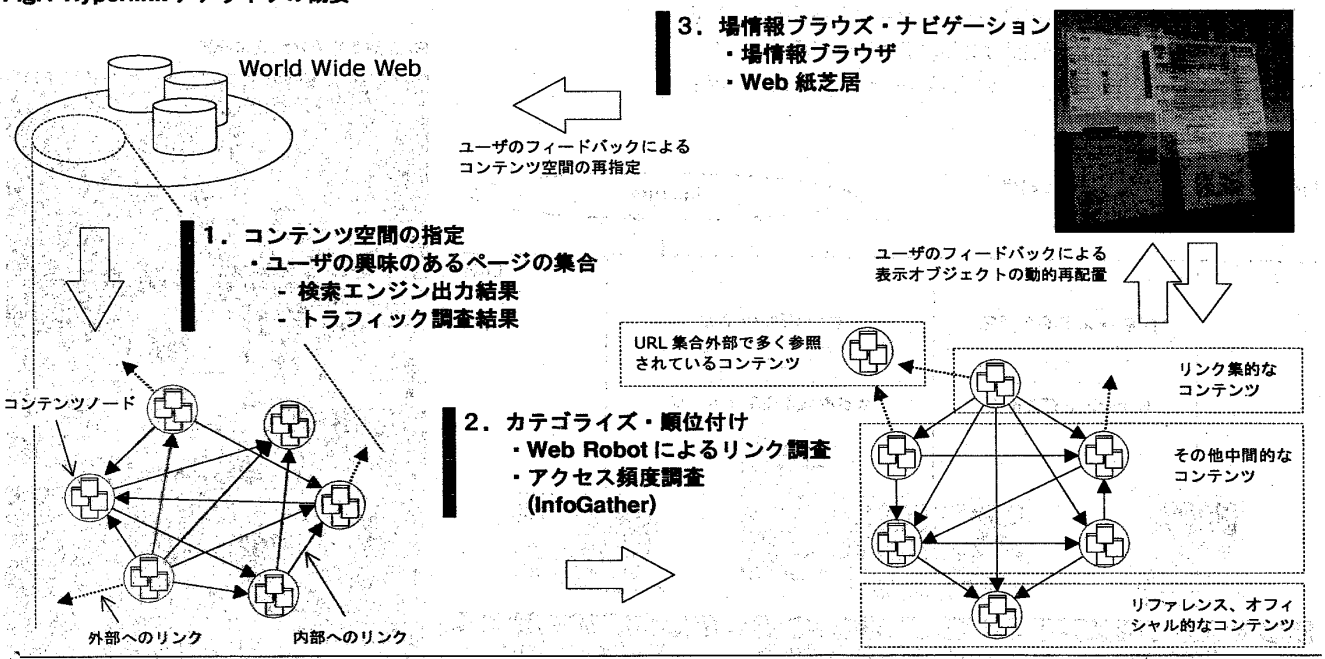
Web ロボットの探索で得られた結果は膨大な情報量にのぼるため、探索ノードの粒度をサイト単位、またはページ単体として扱うことが多い。本研究では、**『コンテンツのインデックスページ（トップページ）以下のディレクトリ階層に属するページ』を Web ロボットが探索結果を管理する粒度（＝コンテンツノード）として定義**する。今回試作した Web ロボットは、このコンテンツノードに含まれる hyperlink すべてを抽出し、他のコンテンツノードへの参照数をカウントする。Web ロボットがすべての探索を終了すると、結果として探索した URL 集合の総当たり表が選られる。これを統計的手法と、ヒューリスティックな経験則によって解析することによって、以下のようなカテゴライズが可能となる。

(1) リンク集的なコンテンツの抽出
外向きのリンク数が多く、広い範囲を参照しているコンテンツノード。

(2) リファレンス的なコンテンツの抽出
多くのコンテンツノードから参照されているコンテンツノード。原典や、オフィシャルページ的なものが多い。

(3) 検索結果以外の関連ページの抽出
多くのコンテンツノードから参照され、かつ探索範囲の外部に存在するもの。検索エンジンの結果には含ま

Fig.1 Hyperlink アナライザの概要



The contents based structuring method by WWW hyperlink.
Tetsuyuki Mutou, NTT Software Laboratories

れていないが、強い関連性を持つノードがわかる。

(4) 検索結果のクラスタリング
お互いにリンクを張り合っているコンテンツノード同士のクラスタリングを行う。たとえば、『酒』というキーワードに対する検索結果の URL 集合から、『ワイン』、『日本酒』、『蒸留酒』などの、分野別のカテゴライズが期待できる。

