

ドメイン名クラスタリングによる WWW 検索サービス

6Z-2

島村 栄 高野 元

NEC C&Cメディア研究所

1. はじめに

現在, WWW 上での情報発見を容易にするために多くの WWW ディレクトリサービスが公開されている。この1年ほどの間に WWW ドキュメントの自動収集能力の向上などによってディレクトリサービスの検索対象となるドキュメントが飛躍的に増加している。しかし, 検索対象ドキュメントが増えるに連れ, 単純なキーワード入力のみでの検索では検索結果としてヒットするドキュメントの数が膨大になりがちで, ある検索に対して数千~数万個の検索結果が返されることも珍しくない。そこで我々は, 膨大になりがちなディレクトリサービスの検索結果を整理するために検索結果 URL のドメイン情報を利用した検索システムを試作した。

2. 従来の問題と解決策

現在多くのディレクトリサービスが持つキーワード検索機能によって, ユーザは求める情報を発見することができる。しかし, 現状の機能には以下のような問題が生じがちである。

■適切な検索条件を考えるのは困難。

論理式や, 追加条件などを組み合わせた複雑な条件式を考えることは困難である。またこのような条件式を入れても, なにも検索にヒットしなくなったり, 逆に候補ドキュメントの数がほとんど絞り込めない場合がある。

■同じようなドキュメントを繰り返しチェックする必要がある。

検索結果全体が整理されていないため, 同じ提供組織による類似したドキュメントが繰り返し現れ, 同じような内容のドキュメントを何度もチェックさせられることが多い。

■検索結果ドキュメントをリストの順にチェックしなければならない。

ディレクトリサーバが出力する検索結果リストの順番決定のアルゴリズムは, 多くの場合ユーザにとってブラックボックスであり, 表示順をあらかじめ予想できない。このため, 必要な情報を発見するまで, はじめから順にチェックする必要がある。

上記のような問題が生じる原因は, ユーザが把握不可能なルールによって順位付けされる検索結果の表示方式と, インタラクティブ性の低い絞り込み検索のインタフェースであると考えられる。

これらの問題に対して, 今回我々は以下のような解決を図った。

- 検索結果表示を Java のアプレットで実現する。
インタラクティブなインタフェースによって容易に検索結果の候補の数を絞り込むことができる。
- 組織を階層化して木構造でユーザに提示する。
検索結果全体の概観が容易であり, 必要に応じてさまざまな詳細度で検索結果をチェックすることができる。
- 検索結果の各ドキュメントを提供組織ごとにクラスタリングして表示する。
明確なルールで検索結果が整理されているため, 必要などころからチェックすることができる。

我々は以上のような機能を持つ組織名ナビゲーションシステムを開発した。次の節ではこのシステムについて説明する。

3. 組織名ナビゲーション

本システムはナビゲーション UI と, その要求に応じて検索結果を返すナビゲーションサーバから成る。以下ではそれぞれについて説明する。

3.1. ナビゲーション UI

図 1 はナビゲーション UI の概観である。

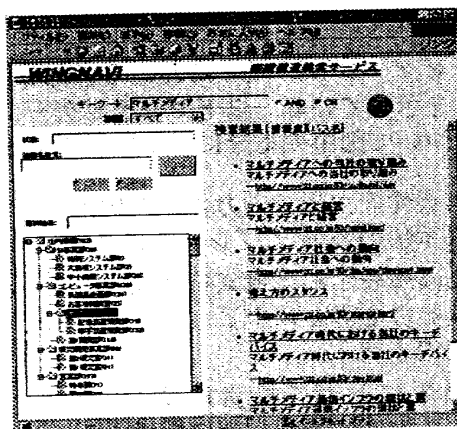


図 1: ナビゲーション UI

"Searching WWW Pages by Domain Clusters"

Hisashi Shimamura, Hajime Takano
C&CMedia Research Labs., NEC

ナビゲーション UI は 3 つのフレームに別れており、上のフレームが検索条件入力部、左下が組織階層表示部、右下が検索結果 URL 表示部である。それぞれの部分の機能は以下のとおりである。

