

4U-01 WWW 情報における構造視覚化と検索機能の統合 *

沢井 宏† 大和田 勇人† 溝口 文雄†

東京理科大学 理工学部†

1 はじめに

ユーザが訪れた Web サイト (ホームページ) のリンク構造を視覚化することで、ユーザは訪れたサイトの規模と構造を容易に理解することができる。我々は WWW 情報構造視覚化ツール *WebMap* を開発してブラウザと連動させることにより、ユーザのブラウジング支援を行なった [1]。しかしながら、情報構造が膨大で複雑な場合、これを単に視覚化しても複雑なマップが表示されるだけで、所望する情報を発見しやすくなるはずではない。

このような問題を克服するために、本稿では WWW ブラウジングで頻繁に利用されている検索機能を *WebMap* に統合して、検索結果の視覚化を行う。具体的には、ユーザが所望する情報が *WebMap* に表示された情報構造上のどの位置に存在するかを理解しやすくする「インタラクティブ情報検索」を提案する。これは単にキーワードを含む文書が羅列される既存の情報検索サービスの欠点も克服できるものと考えられ、検索と視覚化の統合による効果を評価実験により明らかにする。

2 インタラクティブ情報検索

インタラクティブ情報検索は、WWW 情報の視覚化と検索を相互に繰り返し行いながら、WWW 情報にアクセスする方法である。

視覚化する WWW 情報構造は Web サイトやホームページを木構造として解析し、表示方法は *Hyperbolic Tree* [1] をモデルにしている。個々のノードは色と高さの 2 つのパラメータを認識できる円錐形にしており、ノードをマウスクリックすると対応するハイパーテキストをブラウザに表示し、ノードにマウスカーソルを合わせるとテキストのタイトル情報を表示できる。

2.1 ノードの高さとフィルタリングによる検索の容易性

インタラクティブ情報検索の特徴は、図 1 に示すように、(1) ノードの高さによる検索の適合度の理解と、(2)

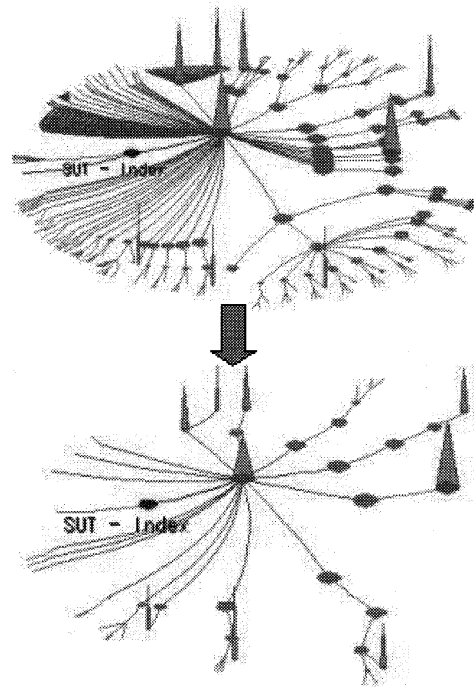


図 1: フィルタリングの様子。(キーワードを含む枝を残すようにフィルタリングを行う。ノードの高さは、検索の適合度を示す。)

フィルタリング (枝刈り) による情報構造の縮小を、繰り返し行いながら所望の情報を発見できる点にある。

検索は、ユーザが属性 (HTML タグ) と値 (キーワード) を入力し、*TFIDF* 理論を利用した各テキストの適合度 [2] を求めるようにした。

検索を行なうと、個々のテキスト t を表すノードの色 c_t が入力したキーワードの有無を示し、ノードの高さ h_t が適合度 FIT (ただし $0 \leq FIT \leq 1$) として示される。ノードの高さから、ユーザは所望のテキストが表示された情報構造のどの位置に多く含まれるかを理解できる。問い合わせ $B_{1, \dots, k}$ が連言として与えられた時、ノードの高さ h_t は、

$$h_t = h_{max} \times \prod_{i=1}^k FIT(t, B_i)$$

と表すことができ、予め定めたノードの高さの最大値

*Integrating Structure Visualization and Retrieval of WWW Information

†Hiroshi SAWAI, Hayato OHWADA, Fumio MIZOGUCHI

‡Dept. of Faculty of Sci. and Tech. Science University of Tokyo

h_{max} の範囲内に必ず納めることができる。

検索の後、ユーザが入力したキーワードを含む枝だけを残すようにフィルタリング(枝刈り)を行なうことができる(図1参照)。ツリーの末端に位置するノードはキーワードを含むテキストを示すので、ユーザは検索結果となるテキストに到達するまでのパス(道のり)を理解することができる。また、WWWの構造は縮小されるので、ユーザにとって意味のあるノードに焦点をあてやすくなる効果が得られる。