Web ロボットに与える探索コンテンツノードの集合には、NTT Directory や Yahoo! JAPAN の検索結果のように、共通の属性（ある検索 keyword に関する結果）を持ち、かつコンテンツのトップページへのリンクが得られるものが適している。

現在、Yahoo! JAPAN の検索結果を探索 URL 集合とした実験を進めており、特徴的なノードの自動抽出アルゴリズムと、Conceptual Clustering などの統計的手法を検討中である。

3. 場情報ブラウザ

現在の Web ブラウザでは、場情報的なものは別フレームやプラグインなどの手法で提供されていることが多い。

本研究では、これらの周辺情報をユーザに提示させるための GUI 機能を持たせた場情報ブラウザを開発中である。この場情報ブラウザでは、たとえば関連するページの情報は、WebPage オブジェクトを利用することによって、ページの表示イメージを任意の位置、サイズで複数個同時に表示することができる。また、透明度を指定することによって、ブラウザを操作しなくても、背後に周辺情報がどの程度あるかを認知させることが可能である。

この WebPage オブジェクトは、3 次元グラフィックス API を利用して実装されているため、他の 3D オブジェクトと混在させることが可能である。たとえば、アニメーションさせたトラフィック情報をオーバーレイさせることによって、WebPage オブジェクト間のユーザ遷移情報などを直感的に視覚化することが可能である。

WebPage オブジェクトを用いると、数百の WebPage を同時に画面に表示することが可能である。各ページの配置メソッドに場情報を利用することによって、表示空間自体に意味を持たせることが可能となる。

たとえば、Yahoo! JAPAN の検索結果を hyperlink アナライザに適用すると、ユーザの指定した優先順位で WebPage オブジェクトがレイアウトされ、検索結果全体を同時にブラウズすることができる。

4. システム構成

Fig.2 に本システムの構成図を示す。

(1) Hyperlink Analyzer
Web ロボットの制御、および hyperlink の解析、一時蓄積を行う。

(2) 場情報ブラウザ
数百個の WebPage オブジェクトを同時に表示し、ストレスなくブラウジングを行うため、専用のグラフィックスワークステーションを利用する。

(3) WebPage Image Handler
与えられた URL の画面表示イメージを生成するサーバ。現状ではもっとも HTML 文書の表現力が高いと思われる Windows プラットフォーム上で動作する。

5. まとめ

本稿では、WWW におけるもっとも基本的な情報である hyperlink のみから場情報を得るための方法と、得られた場情報を効果的にブラウズするためのアプリケーションを提案した。

現状では、コンテンツノード間の探索には膨大な時間を必要とする。しかし、検索エンジンがドキュメントだけではなく、hyperlink の状態も検索可能になれば、非常に実用的なツールとして利用価値が高いのではないかと思われる。

今後は、汎用的な URL カテゴライズのアルゴリズムの確立、および場情報ブラウザへのトラフィック情報表示機能の追加、場情報に適したブラウジング GUI など、実用的なツールとしての完成度を高めていく予定である。

Fig.2 場情報ブラウザのシステム構成

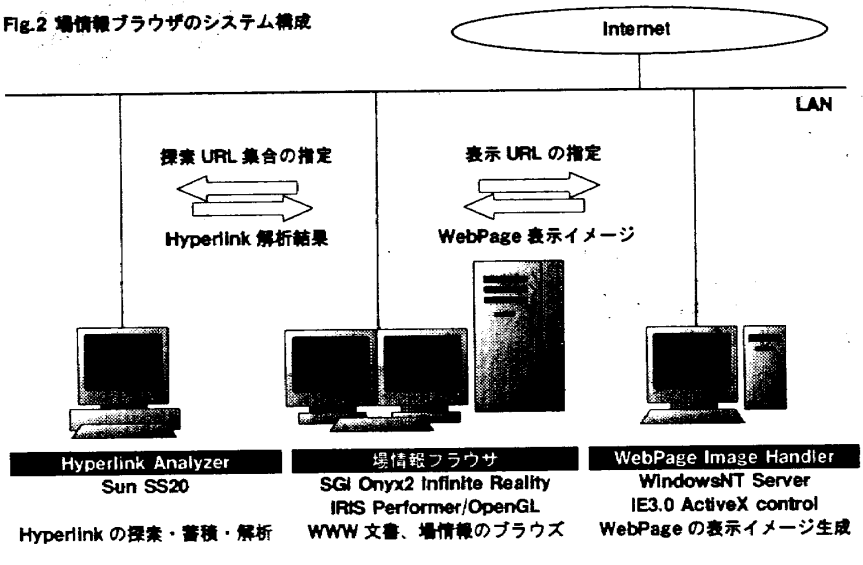


Fig.3 場情報ブラウザの実行例