検索条件入力部：

キーワード、更新期間などの検索条件を入力し、この検索条件を組織階層表示部に渡す。

組織階層表示部：

サーバに対して検索を行い、返された結果を、組織名で整理して階層木の形式で表示する。

図 2 は階層木表示の例である。

検索結果 URL 表示部：

組織階層表示部でマウスクリックされた組織内の検索結果 URL を表示する。

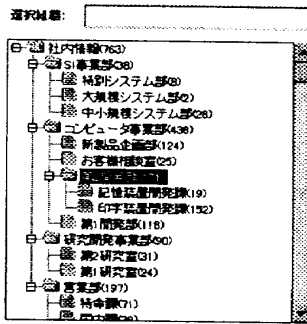


図 2: 組織名に基づく検索結果表示例

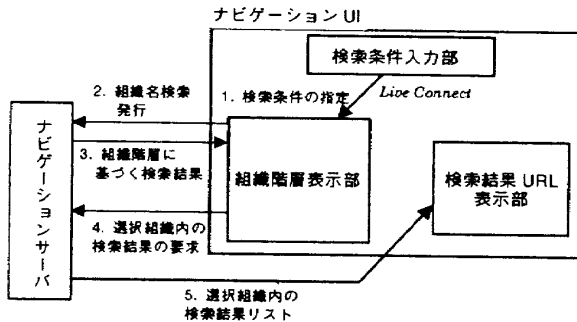


図 3: ナビゲーション UI でのデータの流れ

図 3 はナビゲーション UI における各部分間のデータの流れを示している。以下ではナビゲーション UI の動作について図中の番号に沿って説明する。

1. 検索条件入力部で入力された検索条件を組織階層表示部の Java アプレットに渡す。
2. 組織階層表示部は検索条件入力部から渡された検索条件をもとにナビゲーションサーバに対して検索を行う。
3. 組織名に基づく検索結果が返され、ツリービューの形式でユーザに提示する。これは、各組織ごとに検索条件に合致するドキュメントが何件存在するかを示す。

4. ツリービューに表示された組織名をダブルクリックすると、選択した組織内にある検索結果ドキュメントの URL をナビゲーションサーバに対して要求する。

5. その結果を検索結果 URL 表示部で表示する。

3.2. ナビゲーションサーバ

ナビゲーションサーバは WWW サーバとして実装されており、ナビゲーション UI から検索条件を引数としてサーバアプリケーションを起動することで組織名に基づいた検索が行われる。

ナビゲーションサーバはキーワード検索の対象となる、あらかじめ収集した WWW ドキュメントの DB のほかに、組織名の DB を持つ。組織名 DB に格納される組織名データには、組織名とインターネットドメイン名、階層の深さなどの情報が含まれる。

ナビゲーションサーバはキーワード検索結果の各 URL のインターネットドメイン名から、この組織名 DB を検索して組織名を求め、その組織名で整列した検索結果をナビゲーション UI へ返す。

4. おわりに

従来のディレクトリサービスにおいて膨大な検索結果から求める情報を発見することが困難であった。我々はこの問題に対する一つの解として、検索結果を個々の候補ドキュメントの提供組織で整理してユーザに提示するナビゲーションシステムを開発した。このシステムの特徴は以下のとおりである。

- 検索結果をそれぞれの提供組織ごとにまとめて表示することで、必要と感じた組織の検索結果のみを閲覧することができる。
- 階層木構造でユーザに提示することで、大量の検索結果全体を概観することができる。
- 木構造のビューを Java アプレットで実現し、GUI 操作のみで検索結果の絞り込みが可能である。

これらの機能によって、ユーザは膨大な検索結果から効率よく求める情報を探ることが可能である。

参考文献

[青柳] 青柳龍也, 「Java API プログラミングガイド」, 工学図書, 1996 年

[JPNIC] 「社団法人日本ネットワークインフォメーションセンターホームページ」, 日本ネットワークインフォメーションセンター, <http://www.nic.ad.jp/index-j.html>

[NC] “The LiveConnect/Plug-in Developer’s Guide”, Netscape Communications Corporation, <http://home.netscape.com/eng/mozilla/3.0/handbook/plugins/>