3 システム構成

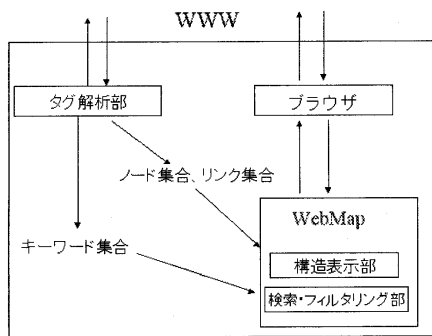


図2: システム構成図

インタラクティブ情報検索を備えた視覚化ツール *WebMap* を図2のようなシステム構成により実装した。実装は全てJavaで行なっているのでOSに依存せず起動できる。*WebMap*のフレームには検索やフィルタリングを行なうための操作パネルを用意してある。ツリーの表示に必要なノード集合、リンク集合はデータベースに保存しているので、フィルタリングを何度行っても元の状態に復元することができる。

4 評価実験

インタラクティブ情報検索の特徴である検索とフィルタリングの効果の評価を行った。*WebMap*に比較的大規模なサイト(本大学のWebサイト、テキスト数2089)を表示し、研究に関するキーワードを入力して、検索とフィルタリングからどのような検索パスが得られるかを調べた(図1参照)。例えば「環境工学」というキーワードで、*WebMap*から「環境工学」を学ぶことができる学科(または研究室)の入口となるページを発見できれば成功である。

評価基準は、(1)各キーワードに対する検索件数と、(2)パスの末端ではなく中間に位置する「中間ノード」、(3)中間ノードの子孫でキーワードを含んでいるノード数の平均、とした。

表1に実験結果を示す。中間ノードの平均は7であり、

表1: 検索-フィルタリングの結果例

問い合わせ	検索件数	中間ノード数	子孫ノード数平均
人工知能	39	4	8
スポーツ	20	2	3
環境工学	36	5	3
環境	222	20	7
地球, 環境	39	4	6
環境, 研究	195	18	7
環境工学, 研究	35	5	3
ロボット	53	4	16
ロボット, 研究	42	3	14
平均	75	7	7

その中には「環境工学」や「ロボット」を研究している研究室の入口のページも含まれていた。また、中間ノードの子孫でキーワードを含むノードが平均7つあった。これらの子孫ノードは、上位に位置する中間ノードの示すテキストをブラウザで閲覧した後にチェックすればよい。また、中間ノードが多数ある場合は、検索の適合度を示すノードの高さの高いものから順にアクセスすればよい。

実験結果から、入力語句があいまいで多数の検索結果が出力される場合、構造表示からキーワードを含む文書の分布を理解し、中間ノードからブラウジングすることにより所望の情報が発見しやすくなるといえる。

5 考察

比較研究として、複数の *Hyperbolic Tree* を3次元に配置したマップ [3] があるが、キャンパスに膨大なノードを表示するよりもユーザにとって意味のあるノードに注目できるマップを提供した方がブラウジング支援の要素が高いと考えている。また、図2のように個々のサイトがリンク集合、ノード集合をXML形式で公開した上で *WebMap* を配布すれば、検索機能付きサイトマップとして効果を発揮できるので、現在開発を進めている。

6 おわりに

WWW情報構造視覚化ツール *WebMap* に検索とフィルタリング機能を加えて、ユーザが所望する情報を発見しやすくした。評価実験からは、階層的に検索結果を把握できる「中間ノード」の数を示すことで、検索と視覚化の統合における有効性を示した。

参考文献

- [1] 沢井 宏, 大和田 勇人, 溝口 文雄, *Hyperbolic Tree* を利用した WWW ブラウジングの支援, 情報処理学会 第58回全国大会, 1999.
- [2] 沢井 宏, 大和田 勇人, 溝口 文雄, WWW情報の構造視覚化と検索機能の統合, 東京理科大学 情報メディアセンターテクニカルレポート, *AIST-TR-99-011*, 1999.
- [3] Tamara Munzner, Exploring Large Graphs in 3D Hyperbolic Space *IEEE Computer Graphics and Applications*, Vol 18. No 4, pp18-23, 1998.